Multi-objective Deployment of WSNs in Underground Sheltered Space

Liangtian Wan[1(✉)], Caiyun Wang[1], Lu Sun[2], Boyu Chen[3], Jibin Zheng[4], and Xianpeng Wang[5]

[1] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China
wanliangtian@dlut.edu.cn

[2] Department of Communication Engineering, Institute of Information Science Technology, Dalian Maritime University, Dalian 116026, China
sunlu@dlmu.edu.cn

[3] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

[4] National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China
jibin_zheng@sina.cn

[5] State Key Laboratory of Marine Resource Utilization in South China Sea and School of Information and Communication Engineering, Hainan University, Haikou 570228, China
wxpeng2016@hainanu.edu.cn

Abstract. Given the increasing population, the underground spaces are widely used. The existing research on wireless sensor networks (WSNs) development has mainly focused on overground space; also, the existing work was based on an ideal deployment environment and simple sensor behaviour. As a consequence, they do not satisfy requirements for practical 3D environment. The 3D wireless sensor networks (WSNs) deployment has new difficulties in the underground space. The signal transmission is impacted by obstacles in the underground space. Thus, we propose a geometry-based 3D signal propagation model to calculate the signal path loss caused by obstacles. Then, taking both coverage and connectivity into account, we transform the WSNs deployment problem into a multi-objective optimization problem (MOP). Finally, we compare some state-of-the-art multi-objective evolutionary algorithms (MOEAs) for the WSNs deployment problem in underground sheltered space. By the experimental results, we optimize the multiobjective deployment problem of the WSNs in the underground sheltered space.

Keywords: Wireless Sensor Networks Deployment · A geometry-based 3D signal propagation model · Multi-objective evolutionary algorithms · Underground sheltered space

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 521–533, 2024.
https://doi.org/10.1007/978-3-031-65126-7_43

6. Cowling, J., Myers, D., Liskov, B., Rodrigues, R., Shrira, L.: HQ replication: a hybrid quorum protocol for byzantine fault tolerance. In: Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI), pp. 177–190. USENIX Association (2006)
7. Kotla, R., Alvisi, L., Dahlin, M., Clement, A., Wong, E.: Zyzzyva: speculative byzantine fault tolerance. ACM Trans. Comput. Syst. **27**(4), 45–58 (2009)
8. Yin, M., Malkhi, D., Reiter, M, K., Gueta, G.G., Abraham, I.: HotStuff: BFT consensus with linearity and responsiveness. In: Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing (PODC), pp. 347–356. Association for Computing Machinery (2019)
9. Liu, J., Li, W., Karame, G.O., Asokan, N.: Scalable byzantine consensus via hardware-assisted secret sharing. IEEE Trans. Comput. **68**(1), 139–151 (2019). https://doi.org/10.1109/TC.2018.2860009
10. RPBFT. https://fisco-bcos-documentation.readthedocs.io/zh_CN/latest/docs/design/consensus/rpbft.html. Accessed 20 Jul 2023
11. FISCO BCOS synchronization optimization. https://fisco-bcos-documentation.readthedocs.io/zh_CN/latest/docs/articles/3_features/31_performance/sync_optimization.html. Accessed 28 Jul 2023
12. PBFT in FISCO BCOS. https://fisco-bcos-documentation.readthedocs.io/zh_CN/latest/en/docs/design/consensus/pbft.html. Accessed 28 Jul 2023
13. FISCO BCOS introduction. https://fisco-bcos-documentation.readthedocs.io/en/latest/docs/introduction.html. Accessed 24 Jul 2023
14. Hyperledger Caliper introduction. https://hyperledger.github.io/caliper/v0.2/getting-started. Accessed 26 Jul 2023
15. Caliper stress test practice on FISCO BCOS platform. https://fisco-bcos-documentation.readthedocs.io/zh_CN/latest/docs/articles/4_tools/46_stresstest/caliper_stress_test_practice.html. Accessed 28 Jul 2023
16. FISCO BCOS adapter. https://hyperledger.github.io/caliper/v0.2/fisco-config. Accessed 28 Jul 2023

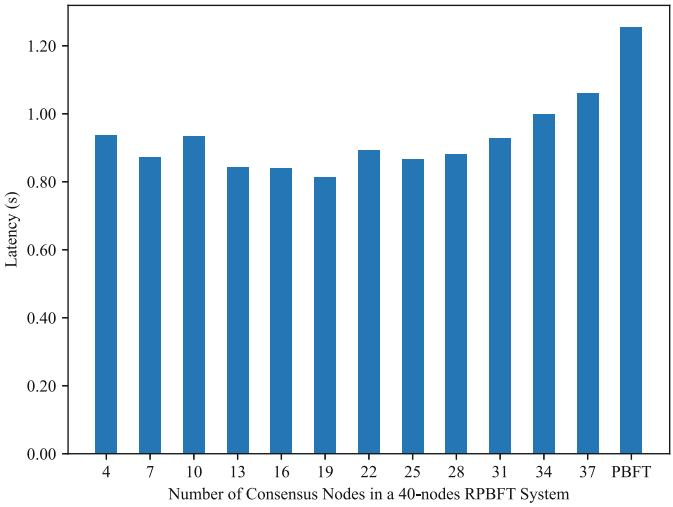


Fig. 10. Latencies of RPBFT systems with different numbers of consensus nodes.

6 Conclusion

In this paper, we introduced RPBFT, a consensus mechanism for blockchain in regulatory scenario. A separate architecture, which divides all nodes into consensus nodes and verification nodes, is designed to improve the scalability of the mechanism and thus facilitate large scale use of blockchain system. Furthermore, RPBFT has a mechanism for consensus node replacement, which strengthens its fault-tolerant capability by avoiding conspiracy. The performance evaluation reveals the advantage of RPBFT over classic PBFT in terms of throughput and latency. Systems running RPBFT gain higher throughput and lower latency when the node scale gets larger, and such advantage remains when the system contains more consensus nodes. The outperformance shows that RPBFT is more suitable for regulatory blockchain use in industries requiring large scale deployment and quick response capability.

References

1. Shrestha, R., Bajracharya, R., Shrestha, A.P., Nam, S.Y.: A new type of blockchain for secure message exchange in VANET. Digit. Commun. Netw. **6**(2), 177–186 (2020)
2. Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H.: An overview of blockchain technology: architecture, consensus, and future trends. In: IEEE International Congress on Big Data, pp. 557–564. IEEE (2017)
3. RPBFT design analysis. https://fisco-bcos-documentation.readthedocs.io/zh_CN/latest/docs/articles/3_features/32_consensus/rpbft_design_analysis.html. Accessed 20 Jul 2023
4. Kang, J., Yu, R., Huang, X., et al.: Blockchain for secure and efficient data sharing in vehicular edge computing and networks. IEEE Internet Things J. **6**(3), 4660–4670 (2019)
5. Castro, M., Liskov, B.: Practical Byzantine fault tolerance. In: Proceedings of the Third Symposium on Operating Systems Design and Implementation (OSDI), pp. 173–186. USENIX Association (1999)

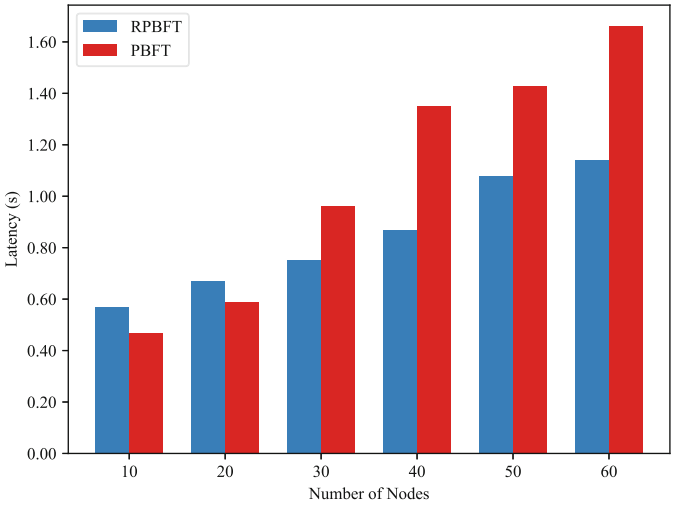


Fig. 8. Comparison of RPBFT and PBFT in terms of latency.

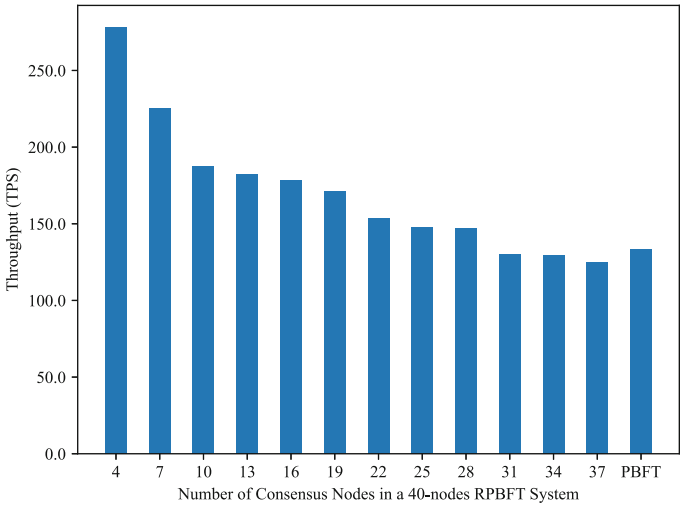


Fig. 9. Throughputs of RPBFT systems with different numbers of consensus nodes.

the node scale expands. Therefore, the design of splitting nodes into different functional parts in RPBFT can improve the scalability of consensus mechanism and realize higher performance in large scale blockchain systems.

We also test the performance of RPBFT systems with different number of consensus nodes while the total number of nodes remains constant. In this experiments we set the total number of nodes to 40, and take 40-nodes PBFT system for comparison. Figures 9 and 10 display the variation of throughput and latency respectively. In Fig. 9 we can see that the throughput keeps decreasing as the number of consensus nodes increases. This is because the consensus nodes inside the system also run PBFT algorithm, thus the increase in the number of consensus nodes also leads to a quadratic increase in communication overhead. Similarly, while the latency keeps stable between 4 and 28 consensus nodes, it starts growing steadily too as the number of consensus nodes increases, which is shown in Fig. 10.

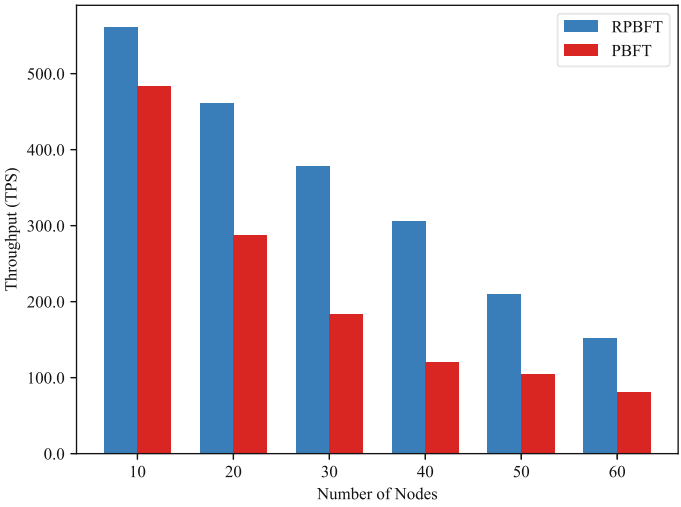


Fig. 7. Comparison of RPBFT and PBFT in terms of throughput.

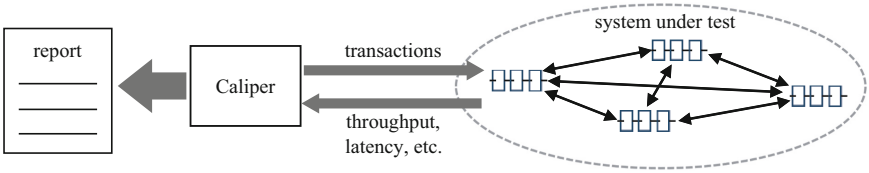


Fig. 5. The form of performance test.

Caliper Report

Basic information

DLT: fisco-bcos
Benchmark: Hello World
Description: This is a helloworld benchmark of FISCO BCOS for caliper
Test Rounds: 2
Details

Benchmark results

Summary
get
set

System Under Test

Version: 2.0.0
Size: 20 Nodes
Distribution: Single Host
Details

Summary

Name	Succ	Fail	Send Rate (TPS)	Max Latency (s)	Min Latency (s)	Avg Latency (s)	Throughput (TPS)
get	10000	0	1000.2	0.02	0.00	0.00	1000.1
set	10000	0	368.8	1.41	0.09	0.49	361.3

get

Test performance of getting name

round 0

Performance metrics

Name	Succ	Fail	Send Rate (TPS)	Max Latency (s)	Min Latency (s)	Avg Latency (s)	Throughput (TPS)
get	10000	0	1000.2	0.02	0.00	0.00	1000.1

Resource consumption

Type	Name	Memory(max)	Memory(avg)	CPU% (max)	CPU% (avg)
Process	node0(avg)	27.6MB	27.4MB	4.80	3.12

Fig. 6. Performance report generated by Hyperledger Caliper.

5 Performance Evaluation

The performance tests were done on a macOS machine with 2.3 GHz 8-Core Intel Core i9 processor and 16 GB memory. Throughput and latency are the main indicators of blockchain system performance. In this paper, we run tests over systems of different numbers of nodes to see how the throughput and latency change with the expansion of the number of nodes, which implies the scalability of consensus mechanism.

Figures 7 and 8 depict the throughput and latency of blockchain systems using different consensus mechanisms in different node scales. In this experiment, the number of RPBFT consensus nodes is set to 4, and we take PBFT for comparison. As illustrated in Fig. 7, the increase in node number leads to a quadratic growth of consensus communication overhead in PBFT system, thus resulting in a rapid decline in throughput. In comparison, the decline for RPBFT system is gentler as the node scale expands due to the much slower increase in communication overhead. For the same system scale, it can also be seen that RPBFT reaches a much higher throughput than PBFT. From Fig. 8, we can observe that the latencies of the two mechanisms are comparable in 10 and 20 nodes systems. However, due to the quadratic communication complexity across the network, the latency of PBFT systems is not as stable as that of RPBFT systems when

collected in the aforementioned process can serve as a proof to restore the consensus process.

3.4 Communication Overhead Analysis

In pre-prepare phase, the leader broadcasts to all nodes, which means $n-1$ messages are sent. Here we exclude the message that is sent to itself. In both prepare and commit phases, each node sends messages to all nodes, meaning $n(n-1)$ messages across the network in each phase. After a round of consensus finishes, the consensus nodes should send new block to verification nodes, meaning additional $m-n$ messages. Therefore, the total number of messages in each round of RPBFT can be expressed as

$$S_{\text{RPBFT}} = (n-1) + n(n-1) + n(n-1) + (m-n) = 2n^2 - 2n + m - 1. \quad (1)$$

For a PBFT system with m nodes, the total number of messages in each round is

$$S_{\text{PBFT}} = (m-1) + m(m-1) + m(m-1) = 2m^2 - m - 1. \quad (2)$$

So the difference of communication overhead caused by RPBFT can be expressed as

$$S_{\text{PBFT}} - S_{\text{RPBFT}} = 2(m-n)(m+n-1). \quad (3)$$

As $n \geq 1$ and $m > n$, we have

$$S_{\text{PBFT}} - S_{\text{RPBFT}} > 0. \quad (4)$$

This shows that the communication overhead of RPBFT is smaller than that of PBFT for the same node scale. From the equation we know that, as the node scale expands, i.e. the value of m increases, the difference of communication overhead will also increases in the case that n is constant. Moreover, for a fixed m, the difference will get smaller if n increases.

4 Implementation

This paper uses FISCO BCOS to build blockchain system. FISCO BCOS is an enterprise-level financial blockchain platform open-sourced by Chinese enterprises [13]. It provides developers with many handy tools to build and connect with blockchain systems. After building and starting a chain locally, we use the built-in console to interact with the blockchain node to verify that the chain we build functions normally.

To conduct performance test on the blockchain system, we install Hyperledger Caliper, a blockchain performance benchmark framework. It supports standardized test results and is compatible with multiple blockchain platforms [14], including FISCO BCOS. As shown in Fig. 5, Caliper plays a role of a client in the performance test [15]. It sends transactions to the system under test at a certain rate, collects performance indicators such as transaction confirmation latency and throughput, and generates a graphic report. By configuring the network condition and smart contract correctly, we connected Caliper to a blockchain system and output a report in HTML format [16], as shown in Fig. 6.

PBFT achieves distributed consistency through voting. As shown in Algorithm 2 and Fig. 4, in each round of consensus, the leader broadcasts a pre-prepare message about a selected transaction to all consensus nodes. After receiving the pre-prepare message, each node will verify whether the transaction is valid. If valid, a prepare message about that transaction is broadcast to all nodes to notify that they have received a pre-prepare message about that transaction. After that, if a node receives up to $2f + 1$ prepare messages from other nodes, it broadcasts a commit message to all nodes. Subsequently, if the node receives up to $2f + 1$ commit messages from other nodes, it confirms the transaction locally. Through the above process, the transaction will finally reach consensus in the whole network.

Algorithm 2: PBFT protocol

```
1  //pre-prepare phase
2  as a leader
3     choose a transaction tx for consensus
4     broadcast pre-prepare(tx) to all nodes (including itself)
5  for every node
6     wait for pre-prepare(tx) message from leader
7     if pre-prepare(tx) is validated then
8        broadcast prepare(tx) to all nodes
9     end if
10 //prepare phase
11 for every node
12    wait for 2f+1 prepare(tx) messages
13    broadcast commit(tx) to all nodes
14 //commit phase
15 for every node
16    wait for 2f+1 commit(tx) messages
17    commit tx
```

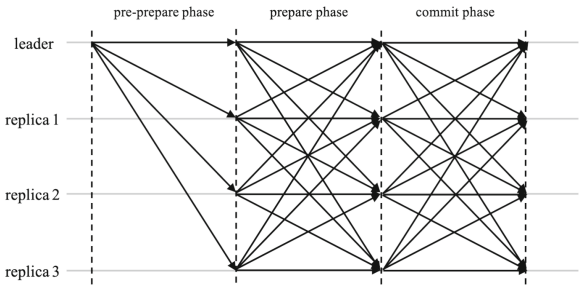


Fig. 4. Consensus process between consensus nodes in RPBFT.

Through multiple rounds of message broadcast and collection, the capacity of malicious nodes is strictly limited. As long as the number of malicious nodes is less than one-third of the consensus nodes, the consensus algorithm can run normally [12]. When the leader is abnormal, the algorithm can also replace it. At this point, the messages

and turn it into a verification node, then select one node from previous verification nodes and add it to the consensus node group. Nodes come in and out according to their node numbers, which is illustrated in Algorithm 1 and Fig. 3.

Algorithm 1: Selection and replacement of consensus nodes of RPBFT

```
1  Initialization
2  initialize m, n and k
3  initialize the nodeID of every node
4  initialize curHeight (current block height) to 0
5  sort the nodes by nodeID using 1, 2, 3,..., m
6  select n nodes whose nodeID ranges from 1 to n as consensus nodes, into con-
   sensus group
7  while the blockchain system is working do
8     wait for a new block to be added to the blockchain through PBFT consensus
   algorithm run by nodes in consensus group
9     curHeight += 1
10    if curHeight % k == 0 then
11       remove the oldest node in consensus group
12       add a new one whose nodeID is right after the latest node in the list
13    end if
14 end while
```

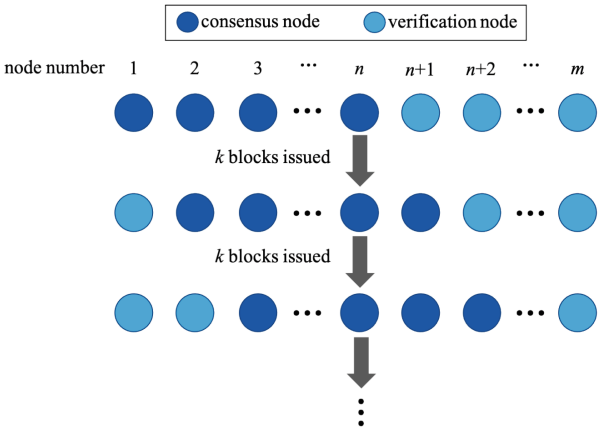


Fig. 3. The replacement of consensus nodes.

3.3 Algorithm for RPBFT Consensus Nodes

In RPBFT, consensus nodes run PBFT algorithm to reach consensus about new blocks. Assuming that the number of malicious nodes is f, according to the Byzantine fault tolerant model, the system can tolerate Byzantine errors when n is bigger than $3f$. Let's say n is equal to $3f + 1$.

3 RPBFT Consensus Mechanism

3.1 Basic Idea

As for the relation between node scale and performance, there is a simple idea that, if no matter how large the node scale is, there are only a fixed number of nodes participating in the consensus algorithm, then the performance of the system will not decline rapidly as the total number of nodes increases. By this way, more nodes could be supported in a blockchain system. According to this idea, we introduce RPBFT, a more scalable consensus mechanism for blockchain [10]. As shown in Fig. 2, RPBFT selects a fixed number of nodes randomly in the whole network as consensus nodes. Consensus nodes participate in each round of consensus algorithm jointly. When a new block is linked to the blockchain, the consensus nodes will synchronize it to other nodes (called verification nodes) using tree broadcast strategy [11]. By eliminating the impact of node scale on communication complexity, RPBFT has better scalability able to reduce latency and improve throughput in large-scale blockchain applications. Moreover, RPBFT periodically replaces the consensus nodes to ensure safety and prevent conspiracy of the consensus nodes.

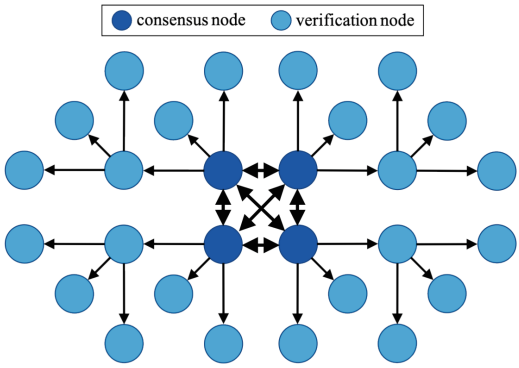


Fig. 2. RPBFT consensus mechanism.

3.2 Selection and Replacement of Consensus Nodes

Suppose the total number of nodes in the system is m. During initialization, the number of consensus nodes n and the period k are needed to be set. The meaning of k is that the replacement of the consensus node takes place every time a collection of k blocks is issued. Meanwhile, the nodeIDs of all nodes (a binary string of fixed length, unique to every node) are sorted. The sort order of each node is called node number (1, 2,…, m). When the system runs for the first time, the first n nodes become consensus nodes automatically.

In order to ensure the safety of the system, every time k blocks are produced by current consensus nodes, the system will remove a node from the consensus node group

needed to design a suitable consensus mechanism to improve scalability. With such goal, we introduce a scalable consensus mechanism called RPBFT, and verify its performance on real blockchain systems.

The contributions of this paper are summarized as follows. We introduce a scalable consensus mechanism to improve the performance of large-scale blockchain systems. The design of the consensus mechanism is based on the idea of selecting a part of the nodes in the system to run consensus protocol. We also introduce the method to select and replace the nodes that participate in consensus procedures to ensure security. To verify the effectiveness of the consensus mechanism, we evaluate its performance in real blockchain systems.

The remainder of this paper is organized as follows. Section 2 discusses related studies. Section 3 presents RPBFT consensus mechanism. Section 4 describes how we implement a blockchain system that runs the consensus mechanism, and how to carry out performance tests. Section 5 provides the test results and the performance analysis of RPBFT. Finally, Sect. 6 concludes the paper.

2 Related Work

In 1999, Miguel Castro and Barbara Liskov proposed Practical Byzantine Fault Tolerance (PBFT) [5]. PBFT runs in an environment where the identity of every node is known to all in advance, making it a natural fit for blockchain systems that aim to be secure and controllable. It can achieve Byzantine fault tolerance, which means it resists malicious behaviors from participants. Besides, it's relatively mature in terms of implementation. Due to these, it is one of the most widely-used blockchain consensus mechanisms.

However, PBFT has the problem of high communication overhead. In each round of consensus, all nodes need to broadcast their consensus messages to the whole network, making the communication complexity of the system $O(n^2)$ (n is the number of nodes participating in the consensus procedures). As a result, with the expansion of node scale, the performance deteriorates rapidly, and therefore the system would be unable to support large-scale blockchain applications.

Many researchers have proposed improved consensus mechanisms based on PBFT. Cowling et al. [6] propose Hybrid-Quorum (HQ), a hybrid Byzantine-fault-tolerant consensus mechanism which has a lightweight Byzantine quorum protocol. In HQ, all replicas have no need to interact with each other to keep consistent. Therefore the communication complexity is reduced. Kotla et al. [7] proposed Zyzzyva that uses speculation to reduce the cost of replication process. In Zyzzyva, replicas reply to the request of a client without first running an expensive three-phase commit protocol. Such design improves the system performance when there is no Byzantine faulty replicas. Yin et al. [8] proposed HotStuff, a consensus mechanism that combines with aggregate signature technique. Compared with PBFT, it simplifies the process of leader replacement and reaches a lower communication complexity, thus has better scalability. Some researchers try to introduce Trusted Execution Environment (TEE) into the design of BFT consensus mechanism, and one of the representative results is FastBFT [9]. By using TEE, the security model is simplified, thus achieving higher security threshold. Besides, the authors also design a tree broadcast strategy to reduce the communication complexity of the mechanism.

digital signature, and distributed application. Unlike traditional centralized architectures, blockchain decentralizes the network structure to avoid a single point of failure. It uses distributed consensus to achieve fault-tolerance against malicious behaviors, and uses a chain structure with cryptographic method to make data tamper-proof. These features make the data on a blockchain consistent, traceable and auditable, which is regarded to have the power of enabling secure data sharing and efficient cooperation [2].

However, the scalability of consensus mechanism limits the employment of blockchain in some industries that requires large scale deployment and quick response capability [3], like intelligent transportation system (ITS) industry. A typical way to integrate blockchain with transportation systems is to deploy nodes on roadside units to harness the hardware capability of roadside infrastructure [4], as depicted in Fig. 1. During high traffic volumes like traffic jams or rush hours, many vehicles may communicate with roadside units to update their condition on the blockchain, thus generating massive amount of transactions in a short time. In this scenario, a less-scalable consensus mechanism would slow down the processing of transactions, causing the pending transactions to gather or even lose.

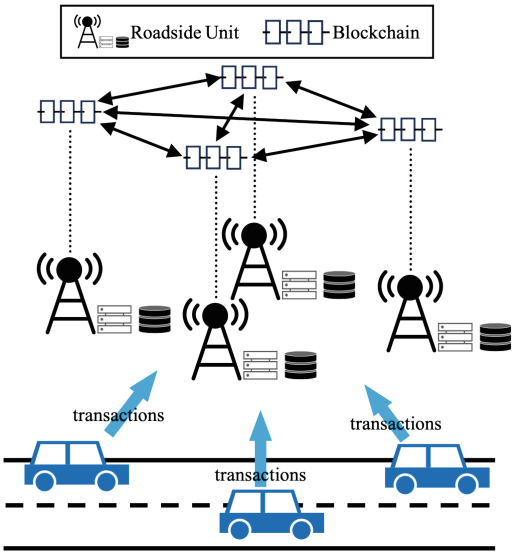


Fig. 1. A typical way to integrate blockchain with transportation systems.

To meet regulatory requirements, the identity of each node in a blockchain network must be known to all, so it's needed to adopt permissioned blockchain, a kind of blockchain that uses classic consensus mechanism to achieve distributed consistency. Such consensus mechanism requires all nodes in the network to communicate with each other frequently. While this messaging method plays an important role in maintaining the safety and liveness of a decentralized system, it incurs a huge amount of communication overhead when the scale of the system is large. Due to this, such consensus mechanism is difficult to support large-scale deployment of blockchain systems. Therefore, it's highly



RPBFT: A Scalable Consensus Mechanism for Large Blockchain Systems

Weizhe Wang[1,2,3], Daxin Tian[1,2,3,4(✉)], Xuting Duan[1,2,3,4], and Jianshan Zhou[1,2,3]

[1] School of Transportation Science and Engineering, Beihang University, Beijing, China
{weizhewang,dtian,duanxuting,jszhou}@buaa.edu.cn
[2] State Key Lab of Intelligent Transportation System, Beijing, China
[3] Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Beijing, China
[4] Zhongguancun Laboratory, Beijing, China

Abstract. As an emerging technology, blockchain has enabled trustworthy data sharing and efficient cooperation in many industries. It can also make data traceable and auditable, which perfectly meets the regulatory demands for data security and privacy. However, the employment of blockchain is limited in some fields due to the poor scalability of the consensus mechanism. Therefore, this paper introduces a scalable consensus mechanism called RPBFT to improve the performance of large-scale systems. To demonstrate the advantage of RPBFT over classic consensus mechanism, we find a way to implement it in blockchain systems and conduct performance tests to evaluate the throughput and latency of the established systems. The results show that RPBFT is superior and more suitable for large-scale blockchain systems.

Keywords: Blockchain · Consensus mechanism · Scalability

1 Introduction

With the rapid development of information technology and the digital transformation of many industries, a large amount of data is generated, collected and shared. As an important way to release the value of data, data sharing breaks the boundaries of cooperation between enterprises and organizations, which is significant for promoting the efficiency of operation. However, in the real business of data sharing, it's inevitable to face a series of data security issues and regulatory concerns. So it's important to take measures to ensure the security of data sharing.

In order to achieve secure data sharing that meets regulatory requirements, many researchers have explored related technologies, and one of the highly promising technologies is blockchain. Blockchain is a new kind of distributed database [1], which integrates various technologies such as peer-to-peer communication, consensus mechanism,

This research was supported by the National Key Research and Development Program of China under Grant No. 2022YFC3803700.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 509–520, 2024.
https://doi.org/10.1007/978-3-031-65126-7_42

36. Yan, L., Yin, C., Chen, H., Rong, W., Xiong, Z., David, B.: Learning resource recommendation in e-learning systems based on online learning style. In: Qiu, H., Zhang, C., Fei, Z., Qiu, M., Kung, S.-Y. (eds.) KSEM 2021. LNCS (LNAI), vol. 12817, pp. 373–385. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82153-1_31
37. Zhang, X., Liu, H., Chen, X., Zhong, J., Wang, D.: A novel hybrid deep recommendation system to differentiate user's preference and item's attractiveness. Inf. Sci. **519**, 306–316 (2020)
38. Zhao, Z.D., Shang, M.S.: User-based collaborative-filtering recommendation algorithms on hadoop. In: 2010 Third International Conference on Knowledge Discovery and Data Mining, pp. 478–481. IEEE (2010)
39. Zheng, Y., Mobasher, B., Burke, R.: CARSKit: a java-based context-aware recommendation engine. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1668–1671. IEEE (2015)

16. Koren, Y., Bell, R.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 77–118. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_3
17. Kupisz, B., Unold, O.: Collaborative filtering recommendation algorithm based on hadoop and spark. In: 2015 IEEE International Conference on Industrial Technology (ICIT), pp. 1510–1514. IEEE (2015)
18. Li, G., Zhang, J.: Music personalized recommendation system based on improved KNN algorithm. In: 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 777–781. IEEE (2018)
19. Lipkus, A.H.: A proof of the triangle inequality for the tanimoto distance. J. Math. Chem. **26**(1–3), 263–265 (1999)
20. Lodderstedt, T., Bradley, J., Labunets, A., Fett, D.: Oauth 2.0 security best current practice. IETF Web Authorization Protocol, Tech. Rep. draft-ietf-oauth-security-topics-16 (2020)
21. Lynn, N., Emanuel, A.: A review on recommender systems for course selection in higher education. In: IOP Conference Series: Materials Science and Engineering, vol. 1098, p. 032039. IOP Publishing (2021)
22. Mahdi, M.N., Ahmad, A.R., Ismail, R., Natiq, H., Mohammed, M.A.: Solution for information overload using faceted search-a review. IEEE Access **8**, 119554–119585 (2020)
23. Nguyen, L.V., Vo, Q.T., Nguyen, T.H.: Adaptive KNN-based extended collaborative filtering recommendation services. Big Data Cogn. Comput. **7**(2), 106 (2023)
24. Owen, S., Friedman, B.E., Anil, R., Dunning, T.: Mahout in Action. Simon and Schuster (2011)
25. Patro, S.G.K., et al.: A hybrid action-related k-nearest neighbour (HAR-KNN) approach for recommendation systems. IEEE Access **8**, 90978–90991 (2020)
26. Pawlicka, A., Pawlicki, M., Kozik, R., Choraś, R.S.: A systematic review of recommender systems and their applications in cybersecurity. Sensors **21**(15), 5248 (2021). https://doi.org/10.3390/s21155248
27. Schneider, S.C.: Information overload: causes and consequences. Hum. Syst. Manag. **7**(2), 143–153 (1987)
28. Sharma, V., Trehan, T., Chanana, R., Dawn, S.: StudieMe: college recommendation system. In: 2019 3rd International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), pp. 227–232. IEEE (2019)
29. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 421425 (2009)
30. Subramaniyaswamy, V., Logesh, R.: Adaptive KNN based recommender system through mining of user preferences. Wireless Pers. Commun. **97**, 2229–2247 (2017)
31. Tan, H., Guo, J., Li, Y.: E-learning recommendation system. In: 2008 International Conference on Computer Science and Software Engineering, vol. 5, pp. 430–433. IEEE (2008)
32. Tarus, J.K., Niu, Z., Mustafa, G.: Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. Artif. Intell. Rev. **50**, 21–48 (2018)
33. Willett, P.: Similarity-based approaches to virtual screening. Biochem. Soc. Trans. **31**(Pt 3), 603–606 (2003). https://doi.org/10.1042/bst0310603. PMID: 12773164
34. Willett, P.: Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today **11**(23–24), 1046–1053 (2006)
35. Xiaojun, L.: An improved clustering-based collaborative filtering recommendation algorithm. Clust. Comput. **20**, 1281–1288 (2017)

relevant and precise recommendations, assisting both graduates and current university students in their college selection process. It effectively addresses the challenges and time costs caused by data overload on the Internet.

Despite the decent accuracy of our recommendation system, there are still challenges to address. Ensuring the authenticity of user-uploaded results is crucial, and we propose using big data techniques for initial identification and manual processing of anomalous data. Additionally, economic costs and school location were found to be significant factors in decision-making. Further research and development are needed to incorporate these factors into our recommender system to improve its comprehensiveness and accuracy.

References

1. Adeniyi, D.A., Wei, Z., Yongquan, Y.: Automated web usage data mining and recommendation system using k-nearest neighbor (KNN) classification method. Appl. Comput. Inf. **12**(1), 90–108 (2016)
2. Adomavicius, G., Zhang, J.: Stability of collaborative filtering recommendation algorithms. Citeseer **10**(1.221), 7584 (2012)
3. Aggarwal, C.C.: Recommender Systems. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29659-3
4. Anand, S.S., Mobasher, B.: Intelligent techniques for web personalization. In: Mobasher, B., Anand, S.S. (eds.) ITWP 2003. LNCS (LNAI), vol. 3169, pp. 1–36. Springer, Heidelberg (2005). https://doi.org/10.1007/11577935_1
5. Anwar, T., Uma, V., Hussain, M.I., Pantula, M.: Collaborative filtering and KNN based recommendation to overcome cold start and sparsity issues: a comparative analysis. Multimedia Tools Appl. **81**(25), 35693–35711 (2022)
6. Awan, M.J., et al.: A recommendation engine for predicting movie ratings using a big data approach. Electronics **10**(10), 1215 (2021)
7. Baskota, A., Ng, Y.K.: A graduate school recommendation system using the multi-class support vector machine and KNN approaches. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 277–284. IEEE (2018)
8. Bawden, D., Holtham, C., Courtney, N.: Perspectives on information overload. In: Aslib Proceedings. vol. 51, pp. 249–255. MCB UP Ltd (1999)
9. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adap. Inter. **12**, 331–370 (2002)
10. Chung, N.C., Miasojedow, B., Startek, M., Gambin, A.: Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data. BMC Bioinf. **20**(15), 1–11 (2019)
11. Edmunds, A., Morris, A.: The problem of information overload in business organisations: a review of the literature. Int. J. Inf. Manage. **20**(1), 17–28 (2000)
12. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: KNN model-based approach in classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) OTM 2003. LNCS, vol. 2888, pp. 986–996. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39964-3_62
13. Hardt, D.: The oauth 2.0 authorization framework. Tech. rep. (2012)
14. Isinkaye, F., Folajimi, Y., Ojokoh, B.: Recommendation systems: principles, methods and evaluation. Egypt. Inf. J. **16**(3), 261–273 (2015)
15. Kajdanowicz, T., Indyk, W., Kazienko, P.: Mapreduce approach to relational influence propagation in complex networks. Pattern Anal. Appl. **17**, 739–746 (2014)

the effectiveness of our recommendation system by examining whether the final choices of these simulated users were adequately covered within our generated recommendation list (Table 1).

Table 1. Test of accuracy

Hit Rate Level	Number	Rate
In the first recommendation	7	14%
In the top three recommendations	29	58%
In the top five recommendations	37	74%
In the top ten recommendations	44	88%

Our recommendation system achieved great results in accuracy tests. Specifically, 58% of the actual final choices were successfully included within the top three recommendation predictions, while 74% of the results were captured within the top five recommendation predictions. Furthermore, the overall hit rate of our recommendation list stood at an impressive 88%.

We speculate that the unsuccessful predictions may have been influenced by other factors, such as economic conditions and region. These factors were not taken into account in our recommendation system, but they could be important factors affecting some students.

5 Conclusion

Our recommendation system efficiently filters internet information, saving users time and addressing information overload. It generates a curated list of colleges based on multiple factors, ensuring objectivity and high success rates. Users can obtain recommended institutions without searching elsewhere, streamlining the application process. Our website offers user-friendly services and comprehensive information to empower applicants in making well-informed decisions for higher education. It aims to be a holistic and effective tool catering to diverse user needs.

We present a recommendation system using KNN-Based Collaborative Filtering to predict suitable programs and universities for users. Our system utilizes data from OpenCS as the foundational dataset and continually updates the database with information from platform users for real-time relevance. To accommodate larger data volumes and further optimize the usability of our recommendation system, we integrate Hadoop-based distributed clustering collaborative filtering. This approach enables efficient and parallel processing of vast datasets, leading to enhanced scalability and improved overall system performance.

Our recommendation system has shown positive outcomes in accuracy testing experiments, proving its effectiveness and usefulness. The system provides

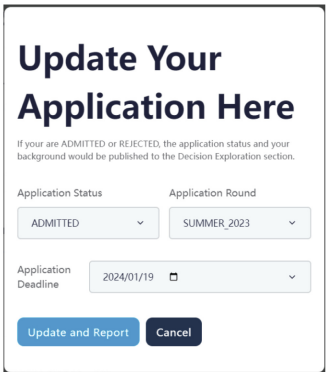


Fig. 7. Report Application

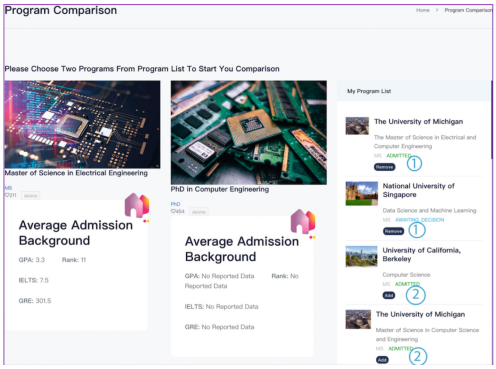


Fig. 8. Comparison

the programs based on their personal preferences and priorities, fostering a more engaged and personalized decision-making experience. By offering this level of control and transparency, our platform enhances the user's understanding and engagement in the college selection process, promoting greater satisfaction and confidence in their final decisions.

Dicision Exploration. Within our platform, applicants have access to an online chat feature that facilitates communication with other applicants. Moreover, applicants can explore detailed information about the schools they have applied to, providing them with comprehensive insights into the institutions of interest. Furthermore, applicants have the capability to review the application background of the decision report's owner. This feature empowers applicants to gauge the credibility and relevance of the decision report.

Dataset. The dataset used in our website was constructed through a two-fold approach. Firstly, we collected relevant data from Open-Source Datasets, ensuring a diverse and comprehensive pool of information. Additionally, we sourced program-related data from OpenCS.app, further enriching our dataset.

Moreover, our platform encourages active user participation, enabling users to contribute to the expansion of our dataset. Users have the option to upload their admission status and application backgrounds, providing valuable real-world data that enhances the accuracy and relevance of our system. This collaborative data-sharing approach ensures the dataset remains up-to-date and reflective of real user experiences, ultimately contributing to the overall effectiveness and reliability of our recommendation system.

Test of Accuracy. To evaluate the accuracy of our recommendation system, we conduct 50 tests using the background of 50 students whose final choice is known, which we simulate as input from website users. In each test, we evaluated

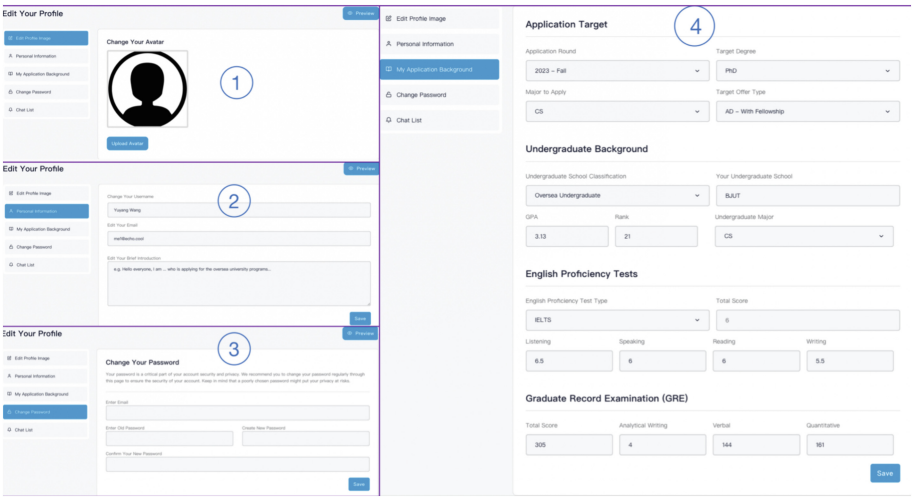


Fig. 5. Personal Information

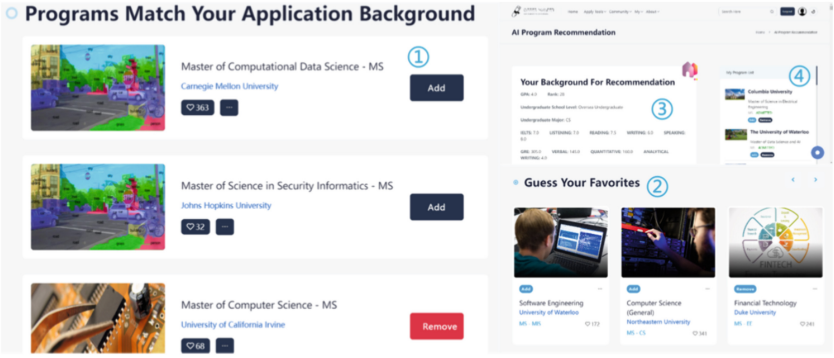


Fig. 6. Recommendation System

Report Application. (Figure 7) Users can upload and change their application status. Only applications with offers will be recorded in our database. However, the approach to guarantee the authenticity of users' data is still a problem that needs to be solved.

Comparison. Figure 8 directs users to compare two different programs selected from their programs list, which display the basic information and average background about this pair of programs.

By incorporating this comparative functionality, users gain deeper insights into the programs of interest, enabling them to make more informed decisions beyond solely relying on the recommendations provided by the system. This interactive and user-driven comparison process empowers applicants to assess

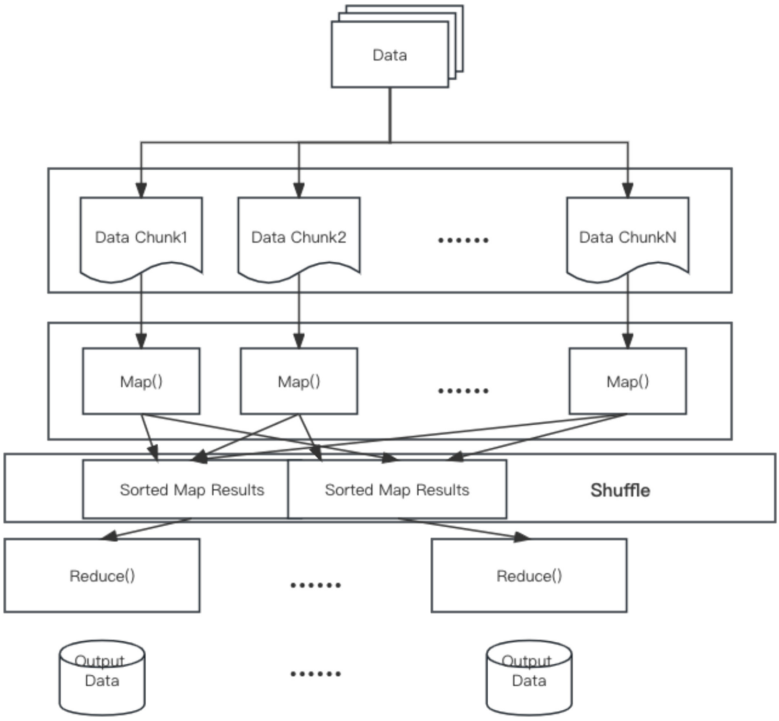


Fig. 4. The MapReduce model

revisions and a user forum for communication and sharing. We aim to revolutionize the college application process by providing comprehensive support and resources for informed decision-making and academic success.

Personal Information. Users can edit their profiles including avatar, username, personal information, password, etc. (1)–(3). They can also upload or update their backgrounds to guide the recommendation system in generating matching recommendations (4) (Fig. 5).

Recommendation System. Figure 6 illustrates the functionality of our recommendation system. Leveraging the user's basic information as input, our system employs KNN-based Collaborative Filtering to generate a personalized recommendation list.

To enhance the user experience and facilitate informed decision-making, our website provides detailed views of recommended programs, offering comprehensive information about the school and specific programs. We also integrate advanced virtual 3D campus technology for virtual campus exploration. Furthermore, our platform includes various functions and resources to enrich the user experience and assist applicants in their college selection process.

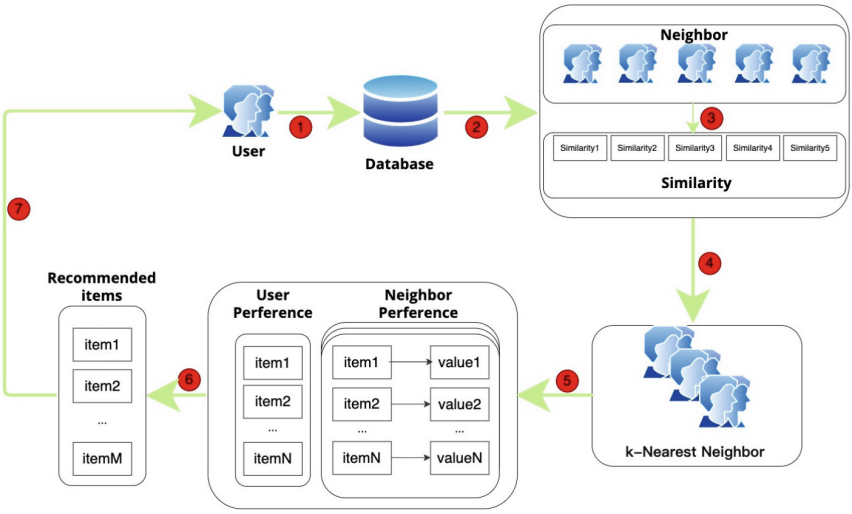


Fig. 3. Recommendation system workflow

3.3 How to Improve CF Performance with Hadoop

The Mahout library [24] provides implementations of paralleled versions of algorithms in the field of machine learning for the Hadoop platform. It focuses on classification, grouping, and collaborative filtering algorithms. The Mahout library was used for creating a recommendation system based on the Apache Hadoop technology. In the Mahout library, the two most important programs which realize the paralleled CF algorithm based on items are RecommenderJob (which calculates recommendations) and ItemSimilarityJob (which calculates the similarity matrix (Eq. 2)) [38]. A full implementation of the paralleled version of the collaborative filtering algorithm based on items, according to the MapReduce paradigm [15] (Fig. 4), is realized in the form of nine consecutive jobs [17].

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}_{ItemSim.} * \begin{bmatrix} x \\ y \end{bmatrix}_{UserPrefs.} = \begin{bmatrix} xa & yb \\ xc & yd \end{bmatrix}_{UserRecs.} \tag{2}$$

4 Result

Our platform offers a highly personalized and comprehensive experience, providing users with their top 10 best-matched colleges and programs. Unlike other sites that rely solely on exam scores, we consider multiple selection parameters for our college recommendations. Collaborative Filtering (CF) is utilized to curate recommendations based on success stories from users with similar backgrounds. Additionally, we enhance the user experience with 3D rendering and Google Earth APIs, allowing virtual campus tours. Our platform also offers paperwork

Efficiency. Efficient data processing is crucial in our recommendation system. While the utilization of Tanimoto Coefficient Similarity narrows down the results to binary form, thus enhancing efficiency to some extent, collaborative filtering efficiency still faces challenges, especially in scenarios involving large datasets. To address this, we adopt Hadoop for distributed processing, implementing a collaborative filtering recommendation algorithm on clustering. This approach significantly enhances the system's efficiency, allowing us to handle large volumes of data more effectively, improving overall performance and scalability [35,36].

Deployment Cost. The KNN algorithm offers two key advantages in our recommendation system. Firstly, it enables fast training, making processing of large datasets efficient. Secondly, its flexibility allows easy expansion of the training set. When users update their backgrounds and decisions, our recommendation model can be promptly retrained with the updated data. This dynamic training process ensures that our system remains up-to-date and responsive to users' changing preferences and needs. By leveraging the speed and adaptability of KNN, we create a recommendation platform that can accommodate a growing user base and continuously improve its performance over time.

The specific workflow of the recommendation system is shown in Fig. 3.

Step (1) user uploads personal background information to our website, Step (2) constructs a data model based on the background information, Step (3) calculates the similarity using Tanimoto Coefficient Similarity, finds K-Nearest-Neighbor users and constructs a similarity model through Step (4). Step (5) Calculate the score based on the weights of different items of all N neighboring users by recommendation engine. Steps (6)–(7) select M highest rated items to recommend to the user.

3.2 Website Architecture

Our website offers more than just a Recommendation System for master program applications. Users can take virtual campus tours using 3D Rendering and Google Earth API. They can compare multiple programs and universities side by side. Additionally, the platform provides resources like application tips, financial aid information, and career prospects. Users can communicate with university representatives and current students through our Instant Messaging (IM) service.

To enhance user experience, we focus on High Availability, Robustness, and user-friendly design. We employ advanced front-end techniques like Vue.js, Bootstrap, and Element UI. Our website is optimized for speed using caching, Content Delivery Network (CDN), and static file compression. We perform rigorous testing to catch and resolve bugs and regularly monitor the website's performance using Portainer and phpMyAdmin. We support Open Authorization (OAuth 2.0) [13] and OIDC [20] for easy login using third-party identity providers like Google, Github, and WeChat, streamlining the authentication process.

grounds. To handle large data volumes efficiently, we utilize Hadoop for improved performance.

3 Framework

In this section, we provide a comprehensive overview of our core recommendation system, presenting its structure, workflow, and underlying motivation. Additionally, we conduct a thorough performance analysis of the recommendation system, evaluating its Applicability, Stability, Efficiency, and Deployment cost (Sect. 3.1). Next, we introduce the overall architecture of our website, and some work other than the recommendation system (Sect. 3.2). Lastly, we present our implementation scheme for KNN-based Collaborative Filtering (CF) utilizing Hadoop to address the challenges posed by large data volumes and enhance system performance (Sect. 3.3).

3.1 Recommendation System Architecture

We choose user-based collaborative filtering [38] to construct a data model based on the user's important background information, use Tanimoto Coefficient Similarity [10,33] (as Eq. 1) to construct a similarity model by calculating the similarity based on the background information, and find out K-Nearest-Neighbors to construct a similarity model. Finally, all rated items in the neighborhood are prioritized by the recommendation engine [6,39] and recommended to the user. The selection of these technologies is guided by factors such as applicability, stability, efficiency, and deployment cost, ensuring the effectiveness and practicality of our recommendation system for real-world applications.

$$s\left(i,j\right)=\frac{n\left(c_i\cap c_j\right)}{n\left(c_i\cup c_j\right)}=\frac{n\left(c_i\cap c_j\right)}{n\left(c_i\right)+n\left(c_j\right)-n\left(c_i\cap c_j\right)} \tag{1}$$

Applicability. In our specific situation, user-based collaborative filtering is the suitable choice, matching user preferences with recommendations from similar situations. KNN-based similarity modeling improves accuracy, leading to precise recommendations. Our system prioritizes schools where successful application is more likely for the user, and we compute similarity using the Tanimoto coefficient [19,34].

Stability. Stability is a crucial factor to consider. Using KNN for similarity modeling allows effective filtering of abnormal information based on the most similar users. However, the presence of inauthentic data can impact the accuracy of KNN-based similarity calculations, affecting the overall precision and reliability of our recommendation system [2].

CF categories	Representative techniques	Main advantages	Main shortcomings
Memory-based CF	*Neighbor-based CF (item-based/user-based CF algorithms with Pearson/vector cosine correlation) *Item-based/user-based top-*N* recommendations	*easy implementation *new data can be added easily and incrementally *need not consider the content of the items being recommended *scale well with co-rated items	*are dependent on human ratings *performance decrease when data are sparse *cannot recommend for new users and items *have limited scalability for large datasets
Model-based CF	*Bayesian belief nets CF *clustering CF *MDP-based CF *latent semantic CF *sparse factor analysis *CF using dimensionality reduction techniques, for example, *SVD*, *PCA*	*better address the sparsity, scalability and other problems *improve prediction performance *give an intuitive rationale for recommendations	*expensive model-building *have trade-off between prediction performance and scalability *lose useful information for dimensionality reduction techniques
Hybrid recommenders	*content-based CF recommender, for example, *Fab* *content-boosted CF *hybrid CF combining memory-based and model-based *CF* algorithms, for example, Personality Diagnosis	*overcome limitations of CF and content-based or other recommenders *improve prediction performance *overcome CF problems such as sparsity and gray sheep	*have increased complexity and expense for implementation *need external information that usually not available

Fig. 2. Overview of collaborative filtering techniques [29].

The usual memory-based collaborative filtering techniques can be separated into item-based, and user-based methods.

- item-based
  The method calculates predictions based on product/item similarity rather than user similarity [4].
- user-based
  The method predicts user behavior by using a weighted sum based on average ratings of users who rated the item in the past and the average user rating.

KNN models are highly accurate and widely used in collaborative filtering (CF) recommendation systems. They have been popular since they were introduced and are known for providing reasonable explanations for their recommendations. Experiments have shown that CF with KNN-based methods significantly reduces error rates [30].

2.4 Overview of Existing School Recommendation Systems

Previous studies on school recommendation systems have utilized cosine similarity and TF-IDF vectorization for matching [28], as well as KNN and Support Vector Machine for filtering decisions [7]. However, these former model may bias recommendations towards popular schools, which might not be the best fit for all students. Moreover, updating the latter model can be computationally expensive.

In contrast, our recommendation system employs KNN-based collaborative filtering, prioritizing successful admission results of students with similar back-

- Collaborative Filtering (CF)
  CF refers to recommending items to target users by identifying users with similar interests. This method is designed to help users get appropriate recommendations through individuals or groups with the same preferences or behaviors.
- Content-Based (CB)
  The content-based method recommends products to users using their historical data. It analyzes the user's past searches and purchases to suggest related items. The method heavily relies on user ratings, making it especially valuable in business, information, and education domains [26,37].
- Knowledge-based
  Recommendation methods are employed to assist users in making informed decisions while purchasing complex items with various attributes. Users often seek items with specific features like car models, engine types, or house interior designs. In certain business contexts, finding ratings for recommendations is challenging due to expensive items and low purchase demand. These methods are particularly useful in cold start situations where traditional rating-based approaches may not be feasible [32].
- Hybrid
  These approaches bring together the advantages of different types of recommendation systems. The aim is to create recommendation systems using techniques that are more efficient and effective in terms of performance [3,9].

2.2 KNN Are Suitable for Recommendation System

The k-Nearest-Neighbours (kNN) is a non-parametric classification method, which is simple but effective in many cases [12].

KNN stands out as one of the most effective neighboring algorithms. Given its proven success, the KNN algorithm finds widespread application in numerous recommendation systems, particularly for computing user similarities [1,3]. However, the KNN algorithm also has many drawbacks. Therefore, many special KNNs for different circumstances have been proposed. Such as Adaptive KNN [30], Improved KNN [18], and A hybrid action-related KNN [25].

2.3 The Characteristics of Collaborative Filtering

Collaborative filtering (CF) is a highly successful method for building recommendation systems. It uses known user preferences to predict unknown preferences [29]. CF can be categorized into three types: Memory-based, Model-based, and Hybrid (see Fig. 2). Memory-based CF calculates user similarities from interaction data to make recommendations. Model-based CF uses machine learning algorithms to build predictive models. Hybrid CF combines both memory-based and model-based techniques for improved and accurate recommendations [29].

Section 2 presents the related work on the recommendation systems, as well as the fundamental principles of KNN (K-Nearest Neighbor) and CF (Collaborative Filtering). In Sect. 3, we present our framework, which comprises essential stages such as Information Collection, Model Construction, Similarity Calculation, k-Neighbors Identification, Recommendation Engine Development, and Performance Enhancement using the Hadoop Cluster. The outcomes and findings of our approach are outlined in Sect. 4. Finally, in Sect. 5, we provide a concise summary and conclusion, highlighting the key contributions and implications of the research presented in this paper.

2 Related Works

This section is dedicated to a comprehensive exploration of existing recommendation systems, where we analyze various implementation methods and algorithms in the context of their applicability (Sect. 2.1). Subsequently, we delve into the details of KNN (K-Nearest Neighbor) and its relevance and suitability for recommendation systems (Sect. 2.2). Moving forward, we examine the distinctive characteristics of collaborative filtering and assess the feasibility of incorporating KNN-based collaborative filtering (Sect. 2.3). Finally, we provide a succinct overview of the current landscape of school recommendation systems (Sect. 2.4).

2.1 Technical Options for Recommendation Systems

The utilization of efficient and precise recommendation techniques holds paramount significance for a system aiming to deliver valuable and relevant recommendations to its individual users. Figure 1 shows the anatomy of different recommendation filtering techniques.

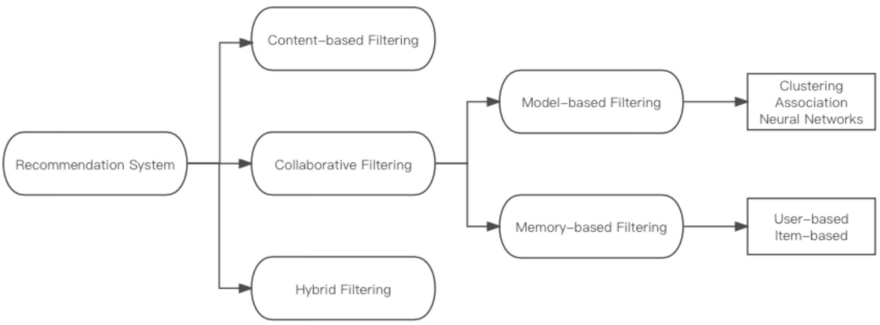


Fig. 1. Recommendation filtering techniques

In general, recommendation systems are categorized according to the methods they use. These methods can be categorized into four main groups [14,21]. The different types of recommendation approaches or methods are briefly discussed below:

1 Introduction

The Internet's rapid growth has led to an overwhelming amount of information, causing information overload [11,27] for users. To address this, two main approaches are used: controlling information generation and filtering information access. However, controlling information generation has become ineffective due to the Internet's speed of development [8]. Thus, there is an urgent need for robust filtering mechanisms to prioritize relevant content and enable efficient communication to tackle information overload [14]. Dynamic faceted filters have shown promise in alleviating information overload in previous studies [22]. Additionally, recommendation systems that filter based on specific requirements have demonstrated their effectiveness in commercial applications [14,31].

Presently, college students encounter difficulties accessing and screening information about graduate programs. The available channels, such as forums and official school websites, are inefficient and require manual screening and comparison. Recommendation systems offer a promising solution to this challenge, providing efficient and accurate information filtering, thus reducing the time and effort required by students [28].

Our recommendation system employs KNN-based Collaborative Filtering (CF) to create personalized recommendation lists for users with similar backgrounds. CF is a highly effective and popular algorithm for recommendation systems, known for its robustness and efficiency [16]. By using KNN in Collaborative Filtering, we address potential personalization issues and generate more reliable recommendation lists by considering multiple similar cases together [5,23]. To handle large volumes of data, we deploy KNN-based CF on Hadoop, significantly improving the system's performance for handling substantial data.

This paper proposes an intelligent fine-grained recommendation system for grad-school application, the contributions are:

- Tanimoto Coefficient Similarity is used in user similarity calculation to focus the similarity on the correlation relationship between users and items, and reduce the focus on specific ratings.
- A recommendation system that focuses on both users' features and interest preferences is proposed, which is more suitable for graduate school recommendation scenarios than the past school recommendation system that only focuses on interest preferences.
- The experimental results verify that the recommendation system focuses on user features while also playing a sizable role in the recommendation of interest preferences.



KNN-Based Collaborative Filtering for Fine-Grained Intelligent Grad-School Recommendation System

Jinfeng Xu[1], Jiyi Liu[2], Zixiao Ma[3], Yuyang Wang[4], Wei Wang[5], and Edith Ngai[1(✉)]

[1] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong 999077, China
chngai@eee.hku.hk

[2] School of Software Engineering, Sun Yat-Sen University, Zhuhai 528406, Guangdong, China
liujy557@mail2.sysu.edu.cn

[3] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
zixiao.ma@duke.edu

[4] Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

[5] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Artificial Intelligence Research Institute, Shenzhen MSU-BITUniversity, Shenzhen 518172, Guangdong, China
ehomewang@ieee.org

Abstract. The development of the Internet has led to information overload, and how to filter and sift information is a rigorous requirement in all fields. In response to this challenge, recommendation systems have emerged as a valuable tool, offering personalized content and services by efficiently searching and processing dynamically generated information. For students applying to grad schools, finding relevant information can be time-consuming and unreliable from official websites or forums. In light of these challenges, we present a novel solution in the form of an application recommendation platform. Our proposed platform leverages specific open-source datasets and real-time information from platform users using KNN (K-Nearest Neighbor) and CF (Collaborative Filtering) techniques to provide recommendations based on users' individual backgrounds, we aim to reduce the complexity inherent in information retrieval while simultaneously enhancing the relevance of the recommendations delivered to users. Specifically, we first collect user behavior data, then we will construct the data model and perform some preprocessing on it. Calculate the user similarity, and find out the K-nearest neighbors and rate based on K-nearest neighbors, finally, the recommendation engine is used to calculate the highest-rated items to be recommended to the users.

Keywords: Recommendation System · Collaborative Filtering · K-Nearest Neighbor

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 494–508, 2024.
https://doi.org/10.1007/978-3-031-65126-7_41

3. Burke, E.K., Smith, A.J.: Hybrid evolutionary techniques for the maintenance scheduling problem. J. IEEE Trans. Power Syst. **15**(1), 122–128 (2000)
4. Kirpatrick, S, Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. In: Readings in Computer Vision, pp. 606–615 (1987)
5. Hoiles, W., Schaar, M.V.D.: Bounded off-policy evaluation with missing data for course recommendation and curriculum design. In: The 33rd ICML, New York, pp. 1596–1604 (2016)
6. Hossain, S.I., Akhand, M.A.H., Shuvo, M.I.R., et al.: Optimization of university course scheduling problem using particle swarm optimization with selective search. J. Expert Syst. Appl. **127**(9), 24 (2019)
7. Tripathy, A.: Computerized decision aid for timetabling—a case analysis. J. Discret. Appl. Math. **35**(3), 313–323 (1992)
8. Ferland, A.J.A.: A large scale timetabling problem. J. Comput. Oper. Res. **16**(1), 67–77 (1989)
9. Lin, Z., Lin, Y.: The application of techniques of artificial intelligence to timetable scheduling. J. Tsinghua Univ. **24**(2), 1–9 (1984)
10. Nengbin, W.: UTSS – a university timetable scheduling system. Chin. J. Comput. (1984)

Table 1. A Class schedule.

Weekly/session	Monday	Tuesday	Wednesday	Thursday	Friday
The first class (Lesson 1–2) 08:00–09:50	Operating system 6-310 Teacher 10	Software architecture and framework 6-408 Teacher 12	Principles of Compiler Construction 6-301 Teacher 8	Software testing technique 6-310 Teacher 15	Principles and applications of microcomputer 6-401 Teacher 9
The second class (Lesson 3–4) 10:10–12:00	Principles of Compiler Construction 6-301 Teacher 8	Software testing technique 6-310 Teacher 15	Operating system 6-310 Teacher 10	Software project management 6-412 Teacher 17	Principles of Compiler Construction 6-301 Teacher 8
The third class (Lesson 5–6) 14:10–16:00	Principles and applications of microcomputer 6-401 Teacher 9	Software project management 6-412 Teacher 17	Principles and applications of microcomputer 6-401 Teacher 9		Software architecture and framework 6-408 Teacher 12
The fourth class (Lesson 7–8) 16:10–18:00					

6 Conclusion

On the basis of studying the problem of college course scheduling, this paper establishes a mathematical model of college course scheduling, and uses improved genetic algorithm to solve the problem of college course scheduling. In this paper, a time-class "course matrix" and a "classroom matrix" are taken as an individual. In the "course matrix", a class-course-teacher "teaching event" is stored, which is represented by the teaching task number. Meanwhile, another unique time-class matrix is used to store the classroom number. The genetic variation operation is carried out on N individuals of the "course matrix", the fitness function is used to select, and the optimal solution is obtained after many iterations. The experiment verifies that the improved genetic course scheduling algorithm proposed in this paper can effectively solve the problem of efficient course scheduling.

References

1. Gotlieb, C.C.: The construction of class-teacher time-tables. J. Commun. ACM **5**(6), 73–77 (1962)
2. Even, S., Itai, A., Shamir, A.: On the complexity of timetable and multicommodity flow problems. SIAM J. Comput. **5**(4), 691–703 (1976)

5 Experimental Results

In this paper, there are 220 classes, 300 classrooms, 700 teachers and 550 courses in four grades in a college. There are 20 classes per week, and each time period is 2 h. The initial population was set to 50. The traditional genetic algorithm and the improved genetic algorithm were used to iteratively evolve 1000 generations, and 10 experiments were conducted. The average value of the optimal individual fitness value was obtained for every 100 generations of evolution. The experimental comparison results were shown in the Fig. 1. The average fitness value of the improved genetic algorithm is higher than that of the traditional genetic algorithm, and the convergence is good. The course scheduling algorithm in this paper is effective in solving the course scheduling problem.

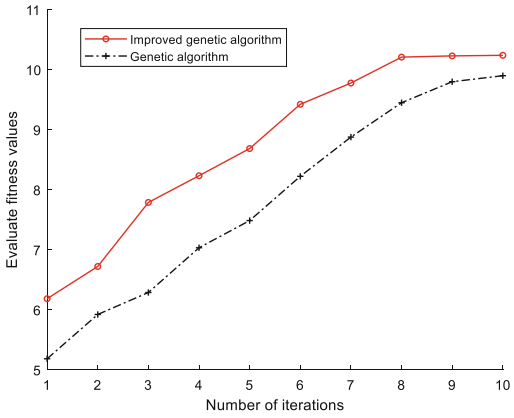


Fig. 1. Optimal course schedule

As shown in the Table 1, basic professional courses are usually arranged in the morning, there are two classes a week with reasonable intervals, and the classes held three times a week are distributed on Monday, Wednesday and Friday. In short, the course distribution is even and reasonable.

function score, which is

$$\overline{c} = \frac{1}{wday} \sum_{i=1}^{wday} c_i \tag{7}$$

$$\sigma_c = \sqrt{\frac{\sum_{i=1}^{wday} (c_i - \overline{c})^2}{wday}} \tag{8}$$

σ_c is the standard deviation of the dispersion of an individual course, c_i is the sum of the number of courses on the i day, $\overline{c}$ is the average of the number of courses on the day, $wday$ corresponds to the sum of the working days of a week, which in this paper refers to Monday to Friday, so $wday = 5$.

For example, when a class has 5 classes a week, if the 5 classes are evenly distributed between Monday and Friday, $\overline{c} = 1, \sigma_c = 0$, which means that the class schedule is evenly distributed; if 4 classes are scheduled on Monday and 1 class is scheduled on Tuesday, $\overline{c} = 1$, $\sigma_c \approx 1.55$, which means that the class schedule is overcrowded. Therefore, the smaller the standard deviation, the higher the excellence of the individual, so a fixed value can be subtracted from the standard deviation and then amplified, that is, the score of this curriculum under this soft constraint condition.

Algorithm 8: Course Interval Weight Algorithm

Input: pi, class schedule individual ;ri ,Classroom individual ;Class_num ,Number of classes;

Output: result ,Evaluation value

```
result = [0,0,...];
for col  1:Class_num
    s=0;
    taskno=pi(:,col);       #Read the teaching task number of each class
    timeno=find(taskno);#Read the non-0 teaching task number
    ut=unique(taskno(timeno)); #ut is a non-repetitive teaching task number
    m=length(ut); #m is the number of classes per week per teaching task number
    for   k1=1:m
            tib=find(ut(k1)==taskno);#tib is the number of classes per week
            if length(tib)==1          #The course meets once a week
                Add the corresponding weights to s
             else if length(tib)==2#The course meets two times a week
                Add the corresponding weights to s
            else if length(tib)==3#The course meets three times a week
                Add the corresponding weights to s
            end if
    end for
   result[col]=result[col]+s;
end for
```

in the schedule are summed. Finally, this value is used as the grading index of the class schedule under the constraint conditions.

$$S_1(x) = \frac{1}{n}\sum_{i=1}^{n} a_m, \quad m \Rightarrow x_i \tag{5}$$

$S_1(x)$ is the evaluation function of the soft constraint condition, and x is a definite class schedule, $m \Rightarrow x_i$ representing the period m corresponding to the class i of the class schedule x. That is, the weights corresponding to the time periods of all courses in a certain class schedule are added successively and then divided by the total number, and the average value obtained is the result of the evaluation function.

Algorithm 7: Course preference evaluation algorithm

Input: population ,class schedule individual ;roomi, Classroom individual ;Class_num ,Number of classes ; T, Time period ;Ce ,Curriculum expectation

Output: result ,Class schedule evaluation value

```
result = [0,0,...];
for col=1:Class_num
    s=0;
    for row=1:T
        if populationi[row][col]!=0
            s+=Ce[row];
        end if
    end for
end for
result. append(s);
```

Course Interval Weight

For the same course, it should be reasonably distributed within a week to avoid crowded or distant courses.

$$S_2(x) = \frac{1}{snumber_x}\sum_{i=1}^{snumber} b_s \tag{6}$$

$S_2(x)$ is the evaluation function of the soft constraint condition, x is a definite curriculum schedule, *snumber* is all the courses in the curriculum schedule, and note that the courses that appear multiple times are counted as one course, which is counted according to the course number. The value of the evaluation function is calculated by adding the weights of all courses according to the contents of the table and then calculating the mean value.

In order to measure the uniformity of the weekly course distribution, the phenomenon of course crowding on several days of the week is avoided. In this paper, the standard deviation of course arrangement is used to judge the degree of dispersion of courses. The higher the standard deviation, the more crowded the courses, the lower the fitness

4.3 The Fitness Function

The fitness function plays a key role in evaluating the strengths and weaknesses of individuals in a population. According to the soft constraints related to the course scheduling problem in colleges and universities, this paper adopts the weighted method to construct the fitness function to ensure the balance performance of the algorithm under various conditions. The fitness function F(s) is designed as follows:

$$F(x_i) = \sum_{i=1}^{n} \alpha_i s_i \tag{3}$$

where, x_i represents an individual (that is, a course scheduling solution), n represents the number of soft constraints and the weight coefficient of a certain soft constraint in the fitness function, reflecting the relative importance of the condition in the optimization process. It represents the score of the individual under the soft constraint condition, which measures the performance of the individual in meeting the soft constraint condition, and s should also satisfy the following constraints:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i = 1 \\ s \in [0, 100] \end{cases} \tag{4}$$

In practical application, we can adjust the weight coefficient α and the scoring method according to the characteristics of the problem and the actual demand, so as to better meet the goal of course scheduling optimization. The weights assigned to the constraints in this paper are 0.4, 0.4, 0.1 and 0.1 respectively.

The evaluation function of soft constraints is defined as follows:

Algorithm 6: Selection Strategy

Input: p, Parent class schedule population; r, Parent classroom population ;p_new, Offspring class schedule population ; r_new, Offspring classroom population ;F(),Evaluation function

Output: p_new and r_new

```
p =p∪ p_new;
index=Sort(p, F(p));
index= 1;
for i in index
    p_new[i]=p[i];
    r_new[i]=r[i];
end for
```

Course Preference Evaluation

Courses in a day should be arranged as far as possible at a time with a high expectation of the time period.

This article assigns class Time Expectations to different time periods. Morning time slots are assigned higher course expectations, while afternoon time slots are assigned lower course expectations. In order to evaluate the degree to which a schedule satisfies this constraint, the expected values corresponding to the time periods of all the courses

Algorithm 4: The crossover operation

Input: population ,Class scheduling individual ;roomi, Classroom individual ;Class_num, Number of classes; T, Time period; CR, cross factor;

Output: populationi and roomi

```
for col=1:class_num
    for row=1:T
        mask = rand(0,1);
        r = rand(1, 20)and r≠row;
        if mask < CR then
            Swaprow(populationi, row, r, col);
            Swaprow(roomi, row, r, col);
        end if
    end for
end for
Conflict handling for populationi; #Algorithm 2
Conflict handling for roomi; #Algorithm 3
```

The Mutation Optimization

We choose F with a value of 0.5, and when accessing an individual, when traversing the columns, randomly generate a number rand data between [0, 1]. If randnum $<$ F, the current column is swapped with the same position column of a randomly selected individual; If randnum $\geq$ F, no exchange operation is performed. At the same time, the same position of the classroom matrix room is exchanged to maintain the consistency of the course arrangement and the classroom arrangement.

Algorithm 5: The mutation operation

Input: N, population size; F, mutation operator, Class_num, number of class; p, Parent class schedule population; r, Parent classroom population;

Output: p_new, Offspring Class schedule population ;r_new, Offspring classroom population

```
for i=1:N
    for col=1:class_num
        p_new = rand(p) % N;
        while p_new == i
            p_new = rand() % N;
        end while
        randnum = rand(0,1);
        if randnum < F then
            swap(p[i][j], p_new[i][j]);
            swap(r[i][j], r_new[i][j]);
        end if
    end for
end for
```

that the capacity of the new classroom number is greater than or equal to the number of teachers.

Algorithm 2: Teacher Teaching Conflict Management

Input: N ,population size; Class_num, number of class; T, Time period; population, Class schedule population; room, classroom population

Output: population and room

```
flag=1; #If flag is equal to 1, there is a teaching task number conflict. Otherwise, no conflict
while flag==1
  flag=0;
  for i=1:N
    for row=1:T
      If a teacher repeats the class in the same period, adjust the individual population's teaching
      time and adjust the Classroom number of individual room at the same time;
      flag=1;
    end for
  end for
end while
```

Algorithm 3: Classroom Conflict Handling Algorithm

Input: N, population size; Class_num ,number of class; T, Time period; room, classroom population

Output: room

```
flag=1;#If flag is equal to 1, there is a classroom number conflict. Otherwise, no conflict
while flag==1
  flag=0;
  for i=1:N
    for row=1:T
      Find if the row has the same classroom number, if there is a duplicate classroom number,
      the next number is reassigned to a new classroom number;
      flag=1;
    end for
  end for
end while
```

The Crossover Operation

The value of CR is 0.2. When col column of the course individual is accessed, the current row is row, the random value mask [0, 1] is set, the random number [1, 20] is taken and $i \neq$ row is taken. If mask $<$ CR, Then, the course individuals population (row, col) and population (i, col) are exchanged, and the above exchange operation is performed for each data in the col column in turn. After the individual mutation is completed, the rows of population and room matrix are re-processed for conflict.

inserted into the matrix, which is called the "curriculum matrix". At the same time, another unique time-class matrix is used to store the classroom numbers, which is called "classroom matrix" in this paper. The two matrices above are called an individual in a population. Since the above matrix can only store the class-course-teacher arrangement, which is called the "classroom matrix" in this paper, the matrix does not contain the time-classroom arrangement, so another matrix is needed to store the time-classroom information, and the position relationship in the matrix corresponds to the course matrix.

In the optimization process of course scheduling by genetic algorithm, the concepts of Population and Individual need to be introduced, and the individual is a class schedule, and the population is a collection of multiple classes with different contents and the same structure. By definition, the curriculum matrix and classroom matrix correspond to individuals in the genetic algorithm, and the matrix set corresponds to the population. (hereinafter referred to as a class of matrix schedule individual population1, population2, … Populationi), a classroom scheduling matrix called classroom individuals (room1, room2, … Roomi), the set of course matrices is called the class schedule population, and the set of classroom scheduling matrices is called the room population.

Algorithm 1: Framework of the course scheduling algorithm.

Input: N, population size; CR, cross factor; F, mutation operator, GEN ,The maximum number of generations ;
Class_num, number of class; T, Time period; teaching_task_data, Teaching task information;
Room_data, Classroom information

Output: A series of best solutions

```
Initialize population ;
while the termination conditions are not met do
    Teacher teaching conflict handling;#Algorithm 2
    Classroom conflict handling;#Algorithm 3
    Mutation;#Algorithm 4
    Crossover;#Algorithm 5
    Selection Strategy;#Algorithm 6
end while
```

4.2 Teacher Teaching Conflict Management and Classroom Conflict

In the process of teacher teaching conflict processing, each row of the individual population is read to find the teacher number corresponding to each teaching task number in the row, and the teacher number is read successively. When there is a duplicate teacher number, the teaching task number corresponding to the duplicate teacher number is exchanged with the teaching task number randomly selected in this column, and the classroom number corresponding to the two teaching task numbers is exchanged at the same time.

In the classroom conflict processing, each line of the individual room is read to find whether there is the same classroom number in the row. If there is a duplicate classroom number, the new classroom number is re-assigned to the following number to ensure

(3) The teaching of each course should be arranged in the same classroom every week;
(4) The number of students in the classroom should be as close as possible to the capacity of the classroom.

Mathematical Models of Constraints

This article will class - class, the teacher Teaching events (would Event), the Teaching events according to the Teaching schedule. $TE = \{(c_a, s_b, t_c\}) \mid c \in C, s \in S, t \in T, a \in cnum, b \in snum, c \in tnum$, (c,s,t) satisfies the "teaching arrangement"}. Will constitute the classroom - time corresponding space-time events (Spatiotemporal Event), namely, $SE = R \times M = (r_1, m_1), (r_1, m_1), (r_1, m_1), (r_1, m_1), \ldots, (r_{rnum}, m_{mnum})$. Cnum is the total number of classes, gnum is the total number of colleges, rnum is the number of classrooms, tnum is the number of teachers, snum is the total number of courses, mnum is the total number of time ranges.

The space-time event is set as a random queue, and when a teaching event needs to be processed, a teacher-time is selected from the space-time event as the class time and place, that is, $TE \cap SE = 1$ should be guaranteed, and if there is no such correspondence, $TE \cap SE = 0$.

The hard constraints include classroom capacity, teacher conflict, curriculum conflict, etc., while the soft constraints include course time allocation, etc. For these constraints, the corresponding mathematical model is established:

(1) Each teacher may only teach in one classroom during a specified period of time.

$$\forall t \in T, m \in M, |\{r \in R : (t, m, r)\}| \leq 1 \tag{1}$$

For any teacher t belongs to set T, and any time period m belongs to set M, that is, given teacher t and time period m, the number of times that teacher t teaches in different classrooms r cannot exceed 1, that is, each teacher can only teach in one classroom within a time period.

(2) Teachers may only teach courses related to their respective faculties.

$$\forall t \in T, s \in S : \text{ Teaching } s \Rightarrow G(t) \in G(s) \tag{2}$$

For any teacher t to belong to set T, and any course s to belong to set S, if a teacher t teaches course s (represented by Professor t s), then the faculty to which the teacher t belongs must be in the set G(s) corresponding to the course s. This means that teachers can only teach courses related to their school.

4 Improve the Genetic Algorithm of College Course Scheduling Algorithm

4.1 Improvement Genetic Algorithms Course Scheduling Problem

This article sets up a matrix where each row represents a time period and each column represents a class. The class-course-teacher combination is regarded as a "teaching event", and each teaching event is represented by a teaching task number, which is

3 Mathematical Model of Scheduling Problem

The Various Entities That Appear in the Scheduling Problem

(1) Class: In college education, classes of different majors may be arranged to take the same course.
(2) College: A university college is a teaching and research institution within a university divided by discipline and specialty. Teachers, courses and classes are all affiliated to the college.
(3) Room: There are many types of classrooms, such as multimedia classrooms, speech classrooms and outdoor venues. The description of the classroom must include the classroom name, building, classroom capacity, and classroom type.
(4) Teacher: A teacher is a lecturer who imparts knowledge to students. The description of teachers includes teacher number, teacher rank, teacher name, teacher gender and other factors.
(5) Task: A course is a series of teaching tasks arranged by the school. Courses have attributes such as course number, course name, teaching period, and school and course type.
(6) Teaching Schedule: Teaching schedule is one of the core elements of course scheduling, which involves the course arrangement required for all classes in a semester, including course name, course number, starting college, grade major and other information.
(7) Time period: The class time is divided into a period according to certain rules. For example, 8:00–10:00 on Monday is the first period, and 16:00–18:00 on Friday is the 20th period.

Hard Constraints That Must be met for Scheduling Problems

(1) A teacher can only arrange one course at a time;
(2) A student can only take one course at a time;
(3) A classroom can only arrange one course at a time;
(4) There is only one type of classroom required for a course;
(5) Must conform to the school's teaching plan, and complete a certain class hour and examination or examination within the specified time.
(6) The number of students in the classroom should be less than the capacity of the classroom.

Soft Constraints

(1) Each course should be arranged at appropriate intervals during the week;
(2) The courses of high difficulty will be arranged in a better time period, and the courses of low difficulty will be arranged in other time periods. The physical education class will be arranged in 3–4 classes every day, or in the afternoon, and there will be no other classes after that;

The core idea of genetic algorithm is to realize efficient search of solution space by simulating natural selection, crossover and variation in the process of biological evolution. Different from traditional numerical optimization methods, genetic algorithm does not depend on the derivative property of the problem and the continuity of the function, so it has higher robustness and applicability. In genetic algorithms, each solution in the solution space is encoded as a chromosome and then evaluated by a fitness function. Based on these evaluation results, the algorithm performs selection, crossover and mutation operations in the solution space to generate a new generation of solutions, and iterates constantly to gradually approach the global optimal solution.

In genetic algorithms, the initial solution population consists of n binary strings, called chromosomes. The binary bits in each chromosome represent genes, which together describe the characteristics of a solution. Genetic algorithm consists of three main operations: selection, crossover and mutation.

2 Preliminaries

In 1963, Gotlibe proposed a mathematical model for the problem of class scheduling [1]. In 1976, S. Even proved for the first time that the scheduling problem is NP-complete [2]. Since scientists do not have an algorithm for solving NP-complete problems, they focus on specific practical problems. In 2000, E. K. Burke and A. J. Smith proposed a hybrid meme algorithm combining local search operators with genetic algorithm, which can better solve scheduling problems such as course scheduling [3]. In 1983, Kirkpatrick et al. proposed the simulated annealing method [4]. Although it has achieved certain results, the simulated annealing algorithm has some shortcomings, such as slow convergence speed, long execution time, algorithm performance is related to the initial value and parameter sensitivity. William Hoiles and Mihaela van der Schaar proposed a personalized course scheduling algorithm based on UCB in 2016, using recorded student data to generate featured course recommendations and course plans [5]. In 2019, Imran Hossain proposed an optimization method for university course scheduling based on particle tour [6]. Indian scholar Arabinda Tripathy proposes to arrange classes based on people and adopt the method of multiple class groups to deal with conflicts [7]. Canadian scholars Jean Aubin, Jacques A. Ferland, Charles Fleutent and others divided the course scheduling problem into schedule problem and grouping problem. Then the SAPHIR course scheduling decision support system, which includes several modules such as data processing, automatic optimization and interactive optimization, is developed [8]. In 1984, Lin Zhang xi and Lin Yao rui published the experimental research result “Application of Artificial Intelligence Technology in Curriculum Scheduling” [9]. They used artificial intelligence technology to conduct heuristic search for and or graphs generated in the curriculum scheduling, and successfully arranged the curriculum according to the two sets of data. Nanjing Institute of Technology developed UTSS (A University Timetable Scheduling System) [10]. The course scheduling system of these universities is designed for the curriculum and teaching resources of the university, and cannot be widely promoted. There are many algorithms to realize the scheduling problem, and there are roughly the following types of algorithms: integer programming algorithm, graph theory based algorithm, heuristic algorithm, genetic algorithm, simulated annealing method and ant colony algorithm.

An Improved Genetic Algorithm for College Course Scheduling

Chenle Wang[1(✉)] and Bin Wang[2]

[1] International Engineering College, Xi'an University of Technology, Xi'an 710048, China
3202241003@stu.xaut.edu.cn

[2] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
wb@xaut.edu.cn

Abstract. The course scheduling is a NP-complete problem. At present, various intelligent optimization algorithms have provided many feasible solutions to the course scheduling problem in colleges and universities, with differentiated advantages and disadvantages. This work tries to design a general and efficient algorithm to solve the large-scale course scheduling. In particular, we analyze and model the problem of course scheduling in colleges and universities, and put forward the improved genetic algorithm to solve the problem of large course scheduling under various constraints. First, the teaching task number is stored in a two-dimensional time-class matrix, which represents the information of teachers, and a two-dimensional classroom matrix is established to store the classroom information of the class. Next, we apply the improved genetic algorithm to cross and mutate the time-class matrix to obtain new individuals, thereby adjusting the corresponding classroom matrix. Then, the excellent individual is selected from the parent and child generation, and iterated until the optimal individual is produced. Finally, experimental results show that the convergence speed of the proposed algorithm is faster and higher fitness values can be obtained.

Keywords: Genetic Algorithm · Scheduling Problem · Classroom Matrix

1 Introduction

Genetic Algorithm (GA) is a randomized, efficient and adaptive search method inspired by the biological genetic mechanisms of survival of the fittest and survival of the fittest. Since the method was proposed by Professor J. Holland of the University of Michigan in 1975, it has been widely used in many fields, such as optimization problem solving, machine learning, data mining, artificial neural network training and intelligent control system, etc., because of its powerful global search and adaptive adjustment ability. Genetic algorithm is regarded as one of the key technologies in modern intelligent computing.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 481–493, 2024.
https://doi.org/10.1007/978-3-031-65126-7_40

References

1. Kreutz, D., Ramos, F.M.V., Verissimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-defined networking: a comprehensive survey. Proc. IEEE **103**, 14–76 (2015)
2. Abdelaziz, A., et al.: Distributed [3] controller clustering in software defined networks. PLoS ONE **12**, e0174715 (2017)
3. Akbar Neghabi, A., Jafari Navimipour, N., Hosseinzadeh, M., Rezaee, A.: Nature-inspired meta-heuristic algorithms for solving the load balancing problem in the software-defined network. Int. J. Commun. Syst. **32**, e3875 (2019)
4. Albowarab, M.H., Zakaria, N.A., Abidin, Z.Z.: Directionally-enhanced binary multi-objective particle swarm optimisation for load balancing in software defined networks. Sensors **21**(10), 3356 (2021)
5. Hamdan, M., et al.: A comprehensive survey of load balancing techniques in software-defined network. J. Netw. Comput. Appl. **174**, 102856 (2020)
6. Praveen, S.P., Sarala, P., Kumar, T.N.S.K.M., Manuri, S.G., Srinivas, V.S., Swapna, D.: An adaptive load balancing technique for multi SDN controllers. In: 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, pp. 1403–1409 (2022)
7. Sridevi, K., Saifulla, M.A.: LBABC: distributed controller load balancing using artificial bee colony optimization in an SDN. Peer-to-Peer Netw. Appl. **16**, 947–957 (2023)
8. Chiesa, M., Kindler, G., Schapira, M.: Traffic engineering with equal-cost-multipath: an algorithmic perspective. IEEE/ACM Trans. Netw. **25**(2), 779–792 (2017)
9. Tanha, M., Sajjadi, D., Ruby, R., Pan, J.: Traffic engineering enhancement by progressive migration to SDN. IEEE Commun. Lett. **22**(3), 438–441 (2018)
10. Novaes, M.P., Carvalho, L.F., Lloret, J., Proenca, M.: Long short-term memory and fuzzy logic for anomaly detection and mitigation in software-defined network environment. IEEE Access **8**, 83765–83781 (2020)
11. Mao, B., Tang, F., Fadlullah, Z.M., Kato, N.: An intelligent route computation approach based on real-time deep learning strategy for software defined communication systems. IEEE Trans. Emerg. Top. Comput. **9**(3), 1554–1565 (2021)
12. Xue, H., Kim, K.T., Youn, H.Y.: Dynamic load balancing of software-defined networking based on genetic-ant colony optimization. Sensors **19**, 311 (2019)
13. Zhu, S., Long, Y., Sun, G., Li, C.: Improved ant colony algorithm for network flow scheduling in SDN data center. J. Harbin Univ. Sci. Technol. **001**, 1–7 (2022)
14. Jamali, S., Badirzadeh, A., Siapoush, M.S.: On the use of the genetic programming for balanced load distribution in softwaredefined networks. Digit. Commun. Netw. **5**, 288–296 (2019)
15. He, X., Ren, Z., Shi, C., Fang, J.: A novel load balancing strategy of software-defined cloud/fog networking in the Internet of Vehicles. China Commun. **13**(Suppl. 2), 140–149 (2016)
16. Belgaum, M.R., Ali, F., Alansari, Z., et al.: Artificial intelligence based reliable load balancing framework in software-defined networks. Comput. Mater. Continuum **70**(1), 251–266 (2022). (in English)
17. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, vol. 4, no. 2 (1995)
18. Zhang, H., Lin, K.-Y., Huang, H.: Multi-objective scheduling model for OpenStack based cloud. In: 2021 8th International Conference on Computational Science/Intelligence and Applied Informatics (CSII) (2021)

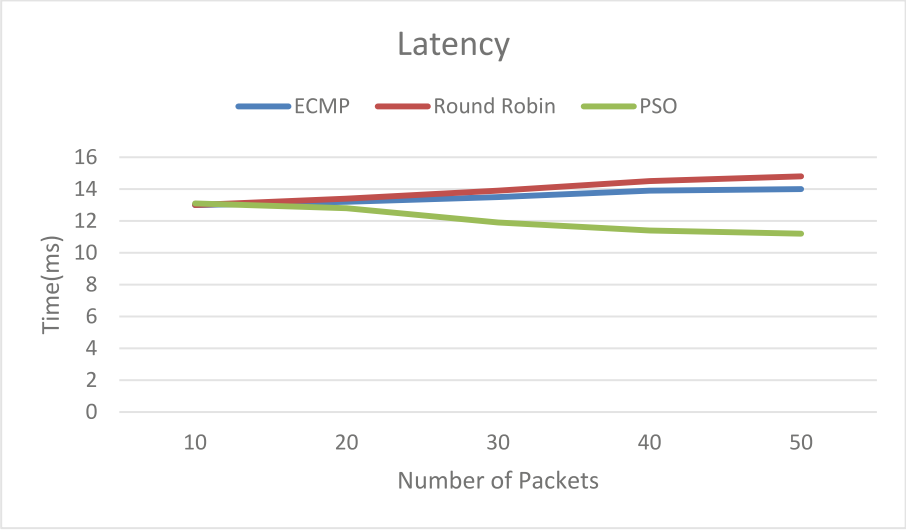


Fig. 2. Comparison of latency.

Table 5. Packet loss ratio

Algorithm	Value
ECMP	0.021
Round Robin	0.025
PSO	0.016

6 Summary

In this paper, we optimized load balancing in software defined networking (SDN) using particle swarm optimization (PSO). We evaluated the performance of PSO and compared with terms of bisection bandwidth, latency, and loss ratio of packet. The results indicate that particle swarm optimization algorithm can dynamically model based on load changes. However, reliability issues have not been addressed in this algorithm, and assuming the performance of all devices is the same, but in actual networks, the performance of devices is different.

In our future work, PSO algorithm must be appropriately improved, and evaluate load balancing while considering real-time networks, such as TCP and UDP.

Fund Project. Zhuhai College of Beijing Institute of Technology 2022 School-level Course (compute network) Beijing Institute of Technology Zhuhai College 2023 School -level Innovation and Entrepreneurship Training Project (Research and Implementation of Load Balancing and Energy-saving Technologies in SDN).

5.2 Results and Analysis

In this study, we compared it with two commonly used load balancing algorithms: loop algorithm and ECMP algorithm. Extract the results of each load balancing algorithm in the following order:

1) In Mininet, create a custom topology.
2) Configurate of all hosts, switches and links.
3) Initialize of the network.
4) Settings for load balancing algorithm in the controller.
5) Transmit hosts requests over the network.
6) Evaluate the performance of load balancing algorithm.

For the average bisection bandwidth, the comparison of experimental results is presented in Table 4. It's evident from the table that PSO algorithm is higher than ECMP and Round Robin in terms of average bisection bandwidth.

Table 4. Average Bisection Bandwidth (Mbps)

Algorithm	Value
ECMP	76
Round Robin	77
PSO	89

The delay results of using different algorithms to transmit different numbers of data packets are shown in Fig. 2. By comparing the results, it can be seen that compared with ECMP and round-robin technology, PSO has lower latency. This is caused by the fast convergence characteristic of the particle swarm algorithm towards the global solution.

Table 5 presents the average packet loss ratio for various techniques. It is evident that the round-robin technique exhibits a higher average packet loss ratio when compared to both ECMP and PSO.

In summary, the PSO algorithm has significant advantages in terms of average bandwidth, latency, and packet loss rate. Among them, the ECMP algorithm does not consider the current network state and only schedules traffic through hashing, it can easily lead to network congestion. Round robin also belongs to stateless scheduling. In this experiment, due to the same performance of each switch, its average bandwidth is better than ECMP. However, in actual networks, the performance load of each switch is different, which can easily result in load imbalance.

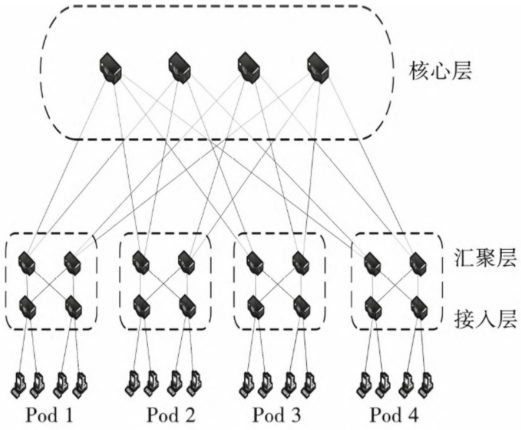


Fig. 1. Fat-Tree network topology

Table 2. Network environment parameters

Parameter	Description
Link Bandwidth	10 Mbps
Switch maximum buffer queue	1000
Traffic timeout	5 s

In the experiment, virtual hosts in the network communicate randomly with equal probability, and the parameter settings of the PSO algorithm directly affect its performance. The parameters of the algorithm have been determined through the experiment as shown in Table 3.

Table 3. PSO algorithm parameters

Parameter	Description
Number of particles	20
Number of iterations	100
w	0.5
c1	1/3
c2	2/3

5 Experimental Results and Analysis

5.1 Experimental Environment Configuration

Mininet, a network simulation tool developed by Stanford University, is a lightweight software-defined platform that supports various southbound protocols like OpenSwitch and OpenFlow. It allows you to build entire networks, consisting of switches, links, and hosts, on a single physical device.

This article describes the use of a Python-based Ryu controller to build an SDN network through Mininet. The experiment is conducted on a HP laptop running a VirtualBox virtual machine with Ubuntu 18.04 as the operating system. Scapy is utilized to inject traffic into the virtual network, simulating network traffic. The simulation parameters for this experimental environment are presented in Table 1.

Table 1. Parameters

Parameter	Description
OS	Ubuntu 18.04
CPU	Ryzen 5-3550
RAM	16G
Simulator Tool	Mininet 2.3.1
Mininet 2.3.1	RYU 4.30
Protocol	OpenFlow V1.3
Traffic generator	Scapy

To validate the performance of the PSO algorithm, the average split bandwidth and end-to-end delay were used as performance indicators. Experimental tests were conducted on the Fat-Tree network topology (Fig. 1), and the PSO algorithm was compared with ECMP and Round Robin.

As depicted in Fig. 1, comprises of 20 switches and 16 hosts interconnected in a random configuration. All these hosts and switches are located within the infrastructure layer, and the controllers are located within the control layer. Any source host communicates with the target host by connecting their switches. If the source host and the destination host in the flow table are not connected, send a packet to the controller_In message. The controller uses the PSO algorithm to determine the optimal and least loaded path, and injects flow entries along that path into the switch to forward user requests. The specific parameters used in the network simulation environment are outlined in Table 2.

Assuming the source host is H1 from pod 1, the destination host is H8 from pod 2, and the data packet is sent from the source to the destination. It is obvious that there are four ways to reach the destination.

(4)–(6) define a traffic conservation constraint, which means that the traffic of f_k from the source host to the destination host is equal to the traffic of any node in the path.

When forwarding traffic, map the model to the particle swarm optimization algorithm to obtain a specific scheduling plan. When the network traffic is at the peak, in order to avoid uneven link responsibility caused by network congestion, a load balance degree of the whole network is set to determine whether the current network is in the load balance state. This article cites the concept of standard deviation in mathematics and uses discrete programs to reflect equilibrium. At the same time, to avoid unnecessary rerouting caused by short-term extreme values, the mean over a period of time is used as the result. The specific definition of load balancing is as follows:

$$\delta(t) = \frac{1}{P}\sum_{t=T-P}^{T}\left(\sqrt{\frac{1}{m}\sum_{l=1}^{m}\left(\overline{load(t)} - load_l(t)\right)^2}\right) \tag{7}$$

Among them, P represents the statistical period, T represents the current time, $\overline{load(t)}$ represents the average load of all d-rated links at time t, and $load_l(t)$ represents the real-time load of link l at time t.

4.2 Particle Swarm Optimization

PSO is mainly used to solve optimization problems and is one of the meta heuristic algorithms based on natural heuristic proposed by Kennedy and Eberhart. In particle swarm optimization, candidate solutions (usually referred to as particles) are continuously refined to seek the optimal solution. This refinement is achieved by adjusting the motion of particles, which are influenced by their understanding of the most determined local position in the search space.

At the initial state, all swarm particles are randomly positioned in a multidimensional search space. The space's dimensions of which depends on the problem being addressed. The particles navigate through the space by adjusting their velocities and subsequently changing their positions. Each particle possesses its own current position and velocity attributes. The location of a particle represents a potential solution to the problem. To update their positions, particles utilize controlled rules to adjust their velocities, enabling them to move toward better locations.

Within PSO, every particle maintains two values: (1) its personal best position (localBest) and (2) the overall best position achieved by the entire group (globalBest). These values are utilized for velocity updates, as illustrated in Eqs. (8) and (9) [17, 18]. Here, C_1 and C_2 are positive numbers known as acceleration coefficients, while r_{1d} and r_{2d} denote two uniformly distributed numbers within the [0, 1]. P_{id} and P_{gd} prefer to the particle's localBest and globalBest positions, respectively. As the velocity is adjusted, the formula for calculating the position of particles is as follows

$$V_{id}(t+1) = V_{id}(t) + C_1 r_{1d}(P_{id} - X_{id}) + C_2 r_{2d}\left(P_{gd} - X_{id}\right) \tag{8}$$

$$X_{id}(t+1) = X_{id}(t) + V_{id}(t+1) \tag{9}$$

3) The business orchestration technology in SDN architecture can use software to develop load balancing strategies.

4 Method

4.1 Optimization Objective Functions

The essence of the traffic load balancing problem is to select an appropriate path according to the path information to schedule the flow from the source node to the target node. Here, the flow path problem can be abstracted into integer linear programming model for solution. The specific model is as follows:

For the classic Fat-Tree data center network topology, it can be abstracted as a directed graph, where H and S distributions represent the set of host nodes and switch nodes, while E denotes the collection of links. The network contains m links, where the links can be expressed l_{ij}, i, j $\in$ H $\cup$ S. Additionally, define the maximum load of the link as C_{ij}. There are n traffic in the network, and the set of traffic can be represented as $F = \{f_1, f_2, \cdots, f_n\}$, One traffic f_k is defined as a triplet $\left(s^k, d^k, b^k\right)$, where s^k represents the source host node, d^k represents the destination host node, and b^k represents the bandwidth occupied by traffic f_k.

The optimization goal is to maximize the average link bandwidth utilization, which means maximizing the utilization of network data while avoiding network congestion. Therefore, the objective function is:

$$\max\left(\frac{\sum_{\{(i,j)\in E\}}\left(\sum_{1\le k\le n} b_{ij}^k / C_{ij}\right)}{m}\right) \tag{1}$$

Among them, $\sum_{1\le k\le n} b_{ij}^k$ represents the actual load of l_{ij}, b_{ij}^k represents the actual bandwidth of traffic f_k on link l_{ij}. The constraints are as follows:

$$\sum\nolimits_{1\le k\le n} b_{ij}^k \le C_{ij}, \forall(i,j) \in E \tag{2}$$

$$0 \le \forall b_{ij}^k \le C_{ij}, i, j \in H \cup S, k \in \{1, 2, \cdots, n\} \tag{3}$$

$$\sum\nolimits_{\{j:(s^k,j)\in E\}} b_{s^k j}^k - \sum\nolimits_{\{j:(j,s^k)\in E\}} b_{j s^k}^k = b^k, k = 1, 2, \cdots n \tag{4}$$

$$\sum\nolimits_{\{j:(d^k,j)\in E\}} b_{d^k j}^k - \sum\nolimits_{\{j:(j,d^k)\in E\}} b_{j d^k}^k = -b^k, k = 1, 2, \cdots n \tag{5}$$

$$\sum\nolimits_{\{j:(i,j)\in E, i\ne s^k, d^k\}} b_{ij}^k = \sum\nolimits_{\{j:(j,i)\in E, i\ne s^k, d^k\}} b_{ji}^k, k = 1, 2, \cdots n \tag{6}$$

Formula (2) represents the traffic capacity constraint, which means that the total flow on any link should not exceed the link capacity. Formula (3) represents the bandwidth constraint of the traffic, which means that the bandwidth occupied by the traffic f_k should be greater than or equal to 0, cannot be negative, and less than the link capacity. Formulas

The Particle Swarm Optimization (PSO) algorithm has been recognized to be effective for its efficiency in SDN load balancing, which to be compared to several other existing optimization techniques [4].

2 Related Work

Load balancing in Software-Defined Networking (SDN) has become an prominent research topic in recent times. In the experiment of [5], the authors provided a classification method for load balancing in SDN, delving into various objectives, such as response time optimization, overall resource optimization throughout, and identification of bottle necks. Their taxonomy distinguished between load balancing in the control plane and load balancing in the data plane. [6] proposed an approach of multiple SDN controllers for dynamic load balancing. And K. Sridevi [7] uses Artificial Bee Colony for distributed controller load balancing.

There are also many studies on load balancing for traffic. The authors of [8, 9] use conventional approaches such as the OSPF and EMCP algorithm, to optimize the network. But somehow, this static hashing method may cause multiple large streams to be divided into the same path, leading to collisions and causing network congestion. Due to a lack of historical experience, numerous studies have utilized supervised learning methods in SDN for achieving intelligent network management [10, 11]. However, there are significant challenges for supervised learning, because of the reliance on abundant training datasets and the sluggish decision-making process in dynamic network scenarios. Heuristic methods have unique advantages in traffic load balancing, such as a study by [12, 13], that proposed a dynamic load balancing algorithm based on genetic ant colony optimization, and a study by [14] that proposed genetic optimization for traffic loading using SDN. In [15], a modified PSO algorithm was proposed to integrate SDN with fog computing, in order to solve the load balancing issue. Two multi-objective particle swarm optimization methodologies are proposed in [4], which are distance Angle Multi-Objective PSO and Angle Multi-Objective Particle Swarm Optimization.

3 Load Balancing in SDN

The architecture of SDN has shifted the control decision process on the controller, whereas the switch relies entirely on the flow rules injected by the controller to forward data to the device [16]. Transfer the load to a specific switch, and the controller makes decisions based on this.

SDN based load balancing technology has three major advantages:

1) The controller in the SDN architecture can collect global network topology, link available bandwidth, and other resources for learning through the OpenFlow protocol, and develop a better load balancing strategy from a global perspective based on the bandwidth requirements of the data flow.
2) Centralized logical control can abstract all OpenFlow forwarding devices into a whole and be uniformly distributed by the controller to forwarding devices such as switches, without the need for complex configurations of devices with different standards in traditional networks.



Load Balancing in Software-Defined Networks Based on Particle Swarm Optimization

Haiyan Zhang(✉), Liren Zou, and Yilong Xie

School of Computer Technology, Beijing Institute of Technology (Zhuhai), Zhuhai 519085, China
2552974138@qq.com

Abstract. Nowadays, as Software-Defined Networking (SDN) gains prominence, Load Balancing (LB) for SDN assumes significant importance. By allocating network traffic among resources efficiently, LB ensures that no individual resource is burdened, thereby optimizing the overall performance. Based on the analysis of SDN Flow network forwarding mechanism, this paper applies Particle Swarm Optimization (PSO) algorithm to the data center's traffic scheduling based on load balancing. First, the SDN load balancing problem is abstracted as an integer Linear programming model, which maximizes the average link bandwidth utilization on the basis of ensuring network delay. PSO algorithm is applied to optimize the load balancing problem, and the optimization algorithm is run in the SDN controller. The simulation experiment by Mininet shows that the SDN load balancing algorithm based on Particle swarm optimization can effectively balance the network load and improve the network performance.

Keywords: SDN · Load Balancing · Particle Swarm Optimization

1 Introduction

Over the past few years, the development of many new technologies, such as the Internet of Things and cloud systems, has experienced rapid growth. Consequently, it becomes imperative to augment the traditional Internet Protocol network to effectively manage the substantial volume of network traffic.

Software-Defined Networking (SDN) [1] has transformed the traditional approach to network management by dismantling vertical integration of network components, divorcing network logic control from underlying switches and routers, promoting centralization of network control, and incorporating programmability of network operations.

However, SDN has a centralized control architecture, which raises concerns about reliability, fault tolerance, scalability and interoperability [2]. Load Balancing ranks among the most crucial tasks for enhancing network performance, robustness and scalability. Therefore, the combination of multi-objective optimization is essential for identifying an optimization point from a set of non-dominated points. SDN load balancing based on metaheuristic has become a hot research topic [3].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 472–480, 2024.
https://doi.org/10.1007/978-3-031-65126-7_39

21. Mohammadi Farsani, R., Pazouki, E.: A transformer self-attention model for time series forecasting. J. Electr. Comput. Eng. Innov. (JECEI) **9**(1), 1–10 (2020)
22. Mozafar, M.R., Moradi, M.H., Amini, M.H.: A simultaneous approach for optimal allocation of renewable energy sources and electric vehicle charging stations in smart grids based on improved GA-PSO algorithm. Sustain. Urban Areas **32**, 627–637 (2017)
23. Roman, D., Saxena, S., Robu, V., Pecht, M., Flynn, D.: Machine learning pipeline for battery state-of-health estimation. Nat. Mach. Intell. **3**(5), 447–456 (2021)
24. Sun, S., Sun, J., Wang, Z., Zhou, Z., Cai, W.: Prediction of battery soh by CNN-BiLSTM network fused with attention mechanism. Energies **15** (2022)
25. Wang, T., Fang, K., Wei, W., Tian, J., Pan, Y., Li, J.: Microcontroller unit chip temperature fingerprint informed machine learning for IIoT intrusion detection. IEEE Trans. Ind. Inf. **19**(2), 2219–2227 (2022)
26. Xu, H., Wu, L., Xiong, S., Li, W., Garg, A., Gao, L.: An improved CNN-LSTM model-based state-of-health estimation approach for lithium-ion batteries. Energy **276**, 127585 (2023)
27. Zhang, D., Li, W., Han, X., Lu, B., Zhang, Q., Bo, C.: Evolving Elman neural networks based state-of-health estimation for satellite lithium-ion batteries. J. Energy Storage **59**, 106571 (2023)
28. Zhang, L., Ji, T., Yu, S., Liu, G.: Accurate prediction approach of soh for lithium-ion batteries based on LSTM method. Batteries **9**(3), 177 (2023)
29. Zhang, M., et al.: A review of soh prediction of li-ion batteries based on data-driven algorithms. Energies **16**(7), 3167 (2023)
30. Zhou, H., Huang, G.B., Lin, Z., Wang, H., Soh, Y.C.: Stacked extreme learning machines. IEEE Trans. Cybern. 2013 (2015)

3. Cai, L., Meng, J., Stroe, D.I., Peng, J., Luo, G., Teodorescu, R.: Multiobjective optimization of data-driven model for lithium-ion battery soh estimation with short-term feature. IEEE Trans. Power Electron. **35**(11), 11855–11864 (2020)
4. Chen, L., Xie, S., Lopes, A.M., Bao, X.: A vision transformer-based deep neural network for state of health estimation of lithium-ion batteries. Int. J. Electr. Power Energy Syst. **152**, 109233 (2023)
5. Dong, G., Chen, Z., Wei, J., Ling, Q.: Battery health prognosis using Brownian motion modeling and particle filtering. IEEE Trans. Ind. Electron. **65**(11), 8646–8655 (2018)
6. Dunn, B., Kamath, H., Tarascon, J.M.: Electrical energy storage for the grid: a battery of choices. Science **334**(6058), 928–935 (2011)
7. Ge, D., Zhang, Z., Kong, X., Wan, Z.: Extreme learning machine using bat optimization algorithm for estimating state of health of lithium-ion batteries. Appl. Sci. **12**(3), 1398 (2022)
8. Guille, C., Gross, G.: A conceptual framework for the vehicle-to-grid (V2G) implementation. Energy Policy **37**(11), 4379–4390 (2009)
9. Hsu, Y.M., Ji, D.Y., Miller, M., Jia, X., Lee, J.: Intelligent maintenance of electric vehicle battery charging systems and networks: challenges and opportunities. Int. J. Prognostics Health Manag. **14**(3) (2023)
10. Hu, X., Jiang, J., Cao, D., Egardt, B.: Battery health prognosis for electric vehicles using sample entropy and sparse Bayesian predictive modeling. IEEE Trans. Ind. Electron. **63**(4), 2645–2656 (2015)
11. Jiang, C., Torquato, R., Salles, D., Xu, W.: Method to assess the power-quality impact of plug-in electric vehicles. IEEE Trans. Power Delivery **29**(2), 958–965 (2013)
12. Kabir, M., Demirocak, D.E.: Degradation mechanisms in li-ion batteries: a state-of-the-art review. Int. J. Energy Res. **41**(14), 1963–1986 (2017)
13. Landi, M., Gross, G.: Measurement techniques for online battery state of health estimation in vehicle-to-grid applications. IEEE Trans. Instrum. Meas. **63**(5), 1224–1234 (2014)
14. Lee, K.M., Lee, S.W., Choi, Y.G., Kang, B.: Active balancing of li-ion battery cells using transformer as energy carrier. IEEE Trans. Ind. Electron. **64**(2), 1251–1257 (2016)
15. Li, S., et al.: Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
16. Ling, L., Wei, Y.: State-of-charge and state-of-health estimation for lithium-ion batteries based on dual fractional-order extended Kalman filter and online parameter identification. IEEE Access **9**, 47588–47602 (2021)
17. Luo, K., Zheng, H., Shi, Z.: A simple feature extraction method for estimating the whole life cycle state of health of lithium-ion batteries using transformer-based neural network. J. Power Sources **576**, 233139 (2023)
18. Luo, Y., Zhu, T., Wan, S., Zhang, S., Li, K.: Optimal charging scheduling for large-scale EV (electric vehicle) deployment based on the interaction of the smart-grid and intelligent-transport systems. Energy **97**, 359–368 (2016)
19. Madawala, U.K., Thrimawithana, D.J.: A bidirectional inductive power interface for electric vehicles in V2G systems. IEEE Trans. Ind. Electron. **58**(10), 4789–4796 (2011)
20. Meda, U.S., Lal, L., Sushantha, M., Garg, P.: Solid electrolyte interphase (SEI), a boon or a bane for lithium batteries: a review on the recent advances. J. Energy Storage **47**, 103564 (2022)

Table 2. Compare the Transformer-based (layers = 4, $m = 20$) Results with those of other prediction models (Experiment I).

Value of m	Model	MAE	RMSE
$m = 20$	Decision Tree Regression	0.0752	0.0937
	Linear Regression	0.0503	0.0661
	KNN Regression	0.0549	0.0720
	Random Forest Regression	0.0494	0.0635
	Adaboost Regression	0.0504	0.0631
	GBRT Regression	0.0483	0.0648
	Bagging Regression	0.0511	0.0656
	ExtraTree Extreme Random Tree Regression	0.0762	0.0947
	Transformer (4 layer)	0.0485	0.0602

6 Conclusion

Li-ion batteries as network nodes for storing energy in energy networks, playing a crucial role in achieving bidirectional energy transmission. The stability, reliability, and security of the power grid are highly dependent on reducing the monitoring data sample and accurately predicting the SOH of multiple Li-ion batteries distributed throughout the network in a timely and effective manner. This article presents a novel data-driven approach that utilises a single-factor Transformer-based technique to predict the State of Health (SOH) of a Li-ion battery energy storage system within a power grid. The method involves consistently monitoring voltage change rates of 3, 5, 10, and 20 as vital health indicators when charging Li-ion batteries in the CC charging mode. These data are parallelised in the Transformer-based model to predict SOH accurately. Experiment I examines and analyses the predicted results against the actual SOH.

Due to the diverse types of Li-ion batteries used within the network, some decommissioned batteries are utilized as buffer power sources in the power grid. Thus, the risk of achieving bidirectional energy transmission is amplified within the network. Further enhancements to the model will be made in the future to cater to a broader range of application areas, involving more complex and diverse topological network nodes.

References

1. Adenusi, H., Chass, G.A., Passerini, S., Tian, K.V., Chen, G.: Lithium batteries and the solid electrolyte interphase (SEI)-progress and outlook. Adv. Energy Mater. **13**(10), 2203307 (2023)
2. Andre, D., Appel, C., Soczka-Guth, T., Sauer, D.U.: Advanced mathematical methods of SOC and SOH estimation for lithium-ion batteries. J. Power Sources **224**, 20–27 (2013)

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(SOH_i - \overline{SOH}_i)^2} \tag{5}$$

K is the total number of regression samples.SOH_i represents the estimationd SOH value from data science regression models while $\overline{SOH}_i$ stand for the real SOH value.

We conducted Experiment I on different values of m using Transformer models with different layers, and calculated the MAE and RMSE of the SOH prediction results. From the table (see Table 1), we can clearly see that the error rate is lowest ($MAE = 0.0485$, $RMSE = 0.0602$) when m is 20.

Table 1. Errors Corresponding to Different Layers and m values (Experiment I).

Number of layers	Value of m	MAE	RMSE
1 layers	$m = 3$	0.0513	0.0639
	$m = 5$	0.0508	0.0645
	$m = 10$	0.0524	0.0631
	$m = 20$	0.0523	0.0635
2 layers	$m = 3$	0.0509	0.0629
	$m = 5$	0.0488	0.0629
	$m = 10$	0.0491	0.0612
	$m = 20$	0.0492	0.0608
3 layers	$m = 3$	0.0509	0.0629
	$m = 5$	0.0488	0.0629
	$m = 10$	0.0491	0.0612
	$m = 20$	0.0490	0.0604
4 layers	$m = 3$	0.0497	0.0631
	$m = 5$	0.0490	0.0626
	$m = 10$	0.0491	0.0606
	$m = 20$	0.0485	0.0602

Based on the results of Experiment I, we selected the Transformer model with $m = 20$ and Layer $= 4$ to repeat Experiment I and compared it with the main algorithm models currently used. The results (see Table 2) clearly indicate that our proposed Transformer-based method has lower error rates ($MAE = 0.0485$, $RMSE = 0.0602$), and therefore, higher prediction accuracy.

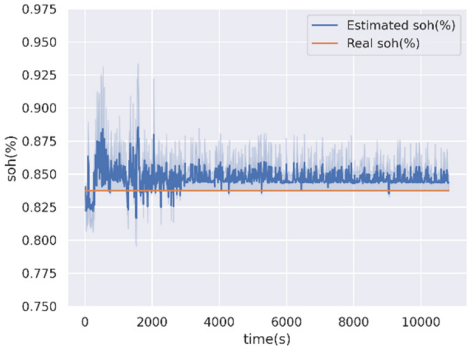


Fig. 6. Estimationd SOH Values ($m = 3$)

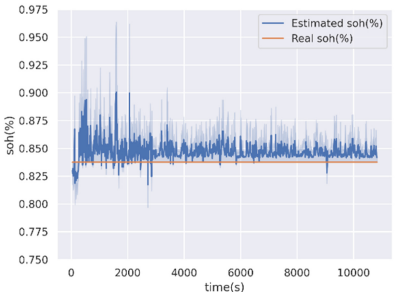


Fig. 7. Estimationd SOH Values ($m = 5$)

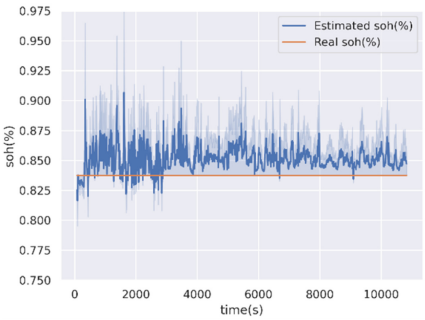


Fig. 8. Estimationd SOH Values ($m = 10$)

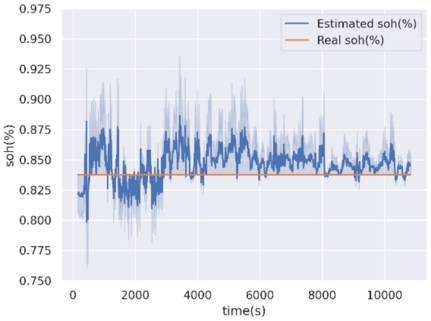


Fig. 9. Estimationd SOH Values ($m = 20$)

the accuracy of the Transformer prediction method in predicting slow decay of lithium batteries throughout their lifecycle.

It was discovered through Experiment I that the prediction accuracy of the proposed Transformer-based technique varies based on the value of m, and the corresponding errors also differ. The accuracy of the prediction results is assessed using MAE (Mean Absolute Error) and RMSE, where MAE stands for Mean Absolute Error.

$$MAE = \frac{1}{K}\sum_{i=1}^{K} |SOH_i - \overline{SOH}_i| \tag{4}$$

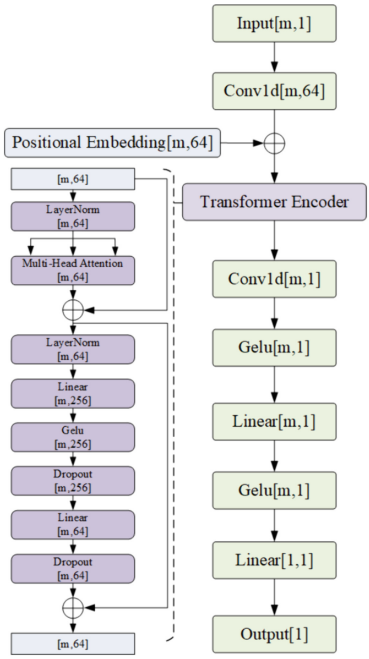


Fig. 5. Data Processing Diagram Based on Transformer.

5 Test Result and Analysis

To accurately simulate a random buffer battery within an energy network node and estimate its State of Health (SOH) for comparison with the actual value, thus demonstrating the accuracy of our model predictions, we utilised a standard charge-discharge test dataset in this paper. For Experiment I, we randomly selected deteriorated lithium batteries ($SOH = 0.837$) from NASA's publicly available 18650 Li-ion battery dataset. The lithium batteries selected were subjected to constant current charging using a current of $I = 1.5\,\mathrm{A}$ and cut-off voltage of 4.2 V. A total of m consecutive data x were obtained randomly from the initial charging process and fed into the prediction model for calculation. To ensure improved verification of the impact of consistently collecting various parameters on the ultimate prediction accuracy. We selected $m = 3$, 5, 10, and 20 to compare the predicted SOH with the actual SOH, and the corresponding graphs are shown respectively (see Fig. 6, Fig. 7, Fig. 8, Fig. 9).

In an operational energy network, it is imperative to carry out real-time calculations of the SOH of buffer lithium batteries in a network node. When the SOH of these buffer lithium batteries drops to 80%, they are removed from the network. The buffer lithium batteries are assumed to be built using the highest-quality new batteries with uniform quality. It is paramount to validate

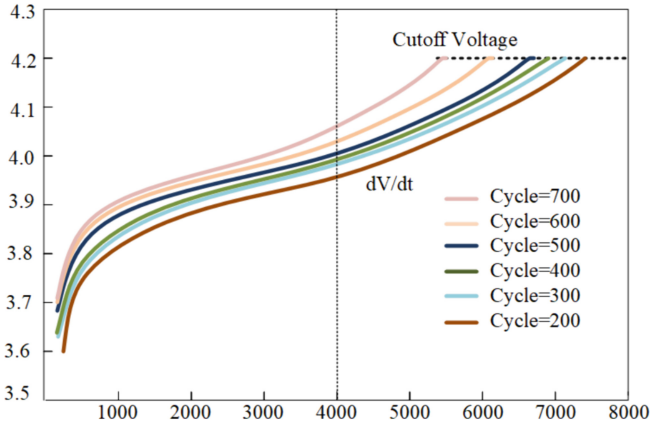


Fig. 4. The Voltage Change Curve of Li-ion Batteries Charged in CC Mode After Different Cycles.

$$x = \frac{dV(t)}{dt} \tag{2}$$

$V(t)$ represents the real-time voltage that changes with time t. Although batteries with lower SOH have faster charging rates, that is, the x value will be larger. In data-driven SOH prediction methods, if the selected features have a higher correlation with SOH changes, the prediction accuracy will be higher [17]. Pearson correlation is calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3}$$

Here, n is the cycle number of the battery, and r takes a value between -1 and $+1$. x_i represents the x value taken at time$i$, and $y_i$represents the SOH value at time i. For the state data during continuous charging of the battery, due to the large randomness of individual x values, we take 3, 5, 10, and 20 consecutive x values at time i, $x_i = [x_i^3, x_i^5, x_i^{10}, x_i^{20}]$. We perform correlation analysis on them separately to obtain the result $r = 0.959$ (when $x_i = x_i^{20}$), where $r = 0.915$ (when $x_i = x_i^3$),$r = 0.924$ (when $x_i = x_i^5$),$r = 0.932$ (when $x_i = x_i^{10}$), that all are higher than other indicators (e.g. temperature, current rate, cut-off voltage, etc.).

4.2 Transformer-Based Model

To enhance parallel computing and the accuracy of time series data processing, this study adopts the Transformer framework [4,15,21]. The framework, illustrated in the figure, comprises essential input, output, and data preprocessing modules. The key component of the Transformer Encoder is the multi-layered Multi-head self-attention, which utilises various sets of attention mechanisms (see Fig. 5).

energy network in supplying energy at a specific location and time is somewhat reliant on the precise assessment of the SOH of buffer batteries. The diverse categories, technical features and abilities of buffer batteries within the energy network points, lead to a considerable variation in SOH. This is especially so for buffer batteries utilizing Li-ion batteries for composite utilization, with a power ranging from 40% to 80% of the original capacity. After estimating the SOH in real-time, the data is transmitted to the communication network layer. This layer then processes and sends the data to the energy network gateway for balancing and dispatch of supply and demand. Traditional SOH estimation methods and models are inadequate for the large volume of sequential data from buffer batteries. They require continuous collection, transmission, and processing of large amounts of data and cannot handle large amounts of data in parallel. Serial calculations for individual Li-ion batteries result in low efficiency and fail to meet the requirements for simultaneous bi-directional energy transmission across multiple nodes.

4 Transformer-Based SOH Prediction Method

4.1 Feature Extraction and Selection

After a period of usage, the capacity of Li-ion batteries diminishes and their internal resistance increases. This is caused by the loss of active materials and the thickening of the Solid Electrolyte Interface (SEI) membrane, among other factors [1,20]. The depletion of active materials in Li-ion batteries is primarily caused by metal ions dissolving from the positive electrode material into the electrolyte. Simultaneously, a film of SEI gradually develops on the negative electrode surface, progressively impacting the lithium ion supply and inducing an uptick in internal resistance. Electrochemical reaction processes are responsible for the internal alterations of Li-ion batteries, which present quantitative monitoring challenges. From the external characteristics of Li-ion batteries, the mentioned internal changes result in increased internal resistance, reduced capacity$Q$, and a stronger inclination towards heat generation. Different SOH Li-ion batteries have charging characteristics as shown (see Fig. 4).

When charging Li-ion batteries by connecting to the power grid, a constant current charging mode is typically utilised. Additionally, factual and unambiguous titles should be used, along with conventional academic sections. Lastly, objective and value-neutral language should be employed, avoiding biased or emotional language and intricate terminology. In such a mode, the charging curve for batteries with different SOH levels demonstrates that with an increase in the cycle times of Li-ion batteries, the battery reaches its cut-off voltage more rapidly. It is essential to explain the abbreviations of technical terms when they are first used in the text. When applying a constant current mode to charge a Li-ion battery, the rate at which its voltage rises varies. Among them, with a constant current charging mode of 0.5 C as an example, the charging rate is faster when the SOH of the battery is worse, making it easier to fully charge. We take the parameter x.

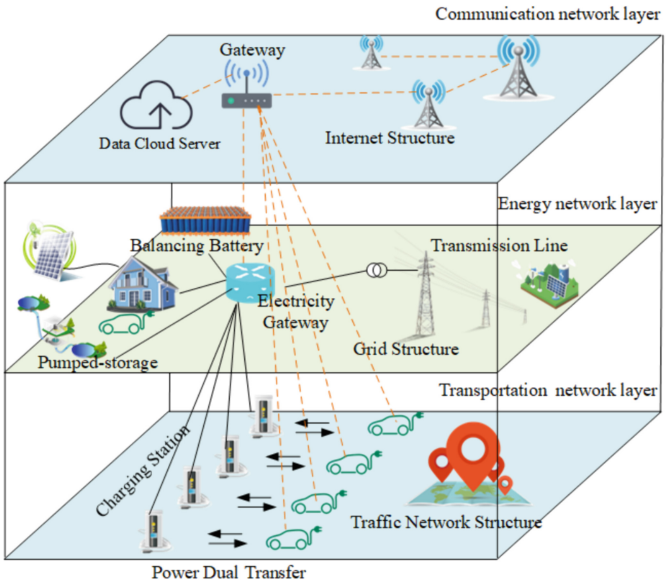


Fig. 3. Integrated Framework for Energy, Communication, and Transportation Networks

each buffer battery determines the optimal charging and discharging strategy. If the SOH judgment of the buffer battery is incorrect, excessive current often leads to overheating of Li-ion batteries, and even internal short circuits, fires, and other issues, threatening the reliability of the entire power system. In the network nodes, a battery management system (BMS) is installed for each buffer battery to monitor the energy flow, power, and temperature changes, especially the SOH of Li-ion batteries [22].

3.2 Stability, Reliability and Safety Requirements in Network

Stability, reliability and safety are crucial measures for evaluating the energy grid. The energy grid possesses traits of bidirectional energy transmission, vast energy distribution and dispatch, as well as substantial volumes of communication data. Buffer batteries are situated in many nodes of the energy grid, and each network node has varying energy storage and supply capabilities. Mobile buffer energy storage systems can determine SOH of the Li-ion batteries onboard by utilising the battery management system (BMS) installed in the vehicle, which can communicate directly with the communication base station [11]. Buffer batteries often draw and release energy from the power network, particularly during emergencies requiring deep or high-power discharge. These batteries prioritize sacrificing the value of Li-ion battery life to maintain grid stability. Nevertheless, the rapid depletion of buffer Li-ion battery life or thermal runaway can reduce the safety and reliability of the energy network. The efficiency of the power

accuracy of the prediction results. It is important to avoid relying too heavily on the completeness of the collected historical data set, especially if the data set is limited, as this can significantly affect the accuracy of the prediction [3]. It is important to avoid relying too heavily on the completeness of the collected historical data set, especially if the data set is limited, as this can significantly affect the accuracy of the prediction.

3 Communication-Energy-Transportation Network

3.1 Topological Structure

In this extensive electricity network, its primary role is to carry the transmission of electrical power. Electricity generation primarily originates from power stations, comprising thermal power stations, nuclear power stations, wind farms, and solar power farms. The resulting electrical power is transported through power lines to the point of consumption of electrical energy. To increase the efficacy of electrical power utilization and promptly fulfil sporadic power needs such as electric vehicle charging, Li-ion batteries have gained significant traction as proficient energy buffer systems with growing application demands [13]. The Communication-Energy-Transportation Network is comprised of three layers: the communication network layer, which is based on the Internet architecture, the energy network layer, which is based on the power grid architecture, and the transportation network layer, which is based on the traffic structure (see Fig. 3). The energy network layer holds a critical role, encompassing the generation, transmission, and dispatch of electrical energy. The dispatch and allocation are dependent on the electricity gateway, which must acquire details of the electric vehicles' charging and discharging requirements from the transportation network layer. When operating in V2G mode, electric vehicles can supply energy back to the power grid and act as portable energy storage stations. When the demand for power falls below the scheduled amount, surplus energy will either be stored in the buffer Li-ion battery (balancing battery) [14] or the onboard Li-ion battery, or a negligible quantity of energy will be absorbed by Pumped Storage. When electricity demand exceeds the scheduled power, the stored energy in the buffer Li-ion battery or the on-board Li-ion battery will discharge into the power grid, supplementing the insufficient power. The energy dispatch and allocation centre, or electricity gateway, is responsible for this function. The gateway transmits scheduling data to the communication network gateway and stores the data in the Data Cloud Server database. The transmission of information regarding vehicles, charging stations, power transmission, and dispatching allocation is facilitated through the communication network layer. This layer operates independently from subjective evaluations, allowing for objectivity and value-neutral communication.

In this complex network, Li-ion batteries are located at the key node position for energy dispatching and allocation, which directly affects the stability and security of the entire energy network [9]. Each buffer battery in the network node has different geographical locations, types, and capacities, and the SOH of

accuracy and prediction precision. There are many types of data-driven methods, such as decision tree regression [23], random forest regression, ExtraTree extreme random tree regression [7], and various methods based on deep learning [26], including BP (Back Propagation), ENN (Elman Neural Network) [27], ELM (Extreme Learning Machine) [30], CNN (Convolutional Neural Network) [24], and long short-term memory (LSTM) [28].

By collecting historical data of Li-ion battery charging and discharging (including voltage, current, temperature and other variables), feature factors affecting SOH are extracted to train neural networks. Test data is used to evaluate the model and calculate its accuracy, error and other indicators to obtain a well-trained neural network. According to the current battery status (such as current, voltage, temperature, etc.), this information is converted into a feature vector that is input into the trained neural network-based prediction model, and the model outputs the predicted SOH value (see Fig. 2).

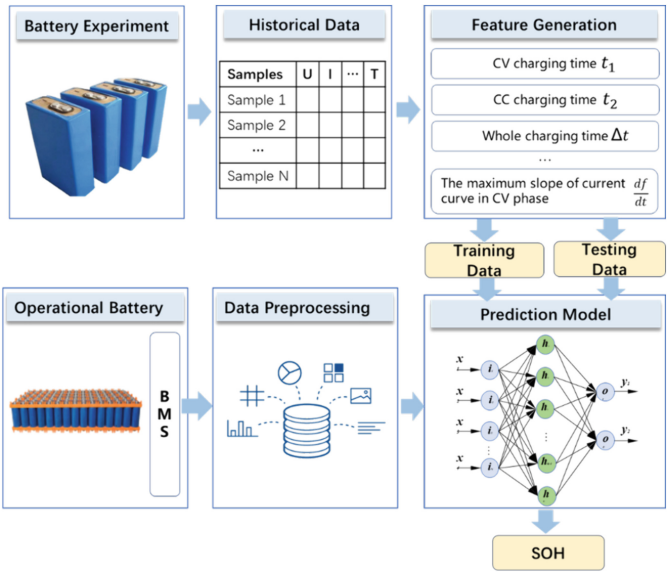


Fig. 2. Data Driven SOH Prediction Process.

Data-driven methods avoid the need for a deep understanding of the electrochemical reaction mechanisms in Li-ion batteries [29]. By using historical data and appropriate neural network algorithms, it is possible to predict State of Health (SOH). This approach is characterised by easy operation, quick modelling, and fast calculations. It is suitable for a range of battery types, including those with complex chemical compositions and structural differences. Based on past data for prediction, it is crucial to have adequate and high-quality historical data. Inadequate or imprecise historical data can have an impact on the

Q_t represents the current battery capacity, and Q_{new} represents the new battery capacity. It quantitatively describes the performance status of the current battery as a percentage from the beginning to the end of its life cycle. The state of health of Li-ion batteries cannot be measured directly. At present, the SOH assessment of lithium batteries can be divided into two main categories.

2.1 Model-Based SOH Prediction

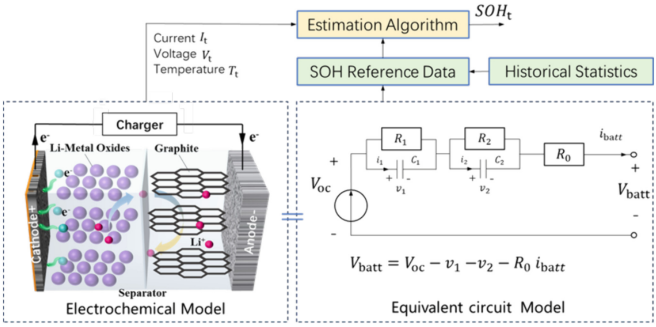


Fig. 1. Electrochemical Model of Lithium Batteries.

There are currently many mainstream types of Li-ion batteries, including lithium cobalt oxide (LiCoO2) batteries, lithium iron phosphate (LiFePO4) batteries, lithium manganese oxide (LiMn2O4) batteries, nickel cobalt manganese (NCM) batteries, nickel cobalt aluminium (NCA) batteries and lithium manganese oxide (LMO) batteries, among others. Although different Li-ion batteries have different discharge characteristics, their electrochemical charging and discharging principles are almost the same. When analysing the electrochemical model of Li-ion batteries, it is necessary to establish a corresponding complex mathematical model or equivalent circuit model [2] For example, Bayesian statistical models [10], Kalman filters [16], particle filters [5] and other estimation algorithms can be used. By establishing a dynamic model based on electrochemical principles, technical parameters (reaction rate, etc.) are determined using experimental battery data. A filter algorithm is used to compare the real-time battery status (real-time current I_t, voltage V_t, and temperature T_t) with the standard data in the model to estimate the SOH value of the current battery (see Fig. 1). The accuracy of model-based SOH prediction is often limited by the accuracy of the model and requires more prior knowledge and complex calculations.

2.2 Data-Driven SOH Prediction

Data-driven methods start from experimental data of the actual charging and discharging process of Li-ion batteries, rely on direct analysis of experimental data, extract specific features for estimating battery SOH, and have higher

inefficient energy bidirectional interaction. The battery module with the lowest SOH in the grid restricts the energy transmission efficiency of the entire network. Adopting diverse SOH-balancing Li-ion batteries and batteries mounted on vehicles can help alleviate this issue. Pure electric vehicles can be used to buffer energy in the grid. However, connecting them directly for bidirectional charging and discharging at high currents and powers may lead to a decrease in safety performance of the energy network terminal. It can even cause fires [24,25]. The usage, discharge patterns, and operating conditions of Li-ion batteries in power energy network nodes differ from those of electric vehicle Li-ion batteries. The degradation rate of this battery type can be complex and challenging to anticipate. Furthermore, with numerous network nodes in the grid, computing the State of Health (SOH) of each battery singly necessitates significant computation and interaction information transfer with low timeliness. Consequently, there is a need to enhance and complement methods and models for approximating the SOH of these batteries throughout the bidirectional energy exchange system.

The above researchers have proposed many estimation methods for Li-ion batteries. However, in the future, it is expected that Li-ion battery estimation methods can accurately, reliably and quickly estimate the state of health (SOH) of Li-ion batteries based on fewer data types and sample sizes. The purpose of this paper is to use the rate of voltage change during Li-ion battery charging as a single influencing factor to predict the SOH of Li-ion batteries. This application involves comparing the characteristics of existing Li-ion battery SOH prediction methods, including two main directions: (1) model-based SOH prediction and (2) data-driven SOH prediction. Details are given in Sect. 2. It's analysed the challenges of estimating the SOH of Li-ion batteries in the networks, which includes three types of heterogeneous networks: power network, communication network and transportation network in Sect. 3. The Li-ion battery is used as a buffer battery in this network, and Sect. 4 focuses on the challenges of estimating the SOH of Li-ion batteries in the energy network mentioned in the third part. It selects the rate of voltage change as the only influencing factor and adopts a transformer-based model to analyse the NASA public battery datasets. Section 5 presents and discusses the verification results. The paper is concluded in Sect. 6.

2 Related Work

In power grids, the capacity degradation of backup Li-ion batteries is a long-term gradual process and the health of the battery is affected by various factors such as temperature, current rate, cut-off voltage, etc. The available capacity can be used to characterise the performance of the battery and the SOH is the percentage of available capacity compared to the initial capacity. The available capacity can be used to characterise the performance of the battery and the SOH is represented by the percentage of available capacity to the initial capacity.

$$SOH = \frac{Q_t}{Q_{new}} \times 100\% \tag{1}$$

1 Introduction

As intelligent transportation systems, 5G communication technologies, smart grid technologies, and energy storage technologies further improved, a new trend of network intersections and fusion is gradually forming, involving vehicles, communication base stations, energy supply stations. As the vehicle in the intelligent transportation system, Vehicle-to-Everything (V2X) and Vehicle-to-Grid (V2G) have emerged as key technologies [8]. These technologies facilitate not only the extraction of energy from the power grid, but also the return of stored energy to the grid, thereby enabling bidirectional energy interaction [19]. Around the grid network, smart grid technology has developed. Moreover, with energy storage technologies as the core, the Energy Management System (EMS) incorporates photovoltaic, energy storage, and charging capabilities.

The widespread use of mobile charging terminals (such as electric bicycles, electric cars, etc.) increases the difficulty of planning and ensuring the safety of the entire power system due to the randomness of the charging time, the network node where the charging takes place, and the high power consumption during charging [18]. Pure electric vehicles are the most common type of mobile charging terminal. Battery types in the on-board energy storage system include lead-acid batteries, nickel-hydrogen batteries, Li-ion batteries, super-capacitors, fuel cells, etc. Among these, Li-ion batteries have become the main storage medium. As the power battery of the vehicle, the life of the Li-ion battery determines the life of the electric vehicle. During the use of the power battery, a series of electrochemical reactions occur inside the battery due to the accumulation of time, changes in ambient temperature, and the increase in charging and discharging cycles, which cause the performance of the Li-ion battery to deteriorate, such as the reduction of lithium ions, loss of active materials, and electrode wear. Battery performance is usually represented by available capacity, and the percentage of available capacity to initial capacity is used to indicate the health of the battery [12]. A gradual decline in battery performance will result in abnormal performance, which may even threaten road safety. Battery safety is key to the design of electric vehicles. Predicting battery life and managing battery health can enable timely replacement or maintenance of batteries based on their SOH and Remaining Useful Life (RUL), ensuring safe availability and stable reliability of battery operation. When the capacity of a Li-ion battery is reduced to 80% of its initial capacity, the battery has reached its End of Life (EOL) and is no longer suitable for on-board use. Indicators for measuring the RUL of on-board batteries include when the SOH drops below 80% or when the internal resistance doubles compared to the original value, which indicates that the on-board Li-ion battery has reached the end of its life.

In the power network, the power grid cannot store electrical energy by itself, and energy storage batteries are utilized as the electrical storage and buffering unit in the system, with Li-ion batteries being the most commonly used [6]. As the primary energy network, the Li-ion batteries in different network nodes often possess dissimilar SOH, which results in variations in their capacity, specifications, power output, and other related characteristics. This, in turn, can cause

SOH Prediction in Li-ion Battery Energy Storage System in Power Energy Network

Xiaofen Fang[1,3(✉)], Kai Fang[2], Lihui Zheng[3], Han Zhu[4], Qichang Zhuo[5], and Jianqing Li[1]

[1] Macau University of Science and Technology, Macau 999078, China
fangxiaofen1985@hotmail.com
[2] Zhejiang Agricultural and Forestry University, Hangzhou 311300, China
[3] Quzhou College of Technology, Quzhou 324000, China
[4] Macao Polytechnic University, Macau 999078, China
[5] Huayou Cobalt Co., Ltd., Quzhou 324000, China

Abstract. The prediction of the State of Health (SOH) of Li-ion batteries is crucial for the system safety and stability of the entire energy network. In this paper, we analyse the role of Li-ion batteries as balancing batteries in the communication-energy-transportation network, which are key nodes for energy exchange. These batteries have different states of health and are constantly in a state of bi-directional energy transfer through charging and discharging. They also require coordinated charging and discharging in the network. Due to these differences, the degradation rates of the different balancing batteries vary, making it necessary to monitor the SOH of each battery in real time. To address the problem of numerous nodes and large computational requirements in the network, we propose a method based on the Transformer architecture for accurate and fast estimation, which can alleviate the communication pressure of the energy network layer. In this method, we select the voltage change rate as the only key health indicator and perform correlation analysis while the battery is in a constant current CC charging mode. According to the test results on NASA public battery data set, we continuously collect 3, 5, 10 and 20 voltage change rate values as inputs to the model and validate 1–4 layer transformer models. The model with 20 voltage change rate values as inputs and 4 layers has the best prediction performance, with a Mean Absolute Error (MAE) as low as 4.85%, while the Root Mean Square Error (RMSE) is 6.02%. In addition, the proposed method can process the time series data of multiple network nodes in real time and in parallel, with high prediction accuracy and stable performance, which makes it suitable for widespread use in power energy networks.

Keywords: Voltage change rate · Transformer-based · Battery energy storage system · State of health prediction · Energy network

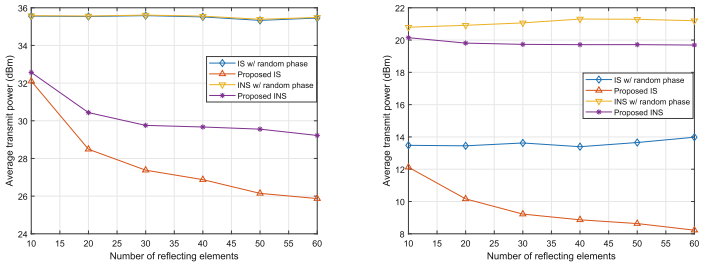
© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 457–471, 2024.
https://doi.org/10.1007/978-3-031-65126-7_38

18. Wu, Q., Zhang, S., Zheng, B., You, C., Zhang, R.: Intelligent reflecting surface-aided wireless communications: a tutorial. IEEE Trans. Commun. **69**(5), 3313–3351 (2021)
19. Zheng, B., Wu, Q., Zhang, R.: Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA? IEEE Commun.s Lett. **24**(4), 753–757 (2020)
20. Zheng, B., Zhang, R.: Intelligent reflecting surface-enhanced OFDM: channel estimation and reflection optimization. IEEE Wirel. Commun. Lett. **9**(4), 518–522 (2020)

method and then acquired the solution with the standard optimization methods. In the reflection design, we developed the MGD-RM based algorithm to obtain the local optimal solution. Simulation results demonstrated the superiority of deploying IRS and employing SLP in HetNet and the effectiveness of our proposed algorithms.

References

1. Alhujaili, K., Monga, V., Rangaswamy, M.: Transmit MIMO radar beampattern design via optimization on the complex circle manifold. IEEE Trans. Signal Process. **67**(13), 3561–3575 (2019)
2. Boyd, S., Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
3. Costa, M.: Writing on dirty paper (corresp.). IEEE Trans. Inf. Theory **29**(3), 439–441 (1983)
4. Désidéri, J.A.: Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. C.R. Math. **350**(5–6), 313–318 (2012)
5. Ericsson: 5G for business: a 2030 market compass. Setting a direction for 5G powered B2B opportunities. White Paper (2019)
6. Li, A., Masouros, C.: Interference exploitation precoding made practical: Optimal closed-form solutions for PSK modulations. IEEE Trans. Wirel. Commun. **17**(11), 7661–7676 (2018)
7. Li, A., et al.: A tutorial on interference exploitation via symbol-level precoding: overview, state-of-the-art and future directions. IEEE Commun. Surveys Tuts **22**(2), 796–839 (2020)
8. Lin, S., Zheng, B., Alexandropoulos, G.C., Wen, M., Renzo, M.D., Chen, F.: Reconfigurable intelligent surfaces with reflection pattern modulation: beamforming design and performance analysis. IEEE Trans. Wirel. Commun. **20**(2), 741–754 (2021)
9. Lopez-Perez, D., Guvenc, I., de la Roche, G., Kountouris, M., Quek, T.Q., Zhang, J.: Enhanced intercell interference coordination challenges in heterogeneous networks. IEEE Wirel. Commun. **18**(3), 22–30 (2011)
10. Marler, R.T., Arora, J.S.: Survey of multi-objective optimization methods for engineering. Struct. Multidisciplinary Optim. **26**(6), 369–395 (2004)
11. Pan, C., et al.: Multicell MIMO communications relying on intelligent reflecting surfaces. IEEE Trans. Wirel. Commun. **19**(8), 5218–5233 (2020)
12. Sawaragi, Y., Nakayama, H., Tanino, T.: Theory of Multiobjective Optimization. Elsevier (1985)
13. Shewchuk, J.R.: An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Carnegie Mellon University (1994)
14. Shewchuk, J.R.: An introduction to the conjugate gradient method without the agonizing pain. Ph.D. thesis, Carnegie Mellon University (1994)
15. Sun, Y., Babu, P., Palomar, D.P.: Majorization-minimization algorithms in signal processing, communications, and machine learning. IEEE Trans. Signal Process. **65**(3), 794–816 (2017)
16. Wen, M., et al.: Private 5G networks: concepts, architectures, and research landscape. IEEE J. Sel. Top. Signal Process. **16**(1), 7–25 (2022)
17. Wu, Q., Zhang, R.: Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming. IEEE Trans. Wirel. Commun. **18**(11), 5394–5409 (2019)



(a) MBS. (b) PBS.

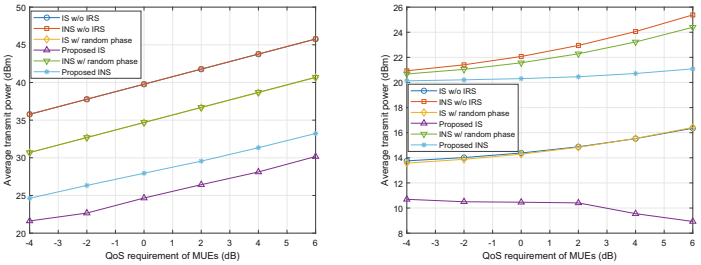
Fig. 5. Average transmit power versus the number of reflecting elements, where $\Gamma_{M,k} = 0$ dB, $\Gamma_{P,n} = 10$ dB and $\xi_1 = \xi_2 = 0.5$, respectively.

the increasing QoS requirements of PUEs. Furthermore, it is also seen that the proposed scheme outperforms the INS scheme thanks to the precoding design for PBS via SLP in the proposed scheme. In Fig. 4(b), we can see that the transmit power at the MBS changes barely with the increasing QoS requirements of PUEs in all schemes. The reason is that the received signal of MUEs is not interfered by PBS, such that the MBS can still transmit the signal at the current power levels despite variations in the QoS requirements of the PUE.

In Fig. 5, we present average power at the MBS and PBS versus the number of reflecting elements. Since the larger number of the reflecting elements at the IRS offers larger reflecting gains, the transmit power at the MBS decreases for the proposed scheme and INS schemes with the increasing number of the reflecting elements, as shown in Fig. 5(a). Moreover, the power-saving performance of our proposed scheme always outperforms that of the other schemes. In fact, the power-saving performance of the PBS can also be improved by optimizing the reflecting coefficients. Figure 5(b) shows that the transmit power at the PBS in the proposed scheme decreases with the increasing size of the IRS. The reason behind this observation is that the reflecting coefficients are designed under the SLP principle, which narrows in a reduced phase shift between the ICI and desired signal intended fro the PUES and renders ICI more constructive for PBS.

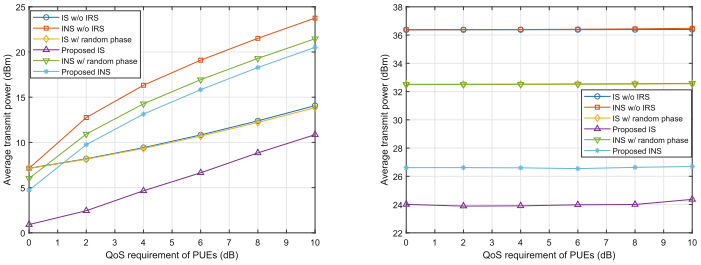
6 Conclusion

In this paper, we investigated a two-tier IRS-enhanced downlink HetNet, where an IRS was deployed for assisting the MBS, which employs the SLP to handle MUI and ICI. The power minimization problems with two objectives were formulated for the joint precoding matrices at the MBS and PBS as well as reflecting coefficients at the IRS. To deal with the non-convexity of the MOO problems, we proposed an AO method to update the precoding matrices and reflecting coefficients alternatively. For optimizing the precoding matrices, we transformed the MOO problem into an SOO problem by adopting the weighted Tchebycheff



(a) MBS. (b) PBS.

Fig. 3. Average transmit power versus QoS requirement of MUEs, where $\Gamma_{P,n} = 10$ dB, $M = 32$ and $\xi_1 = \xi_2 = 0.5$, respectively.



(a) PBS. (b) MBS.

Fig. 4. Average transmit power versus QoS requirement of PUEs, where $\Gamma_M = -4$ dB, $M = 32$ and $\xi_1 = \xi_2 = 0.5$, respectively.

the reflecting coefficients and without the assistance of the IRS are denoted as "INS w/ random phase" and "INS w/o IRS", respectively. From Fig. 3(a), it is noticed that the transmit power-saving at the MBS of the proposed scheme outperforms the INS scheme and the schemes with the random reflecting coefficients or without the assistance of the IS. It is also observed that the power-saving performance of "IS w/ random phase" and "INS w/ random phase", the lines of "IS w/o IRS" and "INS w/o IRS" are almost overlapping, but both are worse than the proposed scheme. This means that the power-saving performance of MBS can be significantly improved by introducing the assistance of IRS in the HetNet. Figure 3(b) presents the comparison of the average transmit power at the PBS in our proposed scheme and the other schemes under different QoS requirements of MUEs. The proposed scheme has a remarkable power-saving benefit over the INS scheme due to the efficient exploitation of MUI and ICI by the PBS via SLP and IRS, and the transmit power at the PBS reduces as the QoS requirements of MUEs increases.

In Fig. 4, we present average power at the MBS and PBS versus the QoS requirements of PUEs. Figure 4(a) shows that the transmit powers at the PBS for the proposed scheme and other baseline schemes increases monotonically with

the path loss at the reference distance of $d_0 = 1$ m, and α and d are denoted as the path loss exponent and link distance, respectively. We assume that an MBS equipped with $N_M = 4$ antennas serves 4 MUEs and a PBS equipped with $N_P = 2$ serves 2 PUEs. The IRS element number is set as $M = 32$. Denote the path loss exponents of MBS-MUE, MBS-IRS, IRS-MUE, MBS-PUE, IRS-PUE, and PBS-PUE as α_{MMU}, α_{MI}, α_{IMU}, α_{MPU}, α_{IPU}, and α_{PPU}, respectively, and let $\alpha_{MMU} = \alpha_{MPU} = \alpha_{PPU} = 3.5$, and $\alpha_{MI} = \alpha_{IMU} = \alpha_{IPU} = 2.2$. The MUEs and PUEs are distributed randomly within a macrocell and a picocell. Define $d_{MMU} \in (0, r_M]$ and $d_{PPU} \in (0, r_P]$ as MBS-MUE and PBS-PUE link distances, respectively. Correspondingly, the IRS-MUE, MBS-PUE, and IRS-PUE link distances are denoted as $d_{IMU} \in (|d_{MM} - d_{MI}|, d_{MMU} + d_{MI})$, $d_{MPU} \in (d_{MP} - d_{PPU}, d_{MP} + d_{PPU})$, and $d_{IPU} \in (d_{IP} - d_{PPU}, d_{IP} + d_{PPU})$, respectively. Furthermore, we consider the Rician fading channel model as the small-scale fading model for all channels involved in this simulation. Hence, the generalized channel is given by

$$\mathbf{H} = (PL(d))^{\frac{1}{2}} \left(\sqrt{\frac{\kappa}{1+\kappa}} \mathbf{H}_{\text{LOS}} + \sqrt{\frac{1}{1+\kappa}} \mathbf{H}_{\text{NLOS}} \right), \tag{27a}$$

$$\mathbf{H}_{\text{LOS}} = \mathbf{a}_M \left(\vartheta^{\text{AOD}} \right) \mathbf{a}_N^H \left(\vartheta^{\text{AOA}} \right), \tag{27b}$$

$$\mathbf{a}_t(\vartheta) = \left[1, e^{j \frac{2\pi D}{\lambda_c} \sin \vartheta}, \cdots, e^{j \frac{2\pi D}{\lambda_c} (t-1) \sin \vartheta} \right]^T, \tag{27c}$$

for $t \in \{N_t, N_r\}$, where N_t and N_r denote the numbers of transmitted and received antennas, respectively. When M or N is equal to 1, the above channel will evolve into MISO or SIMO channel. In (27a), $\mathbf{H}_{\text{LOS}}$ and $\mathbf{H}_{\text{NLOS}}$ represent the deterministic light of sight and Rayleigh fading components, respectively. The Rician factor κ is set as a huge value for the channel between the MBS to IRS and 3 for the other channels involved. In (27b), ϑ^{AOD} and ϑ^{AOA} denote the angle of departure from the transmitter and that of arrival from the receiver, respectively. In (27c), D is the distance traveled by the path and λ_c is the carrier wavelength. For the sake of simplicity, it is reasonable to assume that ϑ^{AOD} and ϑ^{AOA} are uniformly distributed between 0 and 2π for all channels involved, and D/λ_c is equal to 0.5. The QoS requirements of the MUEs and PUEs are denoted as $\Gamma_{M,k} = \Gamma_M$ and $\Gamma_{P,n} = \Gamma_P$, respectively. All transmitted symbols are generated randomly with QPSK modulation, and all simulation results are averaged over 1000 independent channel realizations.

In Fig. 3, we present the average transmit power at the MBS and PBS versus the QoS requirements of MUEs. In the legend, the considered HetNet using our proposed algorithm is denoted as "proposed information sharing (IS)", and the scheme regarding information not sharing (INS) between the MBS and PBS is denoted as "INS", where the precoding matrix for the MBS is designed by SLP while the precoding matrix for the PBS is adopted by the conventional precoding method due to no knowledge of the information symbols of the MUEs. The considered HetNets with the random phase setup of the reflecting coefficients and without the assistance of the IRS are denoted as "IS w/ random phase" and "IS w/o IRS", respectively. The INS schemes with the random phase setup of

4 Computational Complexity Analysis

In this section, we analyze the computational complexity of the proposed algorithm. The convex MOO problem (8) contains CI constraints for MUEs and PUEs, which can be solved by the interior point method [2]. The complexity of solving this MOO problem in the worst case can be approximated as $\mathcal{O}((K(N_M+2)+N(N_P+2)+2)^{3.5})$, where the numbers of optimized variables and constraints are $KN_M + NN_P$ and $2K + 2N + 2$, respectively. The computational complexity of the reflecting design mainly lies in the calculation of the Euclidean gradient of $\mathbf{u}$. The reflecting design in the proposed MGD-RM algorithm requires computing the Euclidean gradient $K + N$ times. Therefore, the complexity of the proposed MGD-RM algorithm is given as $\mathcal{O}((K+N)M^2)$. As a result, the complexity of jointly optimizing SLP and reflecting design is given as $\mathcal{O}(T_{iter}((K(N_M+2)+N(N_P+2)+2)^{3.5}+(K+N)M^2))$, where T_{iter} denotes as the iteration times of the AO algorithm, which is generally less than 10 in the simulations.

Algorithm 1. Proposed multiple gradient descent based on Riemannian manifolds algorithm.

1: **Initialization:** Set the iteration number $q = 1$ and initialized point $\mathbf{u}_1$.
2: Calculate the Riemannian gradient of each objective function in (19), and initialize the common search direction $\mathbf{d}_1 = -\sum_{l=1}^{K+N} \alpha_l \mathbf{r}_l(\mathbf{u}_1)$.
3: **Repeat:**
4: Choose setup-length ς_q by using the Armijo backtracking line search method [13].
5: Update $\mathbf{u}_{q+1}$ by (20) with ς_q, $\mathbf{u}_q$ and $\mathbf{d}_q$.
6: Update the Riemannian gradient $\text{grad}J_l(\mathbf{u}_{q+1})$ according to (21) with $\mathbf{u}_{q+1}$
7: Calculate $\mathbf{r}_l(\mathbf{u}_{q+1})$ by Gram-Schmidt orthogonalization of $\text{grad}J_l(\mathbf{u}_{q+1})$.
8: Update the Riemannian gradient $\hat{\mathbf{r}}_{q+1}$ according to (23) with $\text{grad}J_l(\mathbf{u}_{q+1})$.
9: Calculate $\mathbf{d}_q^t$ according to Riemannian transport operation (26) with $\mathbf{d}_q$ and $\mathbf{u}_{q+1}$.
10: Choose the Polak-Ribiere parameter η_q by [13].
11: Update common search direction $\mathbf{d}_{q+1}$ according to (25) with $\hat{\mathbf{r}}_{q+1}$, $\mathbf{d}_q^t$ and η_q.
12: **Until:** Convergence.

5 Simulation Results

In this section, we investigate the performance of our proposed scheme via simulations. Assume the MBS is located at the center of a macrocell, the IRS and MBS are separated by $d_{MI} = 5$ m in the perpendicular direction, and the PBS and MBS are separated by $d_{MP} = 300$ m in the horizontal direction. The distance between the IRS and PBS is $d_{IP} = \sqrt{d_{MI}^2 + d_{MP}^2}$ m. The serving radius of the macrocell and picocell are set as $r_M = 500$ m and $r_P = 100$ m, respectively. The path loss model is given by $PL(d) = C_0(d/d_0)^{-\alpha}$, where $C_0 = -30$ dB is

2) Common Riemannian gradient: Secondly, we find a proper gradient that lets all objective functions share a common descent direction in the next iteration. The concept of the Riemannian gradient is introduced in the calculation. The Riemannian gradient is computed based on the projection of the smooth objective function onto the tangent space $T_{\mathbf{u}}S$, which is given by [1]

$$\text{grad}J_l(\mathbf{u}_{q+1})=\bigtriangledown J_l(\mathbf{u}_{q+1})-\Re\left\{\bigtriangledown J_l(\mathbf{u}_{q+1})\odot\mathbf{u}_{q+1}^*\right\}\odot\mathbf{u}_{q+1}, l=1,2,\cdots,K+N, \tag{21}$$

where $\bigtriangledown J_l(\mathbf{u})$ denotes Euclidean gradient of $J_l(\mathbf{u})$, which is given by

$$\bigtriangledown J_l(\mathbf{u})=\begin{cases}\dfrac{\exp(\hat{f}_{2l-1}/\varepsilon)\tilde{\mathbf{a}}_{2l-1}+\exp(\hat{f}_{2l}/\varepsilon)\tilde{\mathbf{a}}_{2l}}{\exp(\hat{f}_{2l-1}/\varepsilon)+\exp(\hat{f}_{2l}/\varepsilon)}, l=1,2,\cdots,K,\\ \dfrac{\exp\left(\hat{g}_{2(l-K)-1}/\varepsilon\right)\tilde{\mathbf{c}}_{2(l-K)-1}+\exp\left(\hat{g}_{2(l-K)}/\varepsilon\right)\tilde{\mathbf{c}}_{2(l-K)}}{\exp\left(\hat{g}_{2(l-K)-1}/\varepsilon\right)+\exp\left(\hat{g}_{2(l-K)}/\varepsilon\right)},\\ l=K+1,K+2,\cdots,K+N.\end{cases} \tag{22}$$

Then, the common Riemannian gradient of $J(\mathbf{u})$ is given by

$$\hat{\mathbf{r}}_{q+1}=-\sum_{l=1}^{K+N}\alpha_{l,q+1}\mathbf{r}_l(\mathbf{u}_{q+1}), \tag{23}$$

where $\mathbf{r}_l(\mathbf{u})$ denotes the Gram-Schmidt orthogonalization of $\text{grad}J_l(\mathbf{u})$, and $\alpha_{l,q+1}$ denotes the weight factor of l-th objective function in the $q+1$-th iteration, which can be calculated as [4]

$$a_{l,q+1}=\frac{1}{1+\sum_{i\neq l}\dfrac{|\mathbf{r}_l(\mathbf{u}_{q+1})|^2}{|\mathbf{r}_i(\mathbf{u}_{q+1})|^2}}, l=1,2,\cdots,K+N. \tag{24}$$

3) Update the search direction: The third step is to calculate the next search direction via the conjugate direction method [14]. The updated rule for the next search direction is given by

$$\mathbf{d}_{q+1}=\hat{\mathbf{r}}_{q+1}+\eta_q\mathbf{d}_q^t, \tag{25}$$

where η_q and $\mathbf{d}_q^t$ denote the Polak-Ribiere parameter [13] and the output vector after the Riemannian transport operation, respectively. The role of the Riemannian transport operation is to project $\mathbf{d}_q$ into the same tangent space as $\hat{\mathbf{r}}_{q+1}$, which can be expressed as

$$\mathbf{d}_q^t=\mathbf{d}_q-\Re\left\{\mathbf{d}_q\odot\mathbf{u}_{q+1}^*\right\}\odot\mathbf{u}_{q+1}, l=1,2,\cdots,K+N. \tag{26}$$

In summary, the proposed algorithm is presented in Algorithm 1.

where

$$\hat{f}_{2l-1} \triangleq \mathrm{Re}\left(\mathbf{u}^H \tilde{\mathbf{a}}_{2l-1} + \tilde{b}_{2l-1}\right) + \sigma_{M,l}\sqrt{\Gamma_{M,l}}\tan\theta_t, \tag{13}$$

$$\hat{f}_{2l} \triangleq \mathrm{Re}\left(\mathbf{u}^H \tilde{\mathbf{a}}_{2l} + \tilde{b}_{2l}\right)\sigma_{M,l}\sqrt{\Gamma_{M,l}}\tan\theta_t, \tag{14}$$

$$\hat{g}_{2(l-K)-1} \triangleq \mathrm{Re}\left(\mathbf{u}^H \tilde{\mathbf{c}}_{2(l-K)-1} + \tilde{d}_{2(l-K)-1}\right) + \sigma_{P,(l-K)}\sqrt{\Gamma_{P,(l-K)}}\tan\theta_t, \tag{15}$$

$$\hat{g}_{2(l-K)} \triangleq \mathrm{Re}\left(\mathbf{u}^H \tilde{\mathbf{c}}_{2(l-K)} + \tilde{d}_{2(l-K)}\right) + \sigma_{P,(l-K)}\sqrt{\Gamma_{P,(l-K)}}\tan\theta_t. \tag{16}$$

In (13)–(16), $\tilde{\mathbf{a}}_{2l-1} = \mathbf{a}_l e^{-j\frac{\pi}{2}} - \mathbf{a}_l \tan\theta_t, \tilde{b}_{2l-1} = b_l e^{-j\frac{\pi}{2}} - b_l$, $\tilde{\mathbf{a}}_{2l} = -\mathbf{a}_l e^{-j\frac{\pi}{2}} - \mathbf{a}_l \tan\theta_t, \tilde{b}_{2l} = -b_l e^{-j\frac{\pi}{2}} - b_l$, $\tilde{\mathbf{c}}_{2(l-K)-1}=\mathbf{c}_{(l-K)}e^{-j\frac{\pi}{2}}-\mathbf{c}_{(l-K)}\tan\theta_t, \tilde{d}_{2(l-K)-1} = d_{(l-K)}e^{-j\frac{\pi}{2}} - d_{(l-K)}$, $\tilde{\mathbf{c}}_{2(l-K)} = -\mathbf{c}_{(l-K)}e^{-j\frac{\pi}{2}} - \mathbf{c}_{(l-K)}\tan\theta_t$ and $\tilde{d}_{2(l-K)} = -d_{(l-K)}e^{-j\frac{\pi}{2}} - d_{(l-K)}$. Accordingly, problem (10) can be reformulated as

$$\min_{\mathbf{u}} \left(f_1(\mathbf{u}), f_2(\mathbf{u}), \cdots f_{K+N}(\mathbf{u})\right) \tag{17a}$$

$$\text{s.t.} \quad |u_m| = 1, m = 1, 2, \cdots, M. \tag{17b}$$

Though all objective functions in (17) are smooth and differentiable after some mathematical operations, the non-convexity of the unit-modulus constraint in (17b) still poses challenges in solving the problem. In this paper, we treat the unit-modulus constraint with the manifold optimization methods [11]. Constraint (17b) can be regarded as an M-dimensional complex circle manifold, which is a manifold space of problem (17) characterized by

$$\mathcal{S}^M = \left\{\mathbf{u} \in \mathbb{C}^M : |u_m| = 1, m = 1, 2, \cdots, M\right\} \tag{18}$$

with the tangent space $T_{\mathbf{u}}S = \left\{\mathbf{p} \in \mathbb{C}^M : \Re\left\{p \odot u_m^*\right\} = \mathbf{0}_M, \forall m\right\}$. Problem (17) can be reformulated as an unconstrained optimization problem on the manifold space, which is given by

$$\min_{\mathbf{u}\in\mathcal{S}^M} J(\mathbf{u}), \tag{19}$$

where $J(\mathbf{u}) \triangleq (f_1(\mathbf{u}), f_2(\mathbf{u}), \cdots f_{K+N}(\mathbf{u}))$. We propose an MGD-RM based algorithm to effectively tackle the above problem. The core idea of this algorithm is to derive the gradient of each objective function based on the manifold space and then obtain a common gradient by weighted summation [4]. The common Riemannian gradient enables each objective function decrease in each iteration. Before starting the iterations, we set the initialized point $\mathbf{u}_1$ randomly on $\mathcal{S}^M$ and compute the initialized search direction $\mathbf{d}_1 = -\sum_{l=1}^{K+N} \alpha_l \mathbf{r}_l(\mathbf{u}_1)$. The main steps of the proposed algorithm at the q-th iteration are as follows:

1) Update the search point: We first update the current point over the manifold space $\mathcal{S}^M$. The search point in the next iteration is given by

$$\mathbf{u}_{q+1} = \mathrm{Retr}_{\mathbf{u}}\left(\mathbf{u}_q + \varsigma_q \mathbf{d}_q\right), \tag{20}$$

where $\mathrm{Retr}\left(\cdot\right)$ is specified by retraction operator mapping the value into the manifold space $\mathcal{S}^M$, and ς_q is a step-length, which can be obtained by using the Armijo backtracking line search method [13].

$$\min_{\mathbf{W},\mathbf{V},\mu} \quad \mu \tag{8a}$$

$$\text{s.t.} \quad (3),(4), \tag{8b}$$

$$\xi_a(P_a - P_a^*) \le \mu, \quad \forall a \in \{1,2\}. \tag{8c}$$

Since the CI constraints (3) and (4) are convex, problem (8) can be readily solved by standard optimization tools [2].

3.2 Reflection Design

Next, a reflecting design method is studied with fixed $\mathbf{W}$ and $\mathbf{V}$. The reflecting coefficients at the IRS are not strongly related to the transmit power at the MBS and PBS, but it impacts the quality of the received signals. We can observe that the optimal precoding matrices $\mathbf{W}$ and $\mathbf{V}$ in problem (8) always enable the CI constraints (3) and (4) to satisfy the inequality conditions. Thus, the values on the left-hand side of the constraints (3) and (4) are boosted by optimizing the reflecting coefficients for given precoding matrices, so that the MBS and PBS reduce their transmit power effectively by aligning the precoding matrices in the next iteration. Based on that, we construct the opposite forms of the left parts in constraints (3) and (4) as objective functions, which is given by

$$f_l(\boldsymbol{\theta}) \triangleq \begin{cases} |\mathrm{Im}(\lambda_l)| - \left[\mathrm{Re}(\lambda_l) - \sigma_{M,l}\sqrt{\Gamma_{M,l}}\right]\tan\theta_t, l = 1,2,\cdots K, \\ |\mathrm{Im}(\gamma_{l-K})| - \left[\mathrm{Re}(\gamma_{l-K}) - \sigma_{P,l-K}\sqrt{\Gamma_{P,l-K}}\right]\tan\theta_t, \\ l = K+1, K+2, \cdots K+N. \end{cases} \tag{9}$$

To this end, the reflecting design problem with $K+N$ is formulated as

$$\min_{\boldsymbol{\theta}} (f_1(\boldsymbol{\theta}), f_2(\boldsymbol{\theta}), \cdots f_l(\boldsymbol{\theta})) \tag{10a}$$

$$\text{s.t.} \quad 0 \le \theta_m \le 2\pi, m = 1,2,\cdots,M. \tag{10b}$$

Let $\mathbf{u} = [u_1, u_2, \cdots, u_M]^H$, where $u_m = e^{j\theta_m}, \forall m$. The l-th objective function (9) can be rewritten as

$$f_l(\mathbf{u}) \triangleq \begin{cases} \left|\mathrm{Im}\left(\mathbf{u}^H\mathbf{a}_l + b_l\right)\right| - \left[\mathrm{Re}\left(\mathbf{u}^H\mathbf{a}_l + b_l\right) - \sigma_{M,l}\sqrt{\Gamma_{M,l}}\right]\tan\theta_t, l = 1,2,\cdots K, \\ \left|\mathrm{Im}\left(\mathbf{u}^H\mathbf{c}_{l-K} + d_{l-K}\right)\right| - \left[\mathrm{Re}\left(\mathbf{u}^H\mathbf{c}_{l-K} + d_{l-K}\right) - \sigma_{P,l-K}\sqrt{\Gamma_{P,l-K}}\right]\tan\theta_t, \\ l = K+1, K+2, \cdots K+N, \end{cases} \tag{11}$$

where $\mathbf{a}_l = \frac{1}{s_l}\mathrm{diag}\{\mathbf{h}_{r,l}^H\}\mathbf{GWs}$, $b_l = \frac{1}{s_l}\mathbf{h}_l^H\mathbf{Ws}$, $\mathbf{c}_{l-K} = \frac{1}{x_{l-K}}\mathrm{diag}\{\mathbf{h}_{RI,l-K}^H\}\mathbf{GWs}$ and $d_{l-K} = \frac{1}{x_{l-K}}(\mathbf{g}_{l-K}^H\mathbf{Vx} + \mathbf{h}_{MI,l-K}^H\mathbf{Ws})$. Since the absolute value part of the objective functions causes difficulties in the problem solving, we convert (11) into the max function by exploiting the principle $|a| + b = \max(a+b, -a+b)$, and then approximate the l-th objective function to a smooth form by log-sum-exp inequality [2], which is given by

$$f_l(\mathbf{u}) \triangleq \begin{cases} \varepsilon\log\left(\left[\exp\left(\frac{\hat{f}_{2l-1}}{\varepsilon}\right) + \exp\left(\frac{\hat{f}_{2l}}{\varepsilon}\right)\right]\right), l = 1,2,\cdots K, \\ \varepsilon\log\left(\left[\exp\left(\frac{\hat{g}_{2(l-K)-1}}{\varepsilon}\right) + \exp\left(\frac{\hat{g}_{2(l-K)}}{\varepsilon}\right)\right]\right), l = K+1, K+2, \cdots K+N, \end{cases} \tag{12}$$

$$\min_{\mathbf{W},\mathbf{V},\boldsymbol{\theta}} \quad \sum_{k=1}^{K} \|\mathbf{w}_k\|^2 \tag{5a}$$

$$\text{s.t.} \quad (3), (4), \tag{5b}$$

$$0 \leq \theta_m < 2\pi,\ m = 1, 2, \ldots, M, \tag{5c}$$

where (5c) denotes the phase angle range of the reflecting elements of the IRS. On the other hand, the power minimization problem for the PBS can be formulated as

$$\min_{\mathbf{W},\mathbf{V},\boldsymbol{\theta}} \quad \sum_{n=1}^{N} \|\mathbf{v}_n\|^2 \tag{6}$$

$$\text{s.t.} \quad (3), (4), (5c).$$

We are concerned about the reduction of the total power consumption, as it is essential for deploying HetNet. Optimizing one of the above two objectives (5b) and (6) alone, however, does not optimize the other effectively. Therefore, it is necessary to optimize the two objectives jointly by MOO [12]. Note that unlike SOO, the solution of MOO is a set with multiple points corresponding to different weights of the objectives that govern their trade-off, and those points in the set realize Pareto optimality if there is no other point that improves at least one objective with given weights. Using the weighted Tchebycheff method [10], we can formulate a min-max MOO problem aiming at minimizing the transmit power at the MBS and PBS jointly as

$$\min_{\mathbf{W},\mathbf{V},\boldsymbol{\theta}} \quad \max\left\{\xi_1(P_1 - P_1^*), \xi_2(P_2 - P_2^*)\right\} \tag{7}$$

$$\text{s.t.} \quad (3), (4), (5c),$$

where $P_1 = \sum_{k=1}^{K} \|\mathbf{w}_k\|^2$, $P_2 = \sum_{n=1}^{N} \|\mathbf{v}_n\|^2$, P_1^* and P_2^* are the optimal solutions to the problems (5) and (6), respectively, and ξ_a satisfies $\xi_a \geq 0$ and $\sum_{a=1}^{2} \xi_a = 1$, which specifies the priority of the a-th objective with $a \in \{1, 2\}$. Changing the value of ξ_a can cope with different transmission strategies and the complete optimal set can be obtained by traversal activity. It is observed that problem (7) is non-convex due to the coupling between the precoding matrices and reflecting coefficients. To solve this problem, we use the AO method to decompose it into two sub-problems, namely precoding matrices optimization and reflection design. The details are presented in the following subsections.

3.1 Precoding Matrices Optimization

In this subsection, we focus on optimizing the precoding matrices $\mathbf{W}$ and $\mathbf{V}$ for given reflecting coefficients $\boldsymbol{\theta}$. By introducing an auxilliary variable μ, the precoding matrices optimization sub-problem is given by

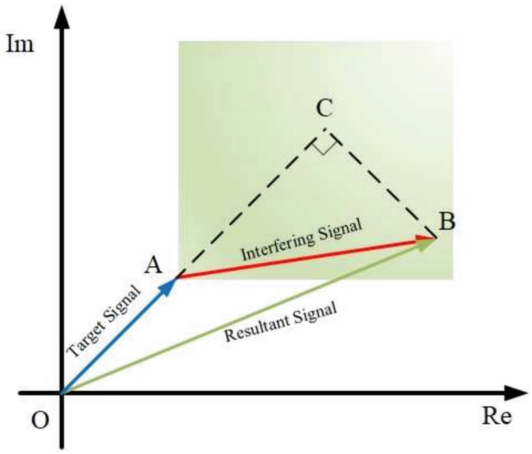


Fig. 2. CI region for $\mathcal{M}$-PSK constellation, where $\overrightarrow{AC}$ is the projection of $\overrightarrow{AB}$ onto $\overrightarrow{OA}$.

driving the received signal further away from the detection threshold of the signal constellation. The CI region for the considered $\mathcal{M}$-PSK constellation, which is in green color, is depicted in Fig. 2. The vector $\overrightarrow{OA}= \sigma_{M,k}\sqrt{\Gamma_{M,k}}s_k$ is the target signal, where $\Gamma_{M,k}$ refers to the SINR requirement of the k-th MUE, and $\overrightarrow{OB}= \lambda_k s_k = (\mathbf{h}_k^H + \mathbf{h}_{r,k}^H \mathbf{\Phi G})\mathbf{Ws}$ is the received signal of the k-th MUE without noise, where λ_k represents the received power level of the k-th MUE. It can be seen from Fig. 2 that the interfering signal $\overrightarrow{AB}$ is constructive as long as the resultant signal lies in the constructive region and drives the target signal beyond the detection threshold. Based on the geometric principle, the condition $\theta_{AB} \leq \theta_t$ should be satisfied in order for the resultant signal to be located in the constructive region, where θ_{AB} denotes the phase angle of the interfering signal and $\theta_t = \pi/\mathcal{M}$ [6]. Correspondingly, the conventional SINR constraints for MUEs can be converted to the following form to guarantee CI [6],

$$\left[\mathrm{Re}(\lambda_k) - \sigma_{M,k}\sqrt{\Gamma_{M,k}}\right] \tan\theta_t \geq \left|\mathrm{Im}(\lambda_k)\right|, \forall k \in \mathcal{K}. \tag{3}$$

In a similar way, the noise-free received signal for the n-th PUE can be formulated as $y_{P,n} = \mathbf{g}_n^H \mathbf{Vx} + \mathbf{h}_{PI,n}^H \mathbf{Ws} = \gamma_n x_n, \forall n \in \mathcal{N}$, where $\mathbf{h}_{PI,n}^H = \mathbf{h}_{MI,n}^H + \mathbf{h}_{RI,n}^H \mathbf{\Phi G}$ and γ_n represents the received power level of the n-th PUE. The CI constraints for PUEs can be expressed as

$$\left[\mathrm{Re}(\gamma_n) - \sigma_{P,n}\sqrt{\Gamma_{P,n}}\right] \tan\theta_t \geq \left|\mathrm{Im}(\gamma_n)\right|, \forall n \in \mathcal{N}. \tag{4}$$

where $\Gamma_{P,n}$ refers to the QoS requirement of the n-th PUE. Therefore, the power minimization problem for the MBS can be formulated as

MBS to communicate with MUEs by adjusting incident signals. It is worth noting that the strength of the reflected signal of the IRS originated from the PUEs is negligible due to the severe double fading, and as a result PUEs passively receive the interference signal reflected by the IRS originated from the MBS. The baseband equivalent channels from the MBS to the IRS, from the IRS to the k-th MUE, and from the MBS to the k-th MUE are denoted as $\mathbf{G}\in\mathbb{C}^{M\times N_M}$, $\mathbf{h}_{r,k}\in\mathbb{C}^{M\times 1}$, and $\mathbf{h}_k\in\mathbb{C}^{N_M\times 1}$ for $k\in\mathcal{K}$, respectively. We assume perfect knowledge of channel state information (CSI) in this paper. Let $\mathbf{s}\in\mathbb{C}^{K\times 1}$ denote the transmitted symbols of the MBS for the K MUEs, which are drawn from the normalized $\mathcal{M}$-ary phase shift keying (PSK) constellation. Define the phase shift matrix of the IRS as $\mathbf{\Phi}=\text{diag}\left(e^{j\theta_1},e^{j\theta_2},\ldots,e^{j\theta_M}\right)$ with $\boldsymbol{\theta}=[\theta_1,\theta_2,\ldots,\theta_M]^T$, where $\boldsymbol{\theta}$ is a set of reflecting coefficients. The received signal of the k-th MUE can be written as

$$y_{M,k}=(\mathbf{h}_k^H+\mathbf{h}_{r,k}^H\mathbf{\Phi}\mathbf{G})\mathbf{W}\mathbf{s}+n_{M,k} \tag{1}$$

with $k\in\mathcal{K}$, where $\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K]$ with $\mathbf{w}_k\in\mathbb{C}^{N_M\times 1}$ and $n_{M,k}\sim\mathcal{CN}(0,\sigma_{M,k}^2)$ denote the precoding matrix and zero-mean additive white Gaussian noise (AWGN) at the k-th MUE with variance $\sigma_{M,k}^2$, respectively. Denote the baseband equivalent channels from the PBS to the n-th PUE, from the MBS to the n-th PUE, and from the IRS to the n-th PUE as $\mathbf{g}_n\in\mathbb{C}^{N_P\times 1}$, $\mathbf{h}_{MI,n}\in\mathbb{C}^{N_M\times 1}$, and $\mathbf{h}_{RI,n}\in\mathbb{C}^{M\times 1}$ for $n\in\mathcal{N}$, respectively. The received signal of PUE n can be thus expressed as

$$y_{P,n}=\mathbf{g}_n^H\mathbf{V}\mathbf{x}+(\mathbf{h}_{MI,n}^H+\mathbf{h}_{RI,n}^H\mathbf{\Phi}\mathbf{G})\mathbf{W}\mathbf{s}+n_{P,n} \tag{2}$$

with $n\in\mathcal{N}$, where $\mathbf{V}=[\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_N]$ with $\mathbf{v}_n\in\mathbb{C}^{N_P\times 1}$ and $\mathbf{x}\in\mathbb{C}^{N\times 1}$ stand for the precoding matrix and the modulated symbols of N PUEs drawn from the same signal constellation as that employed by MUEs, respectively. In (2), the first term is the signal sent from the PBS, which includes the MUI, the second term is the ICI from both the MBS and IRS, and the last term is AWGN, satisfying $n_{P,n}\sim\mathcal{CN}(0,\sigma_{P,n}^2)$, where $\sigma_{P,n}^2$ is the noise variance.

3 Power Minimization Problem via Symbol Level Precoding

In this section, we study the transmit power minimization problem by jointly optimizing the precoding matrices at the MBS and PBS as well as the reflecting coefficients at the IRS for the IS scenario. Particularly, the MBS transmits precoded signals with the aid of the IRS, where the precoding matrices are designed based on the SLP principle. Since the MBS and PBS are aware of each other's information that will be sent. Therefore, the PBS can design the precoding matrix to validly transduce ICI and MUI into CI by SLP.

Unlike conventional precoding techniques, the SLP exploits the CSI and information symbols of the users to convert the harmful interference into the CI,

simulation results to demonstrate the performance advantages of our proposed algorithms. Finally, the paper is concluded in Sect. 6.

Notations: Boldface lower case and upper case letters denote vectors and matrices, respectively. $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real part and imaginary part of a complex scalar value, respectively. $|\cdot|$ and $\|\cdot\|$ are the magnitude of a scalar value and the norm of a vector, respectively. The transpose, conjugate, and transpose-conjugate operations are denoted by $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^H$, respectively. $\odot$ and $\mathrm{diag}(\cdot)$ denote the Hadamard product and diagonalization operation, respectively. $\bigtriangledown f(\mathbf{x})$ denotes the Euclidean gradient of the function $f(\cdot)$ with respect to $\mathbf{x}$.

2 System Model

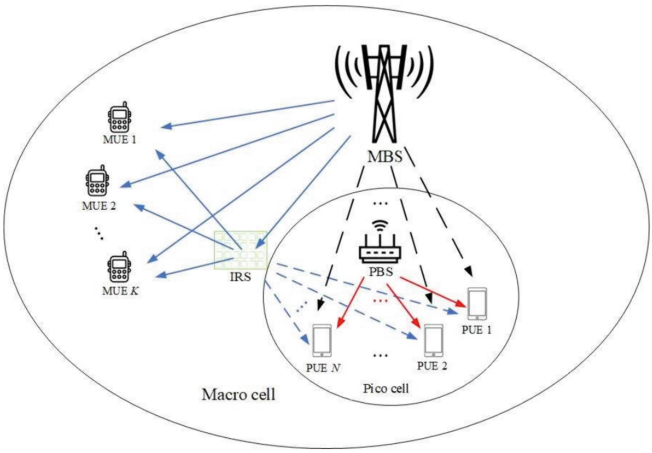


Fig. 1. The proposed HetNet and signal flows, where the blue and red solid lines represent the desired signals of MUEs sent by the MBS including those reflected by the IRS and the desired signals of PUEs sent by the PBS, respectively, and the blue and black dashed lines stand for the interference signals to PUEs reflected by the IRS and from directly the MBS, respectively. (Color figure online)

We consider a two-tier IRS-enhanced downlink HetNet, which consists of a macrocell and a picocell, as shown in Fig. 1. The MBS equipped with N_M antennas serves K single-antenna MUEs, whose indices are collected in $\mathcal{K} = \{1, \ldots, K\}$, and the PBS equipped with N_P antennas serves N single-antenna PUEs, whose indices are collected in $\mathcal{N} = \{1, \ldots, N\}$. Assume that both the MBS and PBS share the same frequency spectrum for communications and all MUEs are located outside the coverage of the PBS. Therefore, the received signals of MUEs will only be affected by the MUI due to the limited power of the PBS, while the received signals of PUEs will be corrupted by both ICI and MUI. The IRS with M reflecting elements is deployed near the MBS and assists the

Motivated by the above, we incorporate the IRS and SLP techniques in the design of HetNet to take their advantages for improving the system performance. By deploying the IRS in HetNet, the ICI can be alleviated, and the exploitation of MUI with SLP can be promoted by reconfiguring the propagation environment of the macrocell. To justify this finding, we exemplify a two-tier HetNet, where a multi-antenna MBS with the assistance of an IRS serves multiple single-antenna macro users (MUEs), and a multiple-antenna PBS serves multiple single-antenna pico users (PUEs), with the objective of minimizing the transmit power of all base stations. The considered HetNet assumes a wired connection between the MBS and the PBS, which implies that there is information sharing (IS) between the MBS and the PBS. In this system, the SLP is enabled at the MBS while the PBS also applies the SLP as the MBS to exploit the ICI from both the MBS and the IRS as well as the MUI. The main contributions of this paper are summarized as follows:

- We formulate a multi-objective optimization (MOO) problem by a joint optimization of the precoding matrices at the MBS and PBS as well as the reflecting coefficients at the IRS. However, it is non-convex with the high coupling of the precoding matrices and reflecting coefficients. To tackle this issue, we propose a productive alternating optimization (AO) based algorithm to solve the precoding matrices and reflecting coefficients separately. For solving the precoding matrices, we transform the MOO problem into a single-objective optimization (SOO) problem via the weighted Tchebycheff method [10] with given reflecting coefficients. The transformed problem is convex, which can be solved by standard optimization methods.
- In the reflection design, we aim to improve the quality of the received signals of both MUEs and PUEs by optimizing the reflecting coefficients with given precoding matrices, so that the MBS and PBS can reduce their transmit power effectively by aligning the precoding matrices in the next iteration. To this end, we formulate new objective functions associated with the received signal quality of MUEs and PUEs subject to unit-modulus constraints on the reflecting coefficients, which turns out to be a new MOO problem. Then, we propose a multiple gradient method on the Riemannian manifold (MGD-RM) based algorithm to solve the MOO problem, which guarantees the convergence to a suboptimal solution.
- Simulation results demonstrate the performance improvement of combining IRS and SLP in HetNet and the effectiveness of our proposed optimization algorithms. In particular, the PBS has the potential to reduce power consumption more with higher quality of service (QoS) requirement of MUEs in the proposed algorithm. In addition, the PBS can also achieve further power saving by introducing the IRS.

The rest of the paper is organized as follows. The proposed IRS-enhanced HetNet system model is described in Sect. 2. The considered power minimization problem is investigated in Sect. 3. The computational complexity analysis of the proposed algorithms is presented in Sect. 4. Section 5 illustrates extensive

annual growth rate of 50 percent over 2020 [5]. Meanwhile, some vertical industry applications are rapidly evolving along with the advancement of 5G, such as smart manufacturing, virtual reality, augmented reality, and automatic drive. Unlike ordinary wireless devices, these applications demand very different transmission rates. To satisfy the diverse rate requirements, one of the best solutions is to deploy picocells for private networks within the macrocell [16], resulting in a so-called heterogeneous network (HetNet). The picocells can be empowered by a pico base station (PBS), femto, relay, etc., which are deployed indoors normally and service the wireless devices with specific data services requirements. Their transmit power is generally less than 30 dBm and the signal coverage is from tens of meters to 300 m [9]. Since these low-cost nodes are within the reach of a macro base station (MBS) and share the same spectrum, they may suffer from severe intra-cell interference (ICI), deteriorating the quality of the received signals, and hence the system performance. Consequently, advanced intelligent wireless resource allocation and interference management (RAIM) technologies are essential to the success of HetNet deployment.

Recently, intelligent reflecting surface (IRS), which consists of a large number of passive and low-power reflective units, has attracted great attention in both academia and industry for its capability of customizing the wireless propagation environment [18]. Some initial studies on the optimization of IRS reflecting coefficients have been carried out for the IRS-enhanced wireless networks [8,11,17,19,20]. Specifically, the IRS aided downlink single-carrier multiple-input single-output (MISO) communication system was proposed, and the transmit power minimization problem was investigated by jointly optimizing the active beamforming at the base station and passive beamforming at the IRS [17]. The authors of [11] solved the reflection optimization problem by the majorization-minimization (MM) algorithm [15] and complex circle manifold (CCM) method [1] in multi-cell multiple-input multiple-output (MIMO) systems. In [20], a channel estimation protocol and reflection optimization problem for IRS-enhanced orthogonal frequency division multiplexing (OFDM) system were proposed. In [8], based on the instantaneous ON/OFF state information of the IRS reflection elements, the average, and instantaneous received signal power maximization problems were studied, respectively.

Another effective technique in dealing with interference, especially multi-user interference (MUI), is the transmitter precoding. With linear zero-forcing (ZF) precoding, the MUI can be eliminated, leading to satisfactory performance at high signal-to-noise ratio (SNR). The dirty paper coding (DPC), as a nonlinear precoding technique, suppresses the MUI by encoding transmit signals sequentially to approximate the Shannon limit [3]. While most conventional linear or nonlinear precoding techniques including the above-mentioned two are intended to suppress or eliminate the MUI, it has not been discovered until recently that the MUI signals can be made beneficial and exploitative by symbol level precoding (SLP), a.k.a., constructive interference (CI) precoding [7]. The SLP can push the received signals away from the detection threshold by converting the MUI into a constructive signal, enjoying magnificent improvement in terms of bit error rate (BER) and transmit power-saving [6].

Joint Symbol-Level Precoding and Reflecting Design for Heterogeneous Networks with Intelligent Reflecting Surface

Haoran Pang, Fei Ji, and Miaowen Wen(✉)

School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China
eemwwen@scut.edu.cn

Abstract. Recently, intelligent reflecting surfaces (IRS) emerge as an effective technique in saving the power consumption by customizing the wireless propagation environment. On the other hand, the symbol level precoding (SLP) technique provides a clever solution to interference exploitation by converting the multiuser interference (MUI) into a beneficial part of the desired signal. In this paper, we propose to jointly exploit IRS and SLP to cope with the power control and interference management issues in a heterogeneous network (HetNet). To this end, the IRS mainly assists the communication link from a macro base station (MBS) to macro users, and the SLP is employed by both the MBS and pico base station (PBS) to process MUI and intra-cell interference. We formulate a multi-objective optimization (MOO) problem to minimize the transmit power of MBS and PBS by jointly optimizing the precoding matrices at the MBS and PBS as well as reflecting coefficients at the IRS. Due to the non-convexity of this problem, the precoding matrices and reflecting coefficients are optimized alternately. In the precoding design, the MOO problem is transformed into a single-objective optimization problem via the weighted Tchebycheff method and then solved by standard optimization solvers with fixed reflecting coefficients. A multiple gradient descent on the Riemannian manifold based algorithm is proposed to obtain the local optimal solution for the reflecting design. Simulation results manifest a significant performance gain achieved by our proposed HetNet over the benchmarks.

Keywords: Heterogeneous network (HetNet) · Intelligent reflecting surfaces (IRS) · Symbol level precoding (SLP) · Multi-objective optimization (MOO)

1 Introduction

According to Ericsson's white paper, the fifth-generation (5G) market in the vertical industries is expected to reach 700 billion dollars in 2030 with a compound

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 441–456, 2024.
https://doi.org/10.1007/978-3-031-65126-7_37

18. Mukhlif, F., Noordin, K.A.B., Mansoor, A.M., Kasirun, Z.M.: Green transmission for C-RAN based on SWIPT in 5G: a review. Wirel. Netw. **25**, 2621–2649 (2019)
19. Hua, M., Li, C., Huang, Y., Yang, L.: Throughput maximization for UAV-enabled wireless power transfer in relaying system. In: 2017 9th International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–5. IEEE (2017)
20. Gul, O.M., Erkmen, A.M., Kantarci, B.: UAV-driven sustainable and quality-aware data collection in robotic wireless sensor networks. IEEE Internet Things J. **9**(24), 25150–25164 (2022)
21. Shen, L., Wang, N., Zhang, D., Chen, J., Mu, X., Wong, K.M.: Energy-aware dynamic trajectory planning for UAV-enabled data collection in MMTC networks. IEEE Trans. Green Commun. Netw. **6**(4), 1957–1971 (2022)
22. Di Franco, C., Buttazzo, G.: Energy-aware coverage path planning of UAVs. In: 2015 IEEE International Conference on Autonomous Robot Systems and Competitions, pp. 111–117. IEEE (2015)
23. Chai, S., Lau, V.K.N.: Multi-UAV trajectory and power optimization for cached UAV wireless networks with energy and content recharging-demand driven deep learning approach. IEEE J. Sel. Areas Commun. **39**(10), 3208–3224 (2021)

References

1. Kurs, A., Karalis, A., Moffatt, R., Joannopoulos, J.D., Fisher, P., Soljacic, M.: Wireless power transfer via strongly coupled magnetic resonances. Science **317**(5834), 83–86 (2007)
2. Kurs, A., Moffatt, R., Soljačić, M.: Simultaneous mid-range power transfer to multiple devices. Appl. Phys. Lett. **96**(4), 044102 (2010)
3. Achtelik, M.C., Stumpf, J., Gurdan, D., Doth, K.-M.: Design of a flexible high performance quadcopter platform breaking the MAV endurance record with laser power beaming. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5166–5172. IEEE (2011)
4. Chen, W., Zhao, S., Shi, Q., Zhang, R.: Resonant beam charging-powered UAV-assisted sensing data collection. IEEE Trans. Veh. Technol. **69**(1), 1086–1090 (2019)
5. Liu, J., Tong, P., Wang, X., Bai, B., Dai, H.: UAV-aided data collection for information freshness in wireless sensor networks. IEEE Trans. Wirel. Commun. **20**(4), 2368–2382 (2020)
6. Wang, J., Jiang, C., Wei, Z., Pan, C., Zhang, H., Ren, Y.: Joint UAV hovering altitude and power control for space-air-ground IoT networks. IEEE Internet Things J. **6**(2), 1741–1753 (2018)
7. Li, X., Yao, H., Wang, J., Xiaobin, X., Jiang, C., Hanzo, L.: A near-optimal UAV-aided radio coverage strategy for dense urban areas. IEEE Trans. Veh. Technol. **68**(9), 9098–9109 (2019)
8. Li, K.-R., See, K.-Y., Koh, W.-J., Zhang, J.-W.: Design of 2.45 ghz microwave wireless power transfer system for battery charging applications. In: 2017 Progress in Electromagnetics Research Symposium - Fall (PIERS - FALL), pp. 2417–2423 (2017)
9. Zhang, Q., Fang, W., Liu, Q., Jun, W., Xia, P., Yang, L.: Distributed laser charging: a wireless power transfer approach. IEEE Internet Things J. **5**(5), 3853–3864 (2018)
10. Zhu, K., et al.: Aerial refueling: scheduling wireless energy charging for UAV enabled data collection. IEEE Trans. Green Commun. Netw. **6**(3), 1494–1510 (2022)
11. Liu, K., Zheng, J.: UAV trajectory optimization for time-constrained data collection in UAV-enabled environmental monitoring systems. IEEE Internet Things J. **9**(23), 24300–24314 (2022)
12. Yang, G., Dai, R., Liang, Y.-C.: Energy-efficient UAV backscatter communication with joint trajectory design and resource optimization. IEEE Trans. Wirel. Commun. **20**(2), 926–941 (2020)
13. Li, J., et al.: Joint optimization on trajectory, altitude, velocity, and link scheduling for minimum mission time in UAV-aided data collection. IEEE Internet Things J. **7**(2), 1464–1475 (2020)
14. Zhan, C., Zeng, Y.: Completion time minimization for multi-UAV-enabled data collection. IEEE Trans. Wirel. Commun. **18**(10), 4859–4872 (2019)
15. Gong, J., Chang, T.-H., Shen, C., Chen, X.: Flight time minimization of UAV for data collection over wireless sensor networks. IEEE J. Sel. Areas Commun. **36**(9), 1942–1954 (2018)
16. Popović, Z., Falkenstein, E.A., Costinett, D., Zane, R.: Low-power far-field wireless powering for wireless sensors. Proc. IEEE **101**(6), 1397–1409 (2013)
17. Li, J., Yin, F., Wang, L., Cui, B., Yang, D.: Electromagnetic induction position sensor applied to anti-misalignment wireless charging for UAVs. IEEE Sens. J. **20**(1), 515–524 (2019)

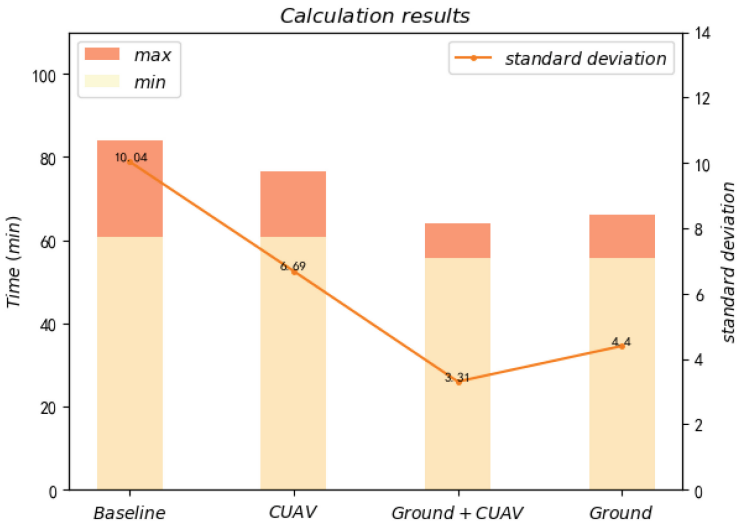


Fig. 6. Data under different modes.

mission UAV, indicating that the loads of each the mission UAV are close to each other. In addition, both the maximum mission completion time and the minimum mission completion time are smaller than other approaches.

4 Conclusion

This paper discusses the problem of minimizing the mission UAV information collection time for a sensor network consisting of mission UAVs, a charging UAV and a fixed charging pile, and sensors. For a known fixed sensor network, the sensors are first classified in an initial way using the *k-means++* clustering algorithm. Then, in order to balance the load of each mission UAV, the approximate time required by each cluster is calculated, and an algorithm is proposed to reduce the difference in the mission time required by each cluster. Path planning is performed within each cluster using a greedy algorithm, and the minimum mission completion time and the standard deviation of the mission completion time are compared under four different approaches. The results show that using charging UAVs combined with the fixed charging pile can minimize the maximum mission UAV completion time.

Table 1. Parameters of DJI UAV in experiments

UAV	MAVIC 3 Pro
Weight/kg	0.958
Flying time/min	43
Battery capacity voltage/mAh	5000
Maximum charging power/W	100
Battery max voltage/V	17.6

node is calculated by using the coordinates of all the sensors. The result of the experiment is shown in the Fig. 5.

The result shows that the mission completion time of all groups of the three methods is less than the *Baseline* except *Group* 1 of the $CUAV$ mode. The mission completion time in the $Ground{+}CUAV$ mode is shorter than the rest methods, which indicates that the use of the combination of the charging UAV and the ground pile is very effective and in line with our preliminary discussion.

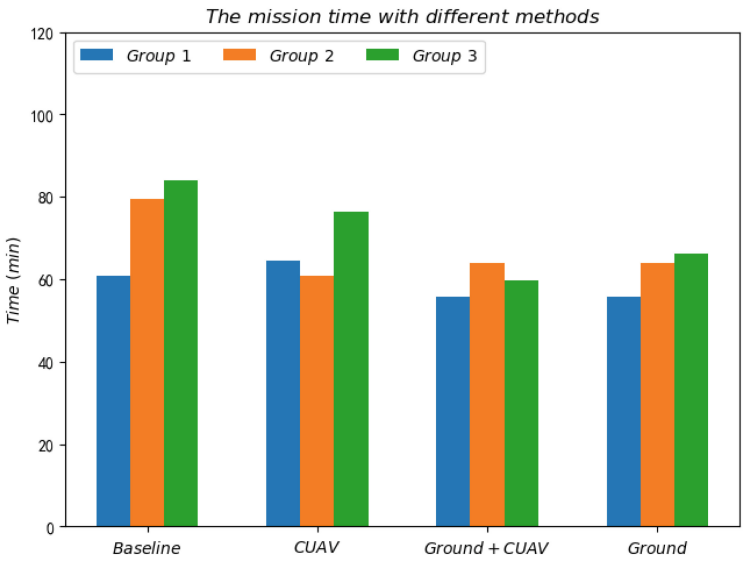


Fig. 5. Improved classification sensors.

Figure 6 shows the maximum and minimum completion time for the different groups under each mode, and the variance of the completion time in the groups under each mode is calculated. Using a combination of charging UAVs and fixed pile is able to minimize the standard deviation of the completion time for each

Algorithm 1. Path Planning

Input: position of charging pile at node p, sensors position set S, sensors distance matrix, mission UAV leftover lifetime T, UAV start point;
Output: Path P.

Initialize minDist $= \infty$, $i = 0$, total cost $R_t = 0$;
function FINDER(p, T, S)
  while $S \neq \emptyset$ **do**
    compute cost c_{ij}
    if $j \in visited$ **then**, $c_{ij} = \infty$
    end if
    find $j \leftarrow \arg\min\ c_{ij}$
    if $t_{i0} + t_{ij} > T_i$ **then**
      fly to charge or wait charging UAV flying to position i
      continue
    else
      $R_t \leftarrow R_t + c_{ij}, S \leftarrow S - j$, $visited \leftarrow j$, update T
    end if
  end while
  compute total cost (R_t)
  if cost(R_t) <minDist **then**
    minDist $\leftarrow$ cost(R_t)
  end if
end function

value, the mission UAV can call a charging UAV to replenish its energy, or travel to the fixed charging pile to recharge.

In the experiment, we utilize the parameters of the latest UAV from DJI, and its energy consumption index is shown in Table 1. Short charging distances are more suited for laser transmissions that can fully charge mission UAV in a short period of time. All groups, in all ways, have shorter time than BASELINE in all cases, except for group 1, which ended up with a slightly higher completion time when charging with the charging UAV alone than the completion time of the mission UAV returning to the base station for charging. This shows that the use of wireless charging technology is able to reduce the mission completion time of the last mission UAV.

The sensor network is divided into three groups, using the improved *k-means++* algorithm from earlier to compare the four methods. *Baseline* mode means the time used by the mission UAV to return to the base station with insufficient power before proceeding to the next sensor and completing the mission. $CUAV$ mode means the time used by the mission UAV to wait for the charging UAV from the selected fixed base station to arrive at the mission UAV's location to charge it and finally complete the mission when it is low on power. *Ground* mode means the time of the mission UAV taking to the selected fixed pile to recharge the mission UAV when its battery is low, then continue the task and return to the origin base station. $Ground + CUAV$ mode means using both *Ground* and $CUAV$ to reduce the mission time. Note that the virtual center

charged, as well as the charging UAV taking longer to reach the mission UAV, which leads to a longer time to complete the mission.

In the following, we will take the last, penultimate, and penultimate third of each cluster–until the number of sensors in a given cluster is all considered as an influencing factor for the virtual center node. As a comparison with the virtual center node formed by the entire sensor network together, the location of the fixed charging pile with the minimum time to complete the information gathering task is found out, as shown in Fig. 4. Note that even though Cluster 2 has only 9 sensors, the number of sensors is equal to 10 in Fig. 4, including all sensors to compute the virtual center node.

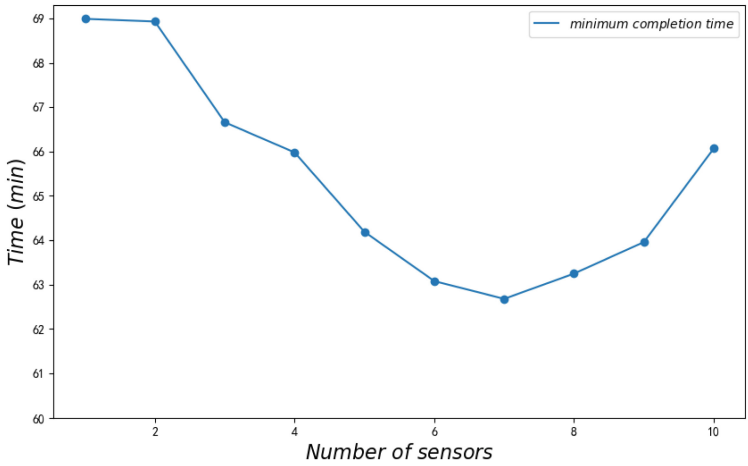


Fig. 4. Relationship between minimum completion time and number of sensors.

3 Experimental Results

In this section, we conduct experiments to verify whether the virtual central node can help reduce the energy consumption of charging UAVs, support more mission UAVs, and ultimately reduce the overall information collection time.

Algorithm 1 illustrates how to perform path planning for mission UAVs within a cluster in a better classified sensor network and select the appropriate charging method, i.e., using a charging UAV or a mission UAV going to a fixed charging pile to charge, to minimize the maximum mission. Before starting the information gathering task, the mission UAV departs from the origin base station and travels to the closest sensor in the cluster to perform the information gathering task, setting the distance to the current sensor to infinity if a sensor has already been visited, and placing the sensors that have already gathered information in the visited set. If the mission UAV's energy is below a certain

the current position to all clustering centers. Suppose that there is a set S and the coordinates of the clustering center point are (x_i, y_i), and the objective is to find the point (x, y) such that the objective function $F(x, y)$ is minimized, which is given as follows:

$$F(x, y) = \sum_{i=1}^{n} \sqrt{(x - x_i)^2 + (y - y_i)^2} \tag{1}$$

Assume that there are n clustering centers, Eq. (1) represents the sum of the distances to all clustering centers. The gradient descent algorithm is used to solve this problem by computing the derivative of the objective function with respect to the variables and updating the values of the variables in the opposite direction of the gradient, gradually approaching the optimal solution. The specific steps are as follows:

$$x = x - \alpha \frac{\partial F}{\partial x} \tag{2}$$

$$y = y - \alpha \frac{\partial F}{\partial y} \tag{3}$$

- *Step 1.* Initializing the coordinates to be optimized (x, y).
- *Step 2.* Calculating the gradient of the objective function with respect to x and y, calculating the partial derivative of $F(x, y)$ with respect to x and y.
- *Step 3.* Updating the values of the variables according to the direction of the gradient and the learning rate ∂. Equation (2) and Eq. (3) represent the updated values of x and y along the direction, respectively.
- *Step 4.* Repeating *step 2* and *step 3* until a predetermined number of iterations is reached or certain termination conditions are qualified.

The gradient descent algorithm gradually approaches the minimum point of the objective function by updating the values of the variables and finds the position of the point with the smallest distance from all the prime points. The location is defined as the virtual center node, the fixed charging pile location is set at the virtual center node, and the charging UAV is set at the virtual center node at the beginning, thus the charging UAV can reach the mission UAV faster, and the mission UAV can reach the fixed charging pile for charging faster.

2.3 Virtual Center Node Optimization

In previous work, it has been discussed how to classify the sensor network in a rational way and a virtual center node has been proposed as a fixed charging pile and the node is found using the gradient descent method. However, even the mission UAV is fully charged when it departs from the origin base station, the mission UAV often needs to be recharged at a later stage. Therefore, taking all points into account when finding the virtual center node may result in the fixed charging pile being farther away from the mission UAV that needs to be

determine the next sensor to be accessed. When equal distance sensors are encountered, the sensor closer to the base station is selected.
- *Step 2*. Calculating the center point of each cluster that has the smallest distance from the points in current cluster.
- *Step 3*. Calculating the standard deviation of the traversal length of each cluster. If the standard deviation is greater than the threshold, the cluster with the maximum traversal length will allocate one sensor to each other cluster, and the allocated sensor is the one closest to the other clusters.
- *Step 4*. Repeating *step 1*, *step 2* and *step 3*, until the standard deviation is less than the threshold or after updating k-1 times.

As shown in Fig. 3, there are two sensors assigned to a new cluster i.e., transferring from Cluster 3 to Cluster 1 and Cluster 3 to Cluster 2, respectively, trying to balance the workload of each mission UAV.

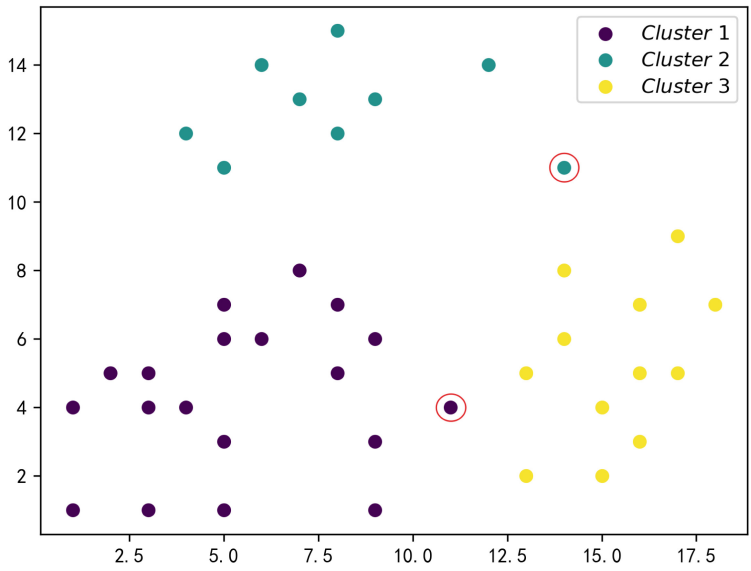


Fig. 3. Improved classification sensors.

The comparison between Fig. 2 and Fig. 3 shows that further classification enables the classes closer to the base station to be assigned more sensors, reducing the load on the mission UAVs responsible for long distance cluster. It also benefits the path planning for charging UAVs, which will be discussed in Sect. 3.

2.2 Virtual Center Node Selection

In order to find the virtual center node, note that it is the point with the smallest sum of distances from all clustering centers, which can be treated as an optimization problem. Define the objective function as the sum of the distances from

- *Step 2.* Calculating the shortest distance between each sample and the currently existing clustering center, denoted as $D(x)$, and calculate the probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$ of each sample being selected as the next clustering center, and select the next clustering center according to the roulette wheel method.
- *Step 3.* Repeating the second step until k clustering centers are found.
- *Step 4.* Assigning each sensor to a cluster, calculating the Euclidean distance between each sensor and clustering center, and assigning it to the cluster corresponding to the closest clustering center.
- *Step 5.* Updating the clustering center of each cluster to the average of all points in that cluster.
- *Step 6.* Iterating *step 4* and *step 5* repeatedly, if the distance between the new clustering center and the previous clustering center is less than a certain threshold, the clustering algorithm is considered to have achieved the desired result and the algorithm stops, otherwise continue iterating.

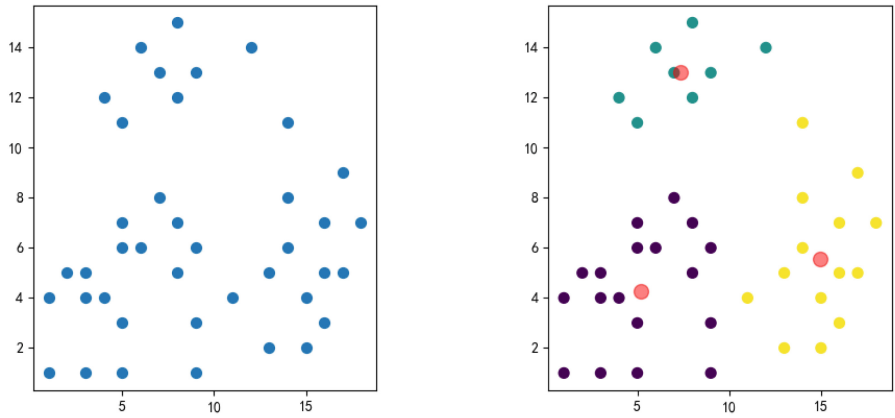


Fig. 2. Sensor classification effect with k-means++.

The use of the *k-means++* algorithm enables the selection of more appropriate initial clustering centers and reduces the influence of the clustering centers on the results due to randomly taken values.

Figure 2 shows the result of *k-means++* classification algorithm. Different from the *k-means* algorithm, by selecting the initial points, *k-means++* significantly improves the classification results. However, in the sensors network, it may lead to a more uneven allocation of nodes to each cluster. The procedure of the improved classification method is summarized as follows:

- *Step 1.* Starting from the base station, traversing the sensors of each cluster and returning the base station. The point closest to the starting point of each cluster is the first sensor to be accessed, and a greedy strategy is used to

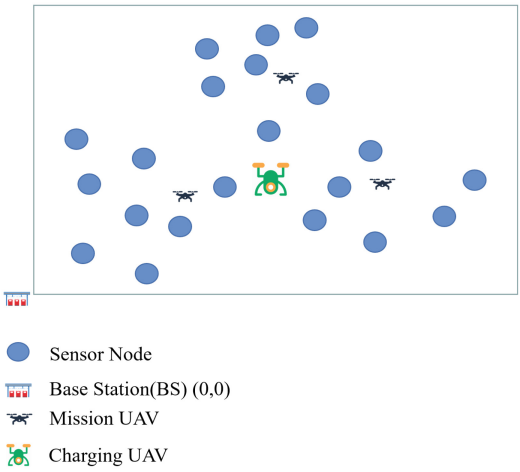


Fig. 1. Illustration of network components.

- *Step 3.* Calculating the variance of the mission time for each cluster and setting a threshold, if the variance is larger than the threshold, assigning a sensor from the cluster with the largest mission time to each of the other clusters, which is the one with the smallest distance from the minimum point of the other clusters' center, and re-calculating the variance until it is smaller than the threshold or after n rounds of loops.
- *Step 4.* The point with the smallest distance from the minimum point in all categories is found and is called the virtual center node.
- *Step 5.* Path planning is performed for the order in which the UAVs in a specified cluster collect sensor information, and the UAVs in each cluster start from the sensor closest to the origin base station and use a greedy strategy to select the next sensor, and if two sensors with the same distance are encountered, the sensor closer to the origin base station is prioritized.
- *Step 6.* The charging UAV flies to the virtual center node and arrives at the location of the mission UAV in time when the mission UAV is low on power, or the mission UAV goes to the fixed charging pile to replenish the power using wireless charging technology.

2.1 Sensor Clustering

The *k-means* algorithm is a simple and practical classification algorithm, but the results are easily influenced by the selection of the initial points. A poorly chosen initial clustering center chosen at random may have a relatively large impact on the results. Therefore, we utilize the *k-means++* algorithm to cluster the sensors. The clustering sensors is as follows:

- *Step 1.* Randomly selecting a sensor from the sensor network as the first cluster center.

the UAV speed and path planning problem by using a continuous convex approximation method and a genetic algorithm. Chai et al. [23] propose a solution for joint path optimization and wireless communication network for multiple UAVs, and obtained a decentralized solution with reduced complexity by mean-field equilibrium analysis, and the simulation results show that the proposed solution is better than the existing methods. All these methods present good solutions for path planning of UAVs in specific scenarios. Different from previous works, this paper consider not only the paths of the mission UAVs, but also the paths of the charging UAV and the deployment of the fixed charging piles. Unreasonable deployment of fixed charging pile can lead to longer waiting time for the charging UAV to charge mission UAVs and longer time for mission UAVs to travel to charging pile, resulting in longer time for the last mission UAV to complete the mission.

The rest of this paper is organized as follows. In Sect. 2, we improve the *k-means++* algorithm by dividing the sensors into proper different clusters to balance the load of the mission UAVs, and use the gradient descent method to find the virtual node. In Sect. 3, we conduct experiments to evaluate the results. Finally the paper is concluded in Sect. 4.

2 System Model

The limited power capacity of UAVs necessitates the need for an efficient approach to collect information from a large sensor network. To achieve this, the sensor network is classified using the *k-means++* algorithm, providing an initial segmentation of the sensors. If the sensors are farther away from the origin base station, the UAV will consume more energy in the process of traveling to the long-distance sensors to collect information. Thus, we assign fewer sensors to the UAV responsible for long-distance sensor data collection than the UAV responsible for the closer sensor data collection to balance the load of the UAVs as much as possible. After the allocation is completed, each UAV uses a greedy strategy to select a suitable path for the sensors in the responsible area to perform the information collection task. The network configuration is illustrated in Fig. 1.

The overall idea of the whole paper from sensor classification, the concept of minimum point, reclassification of individual sensors, the concept of the virtual center node, mission UAVs path planning approach, and mission UAVs charging method is as follows:

- *Step 1.* The sensors are classified using the *k-means++* algorithm, which reduces the negative impact of randomly selecting the initial clustering centers compared to the *k-means* algorithm.
- *Step 2.* The coordinate point in each category with the smallest distance from all sensor locations, called the minimum point, is found by gradient descent, and the mission time required to collect sensor information for each cluster is calculated using the greedy algorithm.

1.3 Related Work

The energy of the UAV is an important limiting factor for the UAV to collect information [11]. For some large wireless sensor networks, a single UAV cannot accomplish the task of collecting information from all sensors in a short period of time. Yang et al. [12] use UAV to provide IoT communications and maximize network energy efficiency by optimizing the trajectory of the UAVs and other practical. Li et al. [13] present the main concern for information collection by UAVs in emergency situations is the time to complete the mission, and propose a method for UAVs to help information collection from ground users, optimizing the trajectory, altitude and speed of the UAVs. Zhan et al. [14] propose a wake-up scheduling approach using jointly optimized UAV flight trajectories and sensors to minimize the UAV information collection time and thus reduce the energy consumption of the UAV. Jie et al. [15] considered the problem of minimizing the time for a UAV to complete a information collection task in a network of sensors on a straight line.

Due to the energy constraint of UAVs, the use of wireless charging is an promising way to address the lack of capability in UAV missions. Wireless charging technology is developing rapidly and is being used in many areas. There are many kinds of wireless charging technologies, including non-directional RF energy transfer [16], electromagnetic induction [17]. Moreover, some new wireless charging technologies emerge in recent years, such as laser power transfer, distributed laser charging [9], simultaneous wireless information and power transfer [18]. Hua et al. [19] use a relay network with malicious amplification and forwarding to replenish the UAV and maximize network throughput. Zhang et al. [9] propose a multi-module distributed laser charging model and derived the maximum power transfer efficiency, which is shown to depend on the power supplied by the transmitter, the laser wavelength, the transmission distance et al. Zhu et al. [10] propose the concept of using one UAV to charge another UAV on a controlled mission by means of wireless charging technology, with the aim of minimizing the time it takes for the mission UAV to complete the mission. Mission UAVs can be used to perform functions such as collecting sensor information or communicating with sensors for data.

Gui et al. [20] use a UAV-assisted approach to collect data from the machine. However, as the power of the UAV is limited, which machine to visit is constrained by the remaining battery power, the location of the machine, and the quality of the data. The problem of mission UAV flight path planning can be converted to a TSP problem in tasks where the goal is to minimize the time of collecting data by the mission UAV. A lot of previous work has discussed the problem of electricity shortage. Shen et al. [21] propose clustering-based service strategies and adynamic trajectory planning algorithms, which dynamically adjust the hovering position of the UAV providing the data collection task to maximize the data collection efficiency. Di et al. [22] propose an energy-aware path planning algorithm to maximize the reduction of energy consumption. Liu et al. [11] consider the flight speed of the UAV, the hovering position and access sequence, the information age, and the recorded energy of the UAV, and solve

In sensor networks where multiple UAVs collect information from sensor nodes, we assume that UAVs cannot reach all sensor nodes at once. We discuss the minimum mission time for the following three scenarios: the UAV returns to the initial base station in time to recharge, returns to the fixed base pile and uses both the fixed base pile and the recharging UAV at the same time.

The goal is to minimize the time taken by the mission UAVs to complete their tasks. As the time taken by the mission UAVs to collect information at the sensors is much less than the flight time, we ignore it in order to simplify the computation. In the paper, we first equalize the amount of tasks for each mission UAV. Since there is only one charging UAV, the mission UAV may wait long time for the charging UAVs' arrival. Therefore, the location of the fixed charging pile is crucial to enable the mission UAVs reach the fixed charging pile faster. The mission UAVs and charging UAV find suitable scheduling strategies to minimize the completion time of the last mission UAV to improve the efficiency of the whole system.

1.2 Main Contributions

In this paper, we consider the location of the fixed Charging piles, data collection from sensors by mission UAVs, and path planning for charging UAVs. The main contributions are as follows:

- *Rational categorization of the sensor nodes.* In previous work, the sensors are usually distributed equally according to their number, which does not take into account the location of the sensors from the origin base station, resulting in a larger load for UAVs that are far from the origin base station. In this paper, distance from the origin base station is used as one of the factors to categorize the sensors, and each mission UAV is responsible for collecting a class of sensor information.
- *Proposition of the concept of virtual center.* While performing information collection tasks, the mission UAVs can call the charging UAVs, especially when they do not have enough power to return to the base station. However, the charging UAVs may not be able to fly from one mission UAV to another in time if the distances between them are far apart. By gradually moving closer to the virtual center while performing the tasks, the charging UAV can reach the mission UAV in time. Additionally, establishing the charging pile at the virtual center node can reduce the travelling time for the mission UAV to charge.
- *Path optimization for the mission UAV.* In each cluster, the mission UAV need to reach every sensors in the cluster once. The path for the mission UAV is optimized to reduce the flight time between the charging UAV and the mission UAVs, as well as the time of mission UAVs flying to the charging pile.

performance is becoming a prominent future trend. For example, magnetic resonance coupling technology works on the principle of resonant coupling, about 40% of non radiant power transmission efficiency can be achieved between two strongly coupled objects when the distance exceed 2 meters [1]. When applied to smaller devices, magnetic resonance coupling can achieve an impressive energy transfer efficiency of 60%, presenting a feasible avenue for using wireless charging technology to recharge UAVs [2]. Another approach to enable wireless charging over long distances involves using a laser beam from a ground station to charge the UAV while it is in flight [3,4].

Utilizing the aforementioned wireless charging technology to replenish the energy of UAVs offers numerous advantages, extending the UAVs' usability and enabling them to perform various tasks more effectively, such as information collection [5], serving as aerial bridging stations to support communications [6,7]. However, in scenarios where UAVs lack sufficient energy to complete their missions, the current mainstream approach involves returning to the base station for recharging, setting up fixed wireless charging piles which allows UAVs to recharge in-flight. Nonetheless, the issue of energy efficiency remains unsolved. The challenge arises because fixed base stations or charging piles are stationary, requiring UAVs to expend significant energy to travel to these fixed locations for recharging. This process not only adds to the mission time but also diminishes the energy efficiency of the UAV's overall operation. Consequently, this limitation hinders the UAV's ability to operate optimally and efficiently during critical missions. As a result, finding innovative solutions to address the energy efficiency problem is imperative to enhance the UAV's performance and extend its operational capabilities.

Recently, with the development of wireless charging technology, modules consisting of receiving antenna arrays, rectifiers and power management circuits have been implemented on UAVs, receiving microwave power and converting it to DC power to charge the UAVs [8]. Distributed laser charging can also be used as a wireless charging solution [9]. These new wireless charging methods make it possible to use one UAV to charge another [10]. UAVs that perform tasks such as data collection are known as mission UAVs, and the one that charges mission UAVs is known as charging UAV. The use of chargeable UAVs to charge mission UAVs eliminates the need for mission UAVs to travel to a base station or fixed base pile for recharging, and improves energy utilization without requiring the mission UAVs to travel to and from the base station. Instead, they can wait for the rechargeable UAVs to arrive at their current location to replenish their energy. The cost of constructing a fixed base station can also be reduced in the case of charging UAVs only.

In this paper, we discuss the UAV charging scenario in the context of wireless sensor networks collecting outdoor information. In this application scenario, the sensor nodes are distributed in a fixed area. We use multiple mission UAVs to collect the sensor network information, and the mission UAVs can either travel to a fixed charging pile to charge, or one charging UAV can charge the mission UAVs to minimize the time for the mission UAVs to complete the mission.



Wireless Charging Based Sensor Network Information Collection Through Unmanned Aerial Vehicles (UAVs)

Guoxin Xu[1], Jiawen Zhao[2], and Xuehe Wang[1(✉)]

[1] Sun Yat-sen University, Zhuhai 519082, China
xugx8@mail2.sysu.edu.cn, wangxuehe@mail.sysu.edu.cn
[2] Mai-Yuan Construction Group Corporation Limited, Shandong 272200, China

Abstract. In recent years, the proliferation of commercial unmanned aerial vehicles (UAVs) has led to the widespread adoption of wireless charging technology, fostering their increasing application in various domains. This trend has made UAVs increasingly suitable for replacing conventional information collection vehicles in wireless sensor networks, particularly in scenarios where sensors possess both sensing and communication capabilities. In this paper, we discuss the minimum information collection time for a large-scale wireless sensor network consisting of multiple mission UAVs and one charging UAV. The mission UAV is responsible for collecting data from each sensor, and the wireless charging pile is used to replenish power to the mission UAVs, in order to minimize the completion time of the mission UAVs. First, a modified *k-means++* clustering algorithm is utilized to assign sensor nodes with the number of clusters equal to the number of mission UAVs. The process of collecting sensor information within a certain range by the mission UAV is modeled as the traveling salesman problem (TSP). Then, we propose the concept of virtual center node which is found by using a gradient descent algorithm. We compare the performance of using a charging UAV to charge a mission UAV with three other methods, i.e., establishing a fixed charging pile to charge mission UAVs, and mission UAVs go back to original base station for charging, and combining both the charging UAV and a fixed charging pile. The experimental results show that the combination of a charging UAV and a charging pile outperforms the other three methods in reducing the completion time of mission UAV.

Keywords: UAV · Wireless Charging · Clustering · Data Collection

1 Introduction

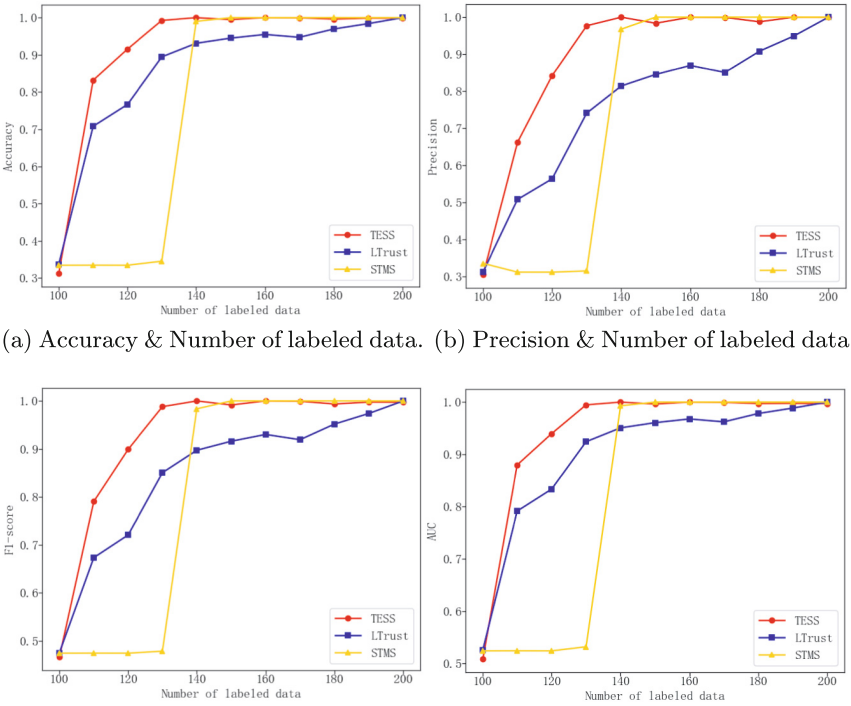
1.1 Background

In recent years, wireless energy transmission technology has seen significant advancements. Due to the relatively limited power capacity of unmanned aerial vehicles (UAVs), the utilization of wireless charging technology to enhance UAV

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 426–440, 2024.
https://doi.org/10.1007/978-3-031-65126-7_36

References

1. Sun, Y., Peng, M., Zhou, Y., Huang, Y., Mao, S.: Application of machine learning in wireless networks: key techniques and open issues. IEEE Commun. Surv. Tutor. **21**(4), 3072–3108 (2019)
2. Ning, Z., Dong, P., Wang, X., et al.: Mobile edge computing enabled 5G health monitoring for Internet of Medical Things: a decentralized game theoretic approach. IEEE J. Sel. Areas Commun. **39**(2), 463–478 (2021)
3. Nguyen, C.T., Huynh, N.V., Chu, N.H., et al.: Transfer learning for future wireless networks: a comprehensive survey. Proc. IEEE **110**(8), 1073–1115 (2022)
4. Wang, X., Ning, Z., Guo, S., et al.: Dynamic UAV deployment for differentiated services: a multi-agent imitation learning based approach. IEEE Trans. Mob. Comput. **22**(4), 2131–2146 (2023)
5. Ning, Z., Yang, Y., Wang, X., et al.: Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE Trans. Mob. Comput. **22**(5), 2628–2644 (2023)
6. Odena, A.: Semi-supervised Learning with Generative Adversarial Networks (2016). https://doi.org/10.48550/arXiv.1606.01583
7. Wang, X., Ning, Z., Guo, L., et al.: Mean-field learning for edge computing in mobile blockchain networks. IEEE Trans. Mob. Comput. 1–17 (2022)
8. Hu, X., Ning, J., Yin, J., Yang, J., Adebisi, B., Gacanin, H.: Efficient malicious traffic classification methods based on semi-supervised learning. In: 2022 9th International Conference on Dependable Systems and Their Applications (DSA), pp. 230–235 (2022). https://doi.org/10.1109/DSA56465.2022.00039
9. Gao, Y., Liu, Y., Jin, Y., Chen, J., Wu, H.: A novel semi-supervised learning approach for network intrusion detection on cloud-based robotic system. IEEE Access **6**, 50927–50938 (2018)
10. Breiman, L., Friedman, J., Olshen, R., et al.: Classification and regression trees. Encyclopedia Ecol. **4**(3), 582–588 (1984)
11. Zhao, R., Tang, T., Gui, G., Xue, Z.: A lightweight semi-supervised learning method based on consistency regularization for intrusion detection. In: IEEE International Conference on Communications, ICC 2022, Seoul, Republic of Korea, pp. 3124–3129 (2022). https://doi.org/10.1109/ICC45855.2022.9838567
12. Cai, S., Han, D., Li, D.: A feedback semi-supervised learning with meta-gradient for intrusion detection. IEEE Syst. J. **17**(1), 1158–1169 (2023)
13. Souissi, I., Ben Azzouna, N., Ben, S.L.: A multi-level study of information trust models in WSN-assisted IoT. Comput. Netw. **151**, 12–30 (2019)
14. Wu, X., Huang, J., Ling, J., Shu, L.: BLTM: beta and LQI based trust model for wireless sensor networks. IEEE Access **7**, 43679–43690 (2019)
15. Shao, N.M.K., Luo, F., Liu, Z.: Normal distribution based dynamical recommendation trust model. J. Softw. **23**(12), 3130–3148 (2012)
16. Han, G., He, Y., Jiang, J., Wang, N., Guizani, M., Ansere, J.A.: A synergetic trust model based on SVM in underwater acoustic sensor networks. IEEE Trans. Veh. Technol. **68**(11), 11239–11247 (2019)
17. Du, J., Han, G., Lin, C., Martínez-García, M.: LTrust: an adaptive trust model based on LSTM for underwater acoustic sensor networks. IEEE Trans. Wirel. Commun. **21**(9), 7314–7328 (2022)



(a) Accuracy & Number of labeled data. (b) Precision & Number of labeled data.

(c) F1-score & Number of labeled data. (d) AUC & Number of labeled data.

Fig. 6. Comparison of performance under different number of labeled data.

case of a small amount of data. Therefore, they can not accurately distinguish normal nodes from malicious nodes.

To sum up, in underwater scenarios with insufficient amount of labeled data, TESS can make use of the advantages of SGAN networks to fully mine the characteristics of labeled data and unlabeled data in the game between generator and discriminator. TESS improves the classification performance of the discriminator by generating fake data, so as to more accurately identify malicious nodes and maintain network security.

6 Conclusion

Aiming at the scenario of insufficient data of underwater wireless sensor networks, we study how to use a small amount of labeled data and large amounts of unlabeled data to improve the detection ability of trust model, accurately identify malicious nodes, and ensure the security of underwater wireless sensor networks. In this paper, we propose a Semi-Supervised learning based trust evaluation method (TESS), which uses the game training process of Generating Adversarial Networks to improve the discriminator classification accuracy and accurately identify malicious nodes.

Table 3. The comparison of accuracy under Grey-hole attack between different trust models

Models	Accuracy
TESS	0.8700
LTrust	0.3589
STMS	0.8632

As can be seen from Table 3, TESS has the highest detection accuracy among the three models, reaching 87.00%. STMS performed slightly worse than TESS, while LTrust has an accuracy of just 35.89%. Combined with the results in Fig. 6, it can be shown that TESS based on Semi-Supervised learning can achieve higher accuracy than the supervised method in scenarios with only a small amount of labeled data. TESS reduces the amount of data required and can better prevent underfitting. TESS is more suitable for trust evaluation of underwater wireless sensor network nodes.

Proportion of Different Malicious Nodes. To compare the performance of TESS with LTrust and STMS at different proportions of malicious nodes, the average accuracy, precision, F1-score, and AUC of the three models are tested under Bad-mouthing attacks. The proportion of malicious nodes in the test set is $[0.05, 0.5]$, and the experimental results are shown in Table 4.

According to the results in Table 4, TESS performs better overall when the model is trained with the same labeled data. Both LTrust and STMS have an accuracy of around 70%. While LTrust and STMS have an precision of 1, the F1-score is below 0.1 and the AUC value is also around 0.5. The performance of LTrust and STMS on normal nodes is better, but the overall performance is lower than TESS. And the classification ability of the model is not high when there is less labeled data. When only a small portion of the data acquired underwater is labeled, TESS can train the model with a large amount of unlabeled data and labeled data at the same time. The generator in SGAN can generate fake samples that are close to the real samples to participate in the training while working against the discriminator. It can improve the discriminator's recognition ability. However, LSTM algorithm adopted by LTrust and SVM algorithm adopted by STMS can only use labeled data for training, which is prone to underfitting in the

Table 4. Comparison of performance under Bad-mouthing between different trust models

Models	TESS	LTrust	STMS
Accuracy	0.9728	0.7083	0.7050
Precision	0.9605	1.0000	1.0000
F1-score	0.9544	0.0541	0.0328
AUC	0.9659	0.5139	0.5083

increases to 110, the classification effect based on semi-supervised learning under both attacks quickly increases to more than 80%. The accuracy and F1-score of supervised classifiers with the same structure is still significantly lower than that of semi-supervised classifiers. Under the Bad-mouthing attack, until the number of labeled data is increased to 500, the supervised classifier's accuracy and F1-score still do not exceed that of the semi-supervised classifier. Under the Good-mouthing attack, the supervised classifier can only achieve an accuracy of more than 80% and an F1-score if it reaches 150 labeled data. In the case of different attacks and the same amount of labeled data, TESS based on Semi-Supervised learning detects better than supervised classifiers of the same structure, which meets the need for scarce data available underwater. TESS compensates for supervised classifiers' reliance on large amounts of labeled data. It utilizes the game training process of GAN to fully mine the information contained in the unlabeled data and improve the performance of the discriminator. TESS greatly improves the accuracy of detecting malicious nodes and ensures network security.

5.2 Comparison of TESS with Other Models

In this section, the accuracy index performance of TESS, LTrust and STMS is compared. For fair comparison, all three trust models use the same labeled data set and are tested on the same test set.

Different Amount of Labeled Data. We test the performance of TESS, LTrust and STMS in Bad-mouthing, Good-mouthing and Grey-hole attacks. The number of nodes in the experiment is set to 100, and the proportion of malicious nodes in the test set is 30%.

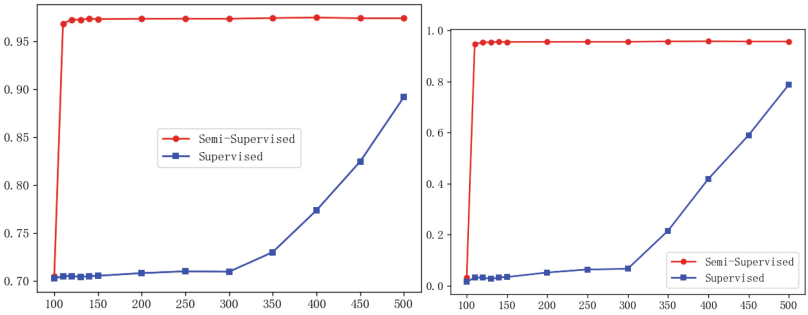
The performance of the three models varies with the amount of labeled data under the Good-mouthing attack, as shown in Fig. 5. On the premise of minimizing the dependence of the trust model on labeled data, $[100, 200]$ is selected as the experiment interval. The proportion of malicious nodes in the test set is 30%.

As we can see from Fig. 6, the initial accuracy and precision of the three models are about 30%, and the F1-score and AUC are about 0.5. This is because the model underfits when the number of labeled data is too small. When the number of labeled data gradually increased to 130, four indicators of TESS reach more than 90%. Although the performance of LTrust is gradually improving, it still does not perform as well as TESS, with accuracy and F1-score not exceeding 90%. STMS only improves its performance to close to 1 when the number of labeled data reaches 140. The combined results of the four performance indicators demonstrate that TESS can significantly improve the detection performance of malicious nodes with a very small amount of labeled data compared to LTrust and STMS. The other two models require more labeled data to learn to achieve the same detection performance as TESS.

Table 3 compares the average accuracy of the three models in detecting malicious nodes in Grey-hole attack, and the amount of labeled data is $[10, 100]$.

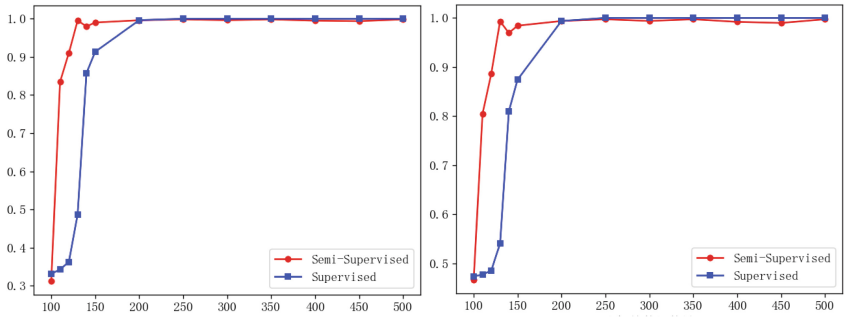
5.1 Performance of TESS

Performance Comparison with Supervised Classifiers. In order to more intuitively verify the detection performance of TESS against malicious nodes in a small amount of labeled data, the performance of TESS was compared with that of supervised classifiers with the same discriminator network structure under Bad-mouthing attack and Good-mouthing attack. The experiment is set to 100 nodes, the number of labeled data in the training set is $[100, 500]$, and the proportion of malicious nodes in the test set is 30%. The comparison results are shown in Fig. 4 and Fig. 5.



(a) Accuracy & Number of labeled data. (b) F1-score & Number of labeled data

Fig. 4. Comparison of performance under Bad-mouthing attack between Semi-Supervised learning and Supervised learning.



(a) Accuracy & Number of labeled data. (b) F1-score & Number of labeled data.

Fig. 5. Comparison of performance under Good-mouthing attack between Semi-Supervised learning and Supervised learning.

Figure 4 and Fig. 5 show the comparison of accuracy and F1-score under Bad-mouthing attack and Good-mouthing attack. As can be seen from the two figures, when the amount of labeled data is 100, the accuracy of semi-supervised classifier and supervised classifier is not ideal. However, when the number of labeled data

5 Simulation Results and Analysis

In this section, we evaluate the performance of the TESS by comparing it to other models. TESS is compared with STMS [16] and LTrust [17]. The classification and evaluation of trust evidence is based on Semi-Supervised learning, and performance of TESS is tested on 3 attack types. Table 2 lists the default simulation parameters. For fair comparison, all three trust models are evaluated based on the same simulated UWSN.

Table 2. Simulation parameters setting

Parameters	Default value
Node	100
Area	$500 \times 500 \times 500\,\mathrm{m}^3$
Communication range	200 m
Initial energy	100 J
The ratio of malicious node	0.3

Our proposed TESS model is implemented in Python 3.6. The environment of TensorFlow 1.15.0 is configured to implement an evaluation method based on SGAN. The sensor layout is 100 nodes randomly deployed in a $500 \times 500 \times 500\,\mathrm{m}^3$ area.

The STMS model and the LTrust model are used as baseline for comparison. Both two trust models use multi-dimensional trust evidence and a supervised learning algorithm to evaluate trust.

Evaluation metrics: In order to evaluate the effectiveness of the TESS model, we use the following metrics to evaluate the performance of TESS.

(1) Accuracy: accuracy refers to the proportion of correctly classified samples in the total number of samples. The higher the accuracy, the higher the overall prediction accuracy and the better the performance of the model.
(2) Precision: precision indicates the accuracy of the prediction in the positive sample result.
(3) F1-score: F1-score is the harmonic average of precision and recall. F1-score is more suitable for unbalanced data sets with a large difference in the proportion of positive and negative samples.
(4) AUC (Area Under Curve): AUC indicates the area under the ROC curve. AUC ranges from 0.5 to 1. The larger the AUC, the better the classification effect.

Data set: This paper uses trust data set generated by simulating underwater real environment, which includes the training set and the testing set. Among them, both the training set and the test set contain node data of 60 time periods in the simulation environment, the proportion of malicious nodes is $[0.05, 0.5]$.

a neuron to generate the probability that the input for D is a fake sample, i.e. $p_n = (y = N+1|x)$. The use of unlabeled data is achieved by maximizing $\log p_n = (y \in \{1, \ldots, N\}|x)$ provide that the input class N has real data. Assuming that the data set contains half the real data and half the fake data, the loss of training classifier can be defined as:

$$\mathcal{F}_{x,y\sim p_d(x,y)} [\log p_n(y|x)] - \mathcal{F}_{x\sim G} [\log p_n(y = N + 1|x)] = \mathcal{L}_s + \mathcal{L}_{us}, \quad (8)$$

$$\mathcal{L}_s = -\mathcal{F}_{x,y\sim p_d(x,y)} \log p_n(y|x, y < N + 1), \quad (9)$$

$$\mathcal{L}_{us} = -\{\mathcal{F}_{x\sim p_d(x)} \log [1 - p_n(y = N + 1|x)] + \mathcal{F}_{x\sim G} \log [p_n(y = N + 1|x)]\}, \quad (10)$$

where, $\mathcal{F}_{x,y\sim p_d(x,y)}$ is the objective function when training with real labeled data, $\mathcal{F}_{x\sim G}$ is the objective function when training with data generated by the generator. The total cross entropy loss is the supervised loss function L_s and the unsupervised loss $\mathcal{L}_{us}$. p_n is the probability of discriminative fake sample. Unsupervised classification only needs to output true and false, so let $D(x) = 1 - p_n(y = N + 1|x)$ and substitute $\mathcal{L}_{us}$ to get:

$$\mathcal{L}_{us} = -\{\mathcal{F}_{x\sim p_d(x)} \log D(x) + \mathcal{F}_{z\sim noise} \log(1 - D(G(z)))\}. \quad (11)$$

The specific steps of SGAN algorithm are as follows: first, the generator G is fixed, and the discriminator D is trained by supervised and unsupervised methods. Then, D is fixed and the G is updated using fake samples generated by random noise. This process is repeated until the model converges.

4.2 Trust Evaluation

Softmax classifier is used to classify sample data in SGAN. The input parameters $T_i = \{T_i^{com}, T_i^{data}, T_i^{energy}\}$ received by the softmax classifier represent the communication trust, data trust, and energy trust of node i, respectively. The parameters have two categories. "0" indicates normal node and "1" indicates malicious node, so $l_i \in \{0, 1\}$. The probability that softmax regression will assign input data T_i to class c is:

$$p(l_i = c|T_i; \theta) = \frac{e^{\theta_c^T T_i}}{\sum_{j=1}^{2} e\theta_j^T T_i}, \quad (12)$$

where, $\theta_0, \theta_1 \in \theta$ are the parameters of the model. Multiplying by $\sum_{j=1}^{2} e\theta_j^T T_i$ is to make the probability in $[0, 1]$. The sum of probability is 1. Let the category of $max(p)$ be l_i, then the final trust evaluation result can be expressed as:

$$Node_{class} = l_i. \quad (13)$$

When $l_i = 0$, the node is identified as a normal node. when $l_i = 1$, the node is identified as a malicious node.

where x is the real data and p_d is the distribution. z is the noise data and p_z is the distribution. $\mathcal{F}_{x\sim p_d(x)}$ represents the objective function when training with real data, and $\mathcal{F}_{z\sim p_z}$ represents the objective function when training with noisy data. $D(x)$ represents the probability of D judging whether real data is true, and $D(G(z))$ is the probability that D determines whether the noisy data generated by the generator is real.

When $p_d = p_G$ is global optimal, p_G generates the distribution of data for G, and the discriminator will achieve the optimal effect by fixing the generator.

SGAN [6] improves the structure of the basic GAN, as shown in Fig. 3. The generator G receives a tensor with input $128 \times 1 \times 3$ through a Dense layer and performs batch normalization after converting the tensor from $128 \times 1 \times 3$ to $1 \times 3 \times 128$ through a transposed convolution layer. After activation with the Leaky ReLU function, the convolution layer is again transposed to $1 \times 3 \times 64$. After the same batch normalization and activation steps, the final transposed convolution changes the tensor to $1 \times 3 \times 1$. The output layer uses hyperbolic tangent (tanh) activation function.

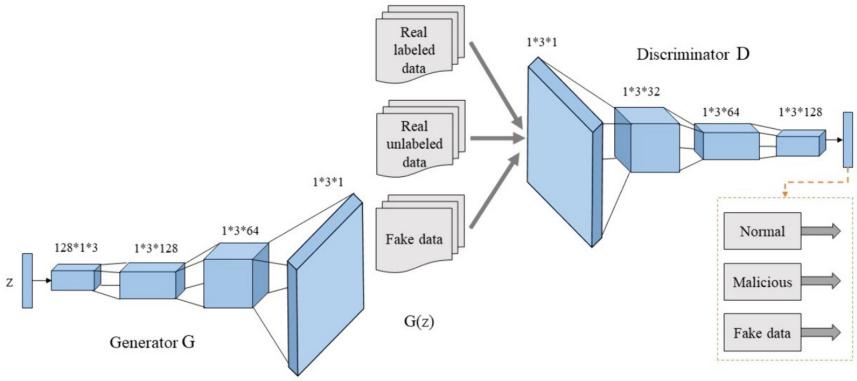


Fig. 3. The structure of SGAN.

Discriminator D use three convolutional layers to change the tensor to $1\times3\times32$, $1\times3\times64$ and $1\times3\times128$, respectively. Batch normalization and Leaky ReLU activation are performed in the middle of every two convolution layers. Finally, regularization is performed using Dropout to randomly drop neurons to prevent overfitting, which is set to 0.5. The output layer of the D is replaced with the softmax classifier, using multiple cross-entropy losses for the supervised task of assigning labels to real samples. The logical function (sigmoid) and binary cross entropy are used for the unsupervised task of classifying real and fake samples.

Training Based on SGAN Algorithm. Assuming that the training data has N classes, the output of the discriminator is $N + 1$, where "1" is the class that discriminates that the input is a fake sample. The softmax classifier adds

Energy Trust. When all the nodes in UWSN are active normally, the energy consumption rate is a stable value. However, when malicious nodes launch attacks, the energy consumption will be different from that of normal nodes. For example, when a node initiates a Grey-hole attack, malicious nodes discard some data packets, resulting in a lower energy consumption rate than normal nodes. Therefore, we use the residual energy rate of the node to measure the reliability of sensor node in terms of energy. The energy trust of a node is defined as:

$$Trust_{energy} = \frac{E_{res}}{E_0}, \tag{5}$$

where E_res is the residual energy and E_0 is the initial energy.

4.1 Semi-Supervised Learning Method Based on GAN

UWSN is deployed in marine environments with poor communication conditions, so there is limited effective data collected underwater. Most of the existing trust evaluation methods adopt machine learning and rely on a large number of labeled data to train models. In order to deal with the problem of scarce underwater data, we use the Semi-Supervised learning method based on GAN to evaluate the trust of nodes.

Data Preprocessing. The trust dataset simulation is based on the design in Sect. 3.1. The format of the data set is an $n \times 3$ matrix, namely $Trust_i = \{Trust_i^{com}, Trust_i^{data}, Trust_i^{energy}\}$. Before trust modeling, the data set is preprocessed. Firstly, both the training set (x_{train}) and the test set (x_{test}) are reconstructed into a three-dimensional matrix of $n \times 1 \times 3$ and extended to the input size of $(1 \times 3 \times 1)$. Then, by pre-defining the number of labeled data, a random batch of labeled data and its labels are obtained from the reconstructed training set. The remaining data is regarded as unlabeled data.

Data Processing and Algorithm Structure of SGAN. In this paper, Semi-Supervised learning based on GAN is used to identify malicious nodes. GAN can continuously learn and simulate the distribution of data and generate samples that are similar to real data. Based on the principle that GAN generator G and discriminator D minimize the loss of D through game play, we use a small amount of labeled and a large amount of unlabeled underwater data to optimize the D. In this method, the classification performance of D will be improved. Therefor, we can achieve accurate identification of malicious nodes.

The game between GAN generator G and discriminator D can be modeled as a two-party min-maximum game problem:

$$\min_G \max_D V(G, D), \tag{6}$$

$$V(G, D) = \mathcal{F}_{x \sim p_d(x)} \log D(x) + \mathcal{F}_{z \sim p_z} \left[\log(1 - D(G(z)))\right], \tag{7}$$

The premise of trust evaluation is the collection of trust parameters, which mainly focuses on the communication behavior, data transmission and energy consumption of nodes. Based on the analysis of the corresponding performance when nodes launch internal attacks, there are three main results: communication failure, packet error and abnormal energy consumption. The consequences of attacks are shown in Table 1.

Table 1. Attacks and consequences

Attack	Consequences
Grey-hole	Lose some packets and the energy consumption is small
Bad-mouthing	The trust value of a normal node is deliberately lowered
Good-mouthing	Deliberately raise the trust value of the malicious node

Communication Trust. The communication behavior of nodes can be divided into normal or abnormal, and the binomial distribution can be used to simulate the communication behavior of nodes. For *Bayesian* analysis, the *Beta* function is a conjugate prior of the binomial likelihood distribution [14]. Therefore, it can be used to simulate the distribution of communication trust.

Assuming that there are $(u+v)$ interactions between nodes. u and v represent the number of cooperative and non-cooperative times, when communicating, respectively. Node i collects the communication behavior of node j. node j's behavior is denoted as δ, which follows a uniform distribution, namely $P(\delta) = uni(0,1) = Beta(1,1)$. The probability of occurrence of this behavior can be obtained by using the *Beta* distribution:

$$P(\delta) = \frac{Bin(u+v,u) \times Beta(1,1)}{u+v+1} = Beta(u+1, v+1). \tag{1}$$

According to the *Beta* distribution, the reputation of node j at node i is:

$$R_{ij} = Beta(u+1, v+1), \tag{2}$$

therefore, the communication trust of node i computes node j is represented by:

$$Trust_{com} = E(R_{ij}) = E(Beta(u+1, v=1)) = Beta(u+1, v+1). \tag{3}$$

Data Trust. Packets sent by adjacent nodes are spatially and temporally related. The values of packets are normally distributed as $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ [15]. However, packets attacked by malicious nodes are different from normal packets. Therefore, we determine whether the node is under malicious attack according to the data packet value. The data trust is defined as follows:

$$Trust_{data} = 2\left(0.5 - \int_{\mu}^{x_d} f(x)dx\right) = 2\int_{x_d}^{\infty} f(x)dx, \tag{4}$$

where x_d is the data value of target node, μ is the average value of this set of data. The larger the difference between x_d and μ, the smaller the data trust value.

3.3 Problem Formulation

The dynamic change of underwater environment will reduce the quality of communication links between sensor nodes. Malicious nodes will exhibit some behaviors different from normal nodes due to launching attacks. For example, they may show abnormal energy consumption, poor packet transmission quality low transmission rates. Due to the difficulty and limited energy of underwater sensor nodes in collecting data, it is unable to provide sufficient data for model training and node identification. To solve these problems, we propose a trust evaluation method based on semi-supervised learning. When there is less labeled data, the unlabeled data can still be used to improve the performance of classifier and identify malicious nodes more accurately.

Definition:

$P(\cdot)$ stands for discrete or continuous probability distribution. $uni(\cdot)$ represents uniform distribution. $Bin(\cdot)$ indicates a binomial distribution.

4 Design of TESS

In this section, we firstly describe the overview of trust evaluation method based on Semi-Supervised Learning. Then we give the detailed methods of TESS, which are described as follows:

The overall flow of the TESS method is shown in Fig. 2.

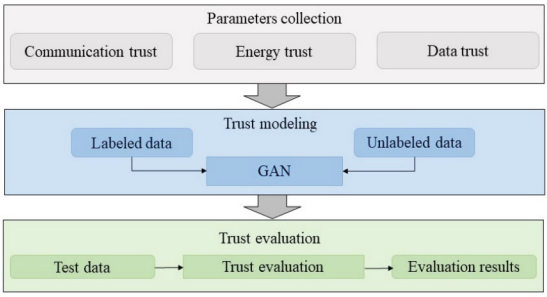


Fig. 2. The process of TESS.

The TESS method consists of three parts: collection of trust parameters, semi-supervised learning process and trust evaluation. In the trust parameter acquisition stage, the underwater sensor node collects three types of trust parameters, including communication, energy and data. They constitute the trust matrix as the input data in the trust modeling stage. In the process of trust modeling, we adopt a Semi-Supervised learning method based on Generative Adversarial Network (GAN). It makes full use of a small amount of labeled data and a large amount of unlabeled data under water. The trained model is used as the detection model in the trust evaluation stage to evaluate the reliability of the detected nodes.

3 Research Scenario and Problem Formulation

This section briefly introduces the system model, communication model, computation model and problem formulation.

3.1 System Model

The structure of UWSN is shown in Fig. 1. The underwater sensor nodes using to collect data are randomly distributed in a certain range of monitoring areas and fixed on the seabed through a tether wire. Sensor nodes can communicate with other nodes within their own communication range through acoustic waves to exchange the collected information. It is assumed that each sensor node has the same initial energy, computing power and storage resources. The sensor node is powered by a battery. The sensor node can communicate with the surface device through the sink node. The surface can communicate with other surface equipment and satellite.

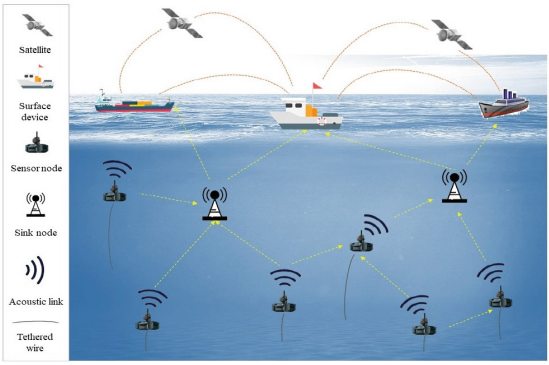


Fig. 1. The structure of UWSN.

3.2 Attack Model

The communication channel in UWSN is unstable and highly susceptible to internal attacks. The UWSN trust management scheme studies a variety of attacks, such as Grey-hole attack, On-off attack, collusion attack and so on. We discuss the following three malicious attack modes [13]:

(1) Grey-hole attack: some malicious nodes intentionally discard some received packets.
(2) Bad-mouthing attack: malicious nodes intentionally offer dishonest recommendations for normal nodes.
(3) Good-mouthing attack: malicious nodes intentionally provide good recommendations to malicious nodes.

data and real labeled data. The goal of the SGAN discriminator is to classify the sample correctly and exclude the fake samples if the input sample is true. In addition, the training goal of the discriminator is to make the discriminator network a Semi-Supervised classifier using only a small part of the label data, and the accuracy is as close as possible to the supervised classifier. The goal of the generator is to help the discriminator learn the knowledge in the sample by providing fake samples that are close to the real sample, thus improving the accuracy of the discriminator's classification.

The overall process of SGAN is as follows: Firstly, the generator is fixed, and the discriminator is trained by supervised and unsupervised methods. Then the discriminator is then fixed and the generator is updated using pseudo-samples generated by random noise. The process is repeated until the model converges.

2.2 Intrusion Detection Method Based on Semi-Supervised Learning Method

With the development of the Industrial Internet of Things (IoT), IoT devices are often under attack. Intrusion detection is an effective method to improve the security of IoT. Semi-Supervised learning has had some applications in the field of intrusion detection. Existing methods based on deep learning often require a large amount of labeled data to support. However, blockchain methods have high requirements for device computing power [7], which poses a challenge to the computing power and resources of IoT devices [8].

Gao *et al.* proposed an integrated Semi-Supervised learning method based on fuzzy [9]. This model used classification regression tree (CART) [10] as the basic learner to mine useful information of unlabeled data based on fuzzy method, and integrated the supervised part and the unsupervised part. The classifier was constructed using unlabeled data to improve the detection effect. In order to solve the problem that IoT devices cannot perform complex calculations, Zhao *et al.* proposed a lightweight intrusion detection method based on consistent regularization Semi-Supervised learning (LSSL) [11]. LSSL improved detection performance by consistently training unlabeled traffic data and used separable convolution for efficient feature extraction. In addition, principal component analysis (PCA) algorithm was used in the pre-processing stage to reduce the complexity and improve the detection performance in complex networks. Aiming at the problem that intrusion detection methods based on deep learning rely on extracting data distribution from a large number of normal data, Cai *et al.* proposed a meta-gradient intrusion detection method (FSMG) based on feedback deep Semi-Supervised learning [12]. FSMG used a lightweight evaluation network for data enhancement and non-processing of inputs and predicted different distributions of training data from small amounts of labeled data. It transformed abnormal data into data for which information can be extracted and dynamically updated models to reduces labeling errors. The double-layer nested optimization gradient update method of the model can ensure that the model converges within $O(C/\sqrt{T})$.

diverse connections of wireless communication [3], the application of machine learning algorithms in practice has certain limitations.

The marine environment of underwater wireless sensor networks (UWSN) is complex and highly dynamic. Due to the instability of acoustic communication, it is difficult for sensor nodes to obtain a lot of effective data. However, the performance of trust models based on traditional machine learning techniques mainly depends on large amounts of training data, but obtaining sufficient data underwater is costly and time-consuming. Edge computing has improved for such problems [4,5], but it is not yet widely used in UWSN. In addition, due to the limited computing power and energy of sensor nodes, it is impossible to train highly complex machine learning models [3]. Therefore, the using of trust models based on machine learning algorithms in UWSN faces the challenge of insufficient data volume.

Based on the above challenges, we propose a trust evaluation method based on Semi-Supervised learning (TESS). The contributions of this paper are as follows:

(1) We propose a trust evaluation method based on Semi-Supervised learning, which uses a small amount of labeled data and a large amount of unlabeled data to build a classifier. This method solves the problem of scarce underwater labeled data and saves the cost of a large amount of data labeling.
(2) We build a Semi-Supervised classifier based on the Generative Adversarial Network, which improve the classification performance of the discriminator through the game training of the generator and discriminator. At the same time, the accuracy of malicious node identification is improved effectively.
(3) Compared with existing trust models based on machine learning, the accuracy of TESS under Bad-mouthing attacks is respectively improved by 26.45% and 26.78%.

The rest of the paper is organized as follows. The Sect. 2 introduces related work. The Sect. 3 introduces the research scenario and problem definition. The Sect. 4 describes the TESS in detail. In Sect. 5, simulation experiments are carried out and the results are analyzed. Section 6 offers conclusions and also discusses limitations, and opportunities for future research.

2 Related Work

2.1 Semi-Supervised Generative Adversarial Network

In this paper, Semi-Supervised learning based on GAN is used to identify malicious nodes, it is an effective application of GAN in practical fields. Semi-Supervised Generative Adversarial Networks (SGAN) are a variant of GAN [6], which extends GAN to the Semi-Supervised domain by using discriminator networks to output class labels.

SGAN includes generator and discriminator, the game between generator and discriminator can be modeled as a two-party min-max game problem. The discriminator receives three inputs: fake samples from the generator, real unlabeled

Semi-Supervised Learning Based Trust Evaluation for Underwater Wireless Sensor Networks

Weicheng Meng[1], Zhenquan Qin[1], Yuxin Cui[1], Hao Lu[2,3], Bingxian Lu[1], and Jianbo Zheng[3,4(✉)]

[1] Dalian University of Technology, Dalian 116024, Liaoning, China
[2] City University of Macau, Macau 999078, China
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
jianbo.zheng@smbu.edu.cn
[4] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

Abstract. In recent years, trust mechanism has gradually become an effective scheme to deal with the internal attacks of underwater wireless sensor networks (UWSN). However, most of the existing trust models are based on traditional machine learning algorithms, which require a large amount of data training to improve the accuracy of the model. Therefore, these models still face the challenge of insufficient data in UWSN. In this paper, we propose a trust evaluation method based on Semi-Supervised learning (TESS). We consider the difficulty of underwater data collection and the lack of valid data. TESS uses a Semi-Supervised classification method based on Generative Adversarial Networks (GAN) to classify the collected trust parameters. This method can train high-precision detection models using a small amount of labeled data and a large amount of unlabeled data. Simulation results show that compared with LTrust and STMS, the accuracy of TESS under Bad-mouthing attacks is respectively improved by 26.45% and 26.78%.

Keywords: Semi-Supervised learning · Trust evaluation · UWSN

1 Introduction

In recent years, the agility and effectiveness of machine learning has been able to cope with large-scale problems [1]. Machine learning promotes the application of the Internet of Things in industrial, medical, transportation and other fields [2]. Therefore, machine learning was also widely used in the trust model of wireless sensor networks. However, due to the high mobility, dynamic environment and

This work was supported in part by the Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 411–425, 2024.
https://doi.org/10.1007/978-3-031-65126-7_35

19. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
20. Vinh, N., Epps, J., Bailey, J.: Information theoretic measures for clustering comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**(10), 2837–2854 (2010)

References

1. Yong, L., Zhou, Z., Xiong, L.: Research on electromagnetic spectrum management and control system of launch site based on big data. In: 21st International Symposium on Communications and Information Technologies 2022, Xi'an, China, pp. 27–30 (2022)
2. Hu, S., Pei, Y., Liang, Y.: Sensing-mining-access tradeoff in blockchain-enabled dynamic spectrum access. IEEE Wirel. Commun. Lett. **10**(4), 820–824 (2021)
3. Wang, C.: A spectrum feature-based security situation awareness algorithm for network coordinate system. In: 11th International Conference on Communications, Circuits and System 2022, pp.13–15 (2022)
4. Mario, B.: A deep learning-based signal classification approach for spectrum sensing using long short-term memory (LSTM) networks. In: 6th International Conference on Information Technology, Information Systems and Electrical Engineering 2022, Yogyakarta, Indonesia, pp. 13–16 (2022)
5. Liu, C., Wu, X., Zhu, L.: The communication relationship discovery based on the spectrum monitoring data by improved DBSCAN. IEEE Access **121**(10), 793–804 (2019)
6. Zeng, Y., Zhang, R.: Wireless information surveillance via proactive eavesdropping with spoofing relay. IEEE J. Sel. Topics Signal Process **10**(8), 1449–1461 (2016)
7. Han, Y., Duan, L., Zhang, R.: Jamming-assisted eavesdropping over parallel fading channels. IEEE Trans. Inf. Forensics Security **14**(9), 2486–2499 (2019)
8. Xu, J., Duan, L., Zhang, R.: Proactive eavesdropping via cognitive jamming in fading channels. IEEE Trans. Wireless Communication. **16**(5), 2790–2806 (2017)
9. Pan, T., Wu, X., Yao, C.: Communication behavior structure mining based on electromagnetic spectrum analysis. In: Proceedings of IEEE TAIC 2019, Chongqing, China, pp. 1611–1616 (2019)
10. Zhang, H., Yao, Y., Lei, L.: Study of end-to-end radio pass-through relationship based on spectrum data. Commun. Technol. **53**(11), 2745–2748 (2020)
11. Liu, C., Wu, X., Zhu, L.: Discover and research of communication relation based on communication rules of ultrashort wave radio station. In: Proceedings of IEEE ICBDA 2019, Suzhou, China, pp. 112–117 (2019)
12. Liu, C., Wu, L., Zhu, L.: Research on communication network structure mining based on spectrum monitoring data. IEEE Access **10**(1109), 3945–3959 (2019)
13. Franceschet, M.: PageRank: standing on the shoulders of giants. Commun. ACM **54**(6), 1–9 (2010)
14. Long, Z., Hong, C., Gke, L.: Recognizing automatic link establishment behaviors of a shortwave radio station by an improved unidimensional DenseNet. IEEE Access **10**(1109), 96055–96064 (2020)
15. Cheng, K., Zhu, L., Yao, C.: DCGAN based spectrum sensing data enhancement for behavior recognition in self-organized communication network. In: China Communications 2021, China, pp. 182–196 (2021)
16. Cheng, K., Zhu, L., Yao, C.: Squeeze-excitation based communication behavior recognition approach for spectrum sensing data. In: International Conference on Computer and Communications. 2021, Chengdu, China, pp. 682–686 (2021)
17. Zhang, H., Zhu, L., Yao, C.: Recognition of communication relationship based on the spectrum monitoring data by improved VGGNET. In: International Conference on Computer and Communications. 2020, Kunming, China, pp. 329–337 (2020)
18. Zhang, H., Zhu, L., Yao, C.: Recognition of connection relationship based on the spectrum monitoring data by deep convolutional neural networks. In: International Conference on Communication Technology, China, pp. 1285–1290 (2020)

network structure can be formed. This approach strikes a balance between performance and resource efficiency.

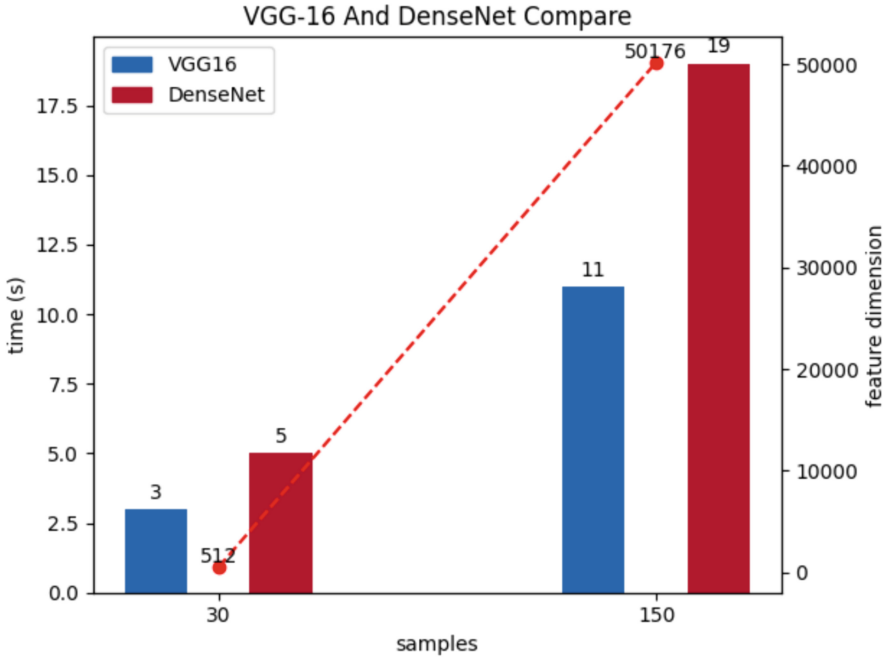


Fig. 7. VGG16 and DenseNet121 comparison diagram.

6 Conclusion

This paper aims to reduce the time and cost of deciphering communication contents by analyzing the characteristics of spectrum data to mine the communication behavior and network topology of nodes. Firstly, the features of spectrum data are selected, and preprocessing techniques are applied to calculate and uniquely determine the network nodes. Then, a CNN model VGG16 is utilized to extract the features of the spectrum data. With the maximum similarity method in hierarchical clustering, the samples with higher similarity are grouped into one class. This process helps to obtain the communication behavior of each node according to the clustering results. A network topology discovery algorithm is designed to mine of communication behavior and obtain wireless network topology from spectrum data. Experimental results demonstrate that the method is able to discover hidden communication behaviors that contribute to reveal the existing wireless network topology.

Acknowledgement. This research was supported by the National Natural Science Foundation of China (No. 62171449).

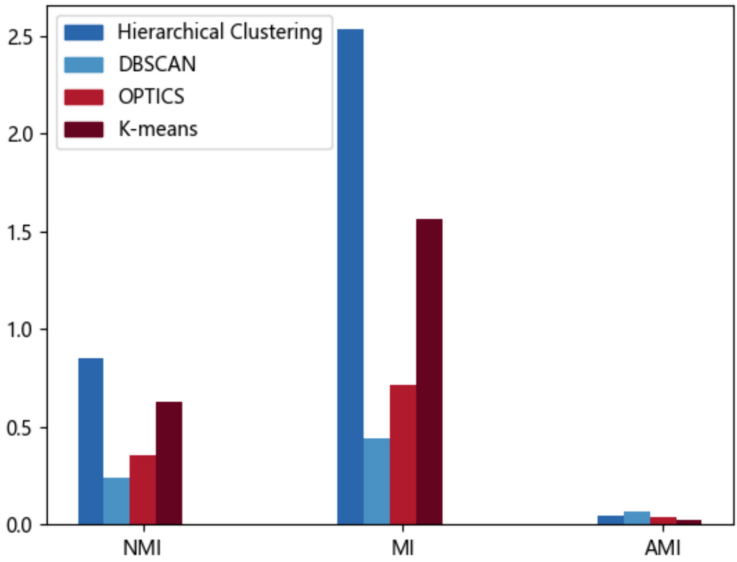


Fig. 5. The VGG16 method comparison diagram.

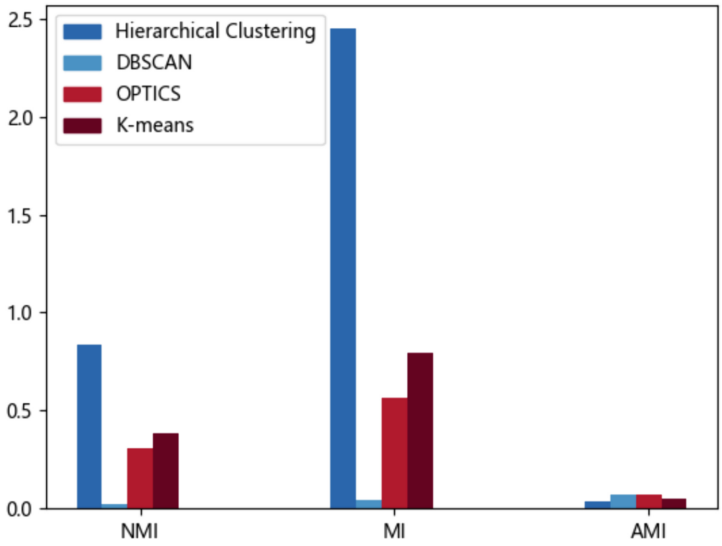


Fig. 6. The DenseNet121 method comparison diagram.

network surpasses that of the DenseNet121 network. Hence, it is preferable to extract the image feature vectors using the VGG16 network and then determine the number of clusters based on the clustering tree. After clustering, the categories can be filtered to identify the through-connections. By connecting the network nodes, a communication

Table 1. Experiment parameters.

Parameter	Value
Transmitter Antenna Wavelength Spacing	5 m
Receiver Antenna Wavelength Spacing	5 m
Bandwidth	46.08 kHz
Number of Subcarriers	384
Number of Picked Carriers	5
Number of Nodes	30

the distribution of node locations obtained from signal power and orientation, representing the nodes comprising this communication network. The procedure involves acquiring the features of spectrum data using the VGG16 model from the dataset. Subsequently, we apply hierarchical clustering to different nodes based on the principle of maximum similarity. Nodes exhibiting high similarity indicate sustained communication behaviors and a generic relationship between them. All nodes sharing a generic relationship are mined to construct the wireless network topology. By analyzing the clustering tree, we identify edges and network nodes, setting the number of categories for clustering and obtaining the clustering effect graph. In the same category, a through relationship between two nodes is established, and based on this relationship, the corresponding nodes are connected to form the network topology.

In this study, feature vector extraction is performed using the VGG16 network and DenseNet121 network [14]. Hierarchical clustering and other clustering methods (DBSCAN, OPTICS, K-means) are compared using evaluation metrics such as Mutual Information (MI), Normalized Mutual Information (NMI), and Adjusted Mutual Information (AMI), as shown in Fig. 5 and Fig. 6. MI was first proposed in [19], NMI is a metric that normalizes MI, and AMI is a metric that adjusts mutual information to account for the effects of random clustering results [20]. These metrics are employed to assess the degree of similarity between clustering outcomes and ground truth labels, where elevated values denote superior performance. The results indicate that hierarchical clustering outperforms other methods under the NMI and MI indexes.

Specifically, under the NMI metric, the image feature vector extraction using the VGG16 network with hierarchical clustering yields 0.81, while K-means achieves 0.68. The remaining two methods result in NMI scores of 0.45 and 0.47, respectively. For the DenseNet121 network, the NMI for hierarchical clustering is 0.84, whereas K-means and OPTICS produce NMI scores of 0.37 and 0.31, respectively. Thus, hierarchical clustering demonstrates superior overall performance in this context.

Figures 5 and 6 illustrate that the hierarchical clustering method exhibits better performance. However, the evaluation metrics MI, NMI, and AMI under the two convolutional networks do not vary significantly. The DenseNet121 network requires more time to extract image feature vectors, and the dimension of the extracted feature vectors is much larger compared to the VGG16 network, as depicted in Fig. 7. Taking into account the time cost and memory considerations, the overall performance of the VGG16

Algorithm 1. Network topology discovery

```
Initialize communication scenarios
Acquisition of spectrum data
Determine distance and orientation information
for epochs = 1, M do
    MLP prediction node location
end for
VGG16 extracts spectrum data feature vectors
Hierarchical clustering for classification
Obtain communication relationships
Obtain network topology
end
```

In military scenarios, information is usually transmitted step by step due to a strict hierarchy, i.e., information is transmitted in steps. Although network nodes are interconnected, communication devices across levels are subject to certain restrictions, i.e., each communication device has a different communication range and authority, and the flow of information from lower to higher levels often requires a step-by-step delivery. Therefore, the communication behavior of nodes at different levels differs. When we analyse network topology, the path contains the direction of information transmission, network hierarchy, and communication order. Determining the network key nodes and sub-networks can also be done by analyzing the network topology.

5 Simulation Experiments

5.1 Scene Setting and Data Collection

We adopt the dataset which was built based on the real map scenarios[1]. We consider 30 nodes are randomly distributed in a square area of 171 m × 171 m. The nodes communicate in the frequency range of 885–909 MHz and 930–954 MHz, and the monitoring station has a scan bandwidth of 20 MHz and a scan rate of 80 GHz/s [5].

Based on the nodes and monitoring station setup in Fig. 3, the communications between nodes are simulated, the spectrum data is monitored using the monitoring station, and the spectrum data characteristics are analyzed, from which the communication relationship and network topology are mined. In the simulation experiment, node terminals communicate with each other and generate feedback. The wireless network topology is determined based on the location of nodes and simulated communication.

The experiment parameters are presented in Table 1, which contains bandwidth, number of pickup carriers, and transmitter and receiver antenna wavelength spacing.

5.2 Analysis of Experimental Results

According to the VGG16 network and hierarchical clustering method, features are extracted from the spectrum data to mine the wireless network topology. Figure 3 displays

[1] https://www.mobileai-dataset.com/html/default/zhongwen/shujuji/1592719963402108929.html?index=1.

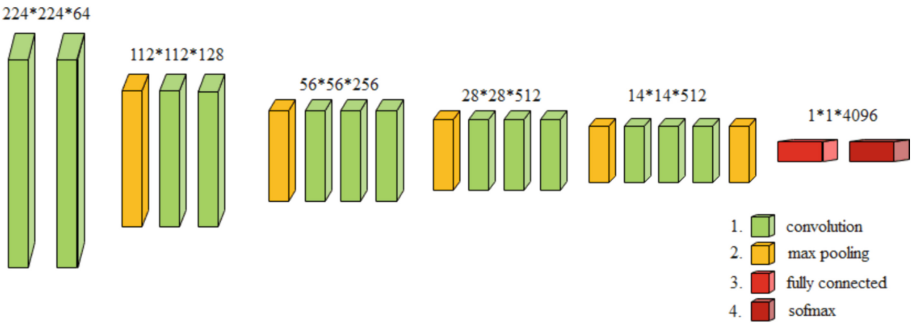


Fig. 4. The structure of VGG16 model.

into different data clusters. On the contrary, cohesive hierarchical clustering treats each individual as a cluster and then keeps merging these clusters until the clusters of different classes are obtained under a particular condition.

We use the feature vectors obtained by VGG16, then clusters the feature vectors based on the maximum similarity. The number of clusters is decided based on the similarity of the clustering tree. As a result, different clusters represent different communication relationships.

4.4 Network Topology Discovery

We obtain network topology according to the PageRank algorithm in the literature [13]. Therefore, if a network node is connected to other nodes, this node behaves as member node in the communication network, i.e., the node has a high PageRank value. Assuming that a node has a high PageRank value, the PageRank value of the nodes connected to it will increase accordingly. Therefore, the laws of edges and nodes should be considered when analyzing the network topology.

Suppose the wireless network topology mined from the spectrum data is $\boldsymbol{C} = \{\boldsymbol{N}, \boldsymbol{R}\}$, where $\boldsymbol{R} = \{(e_1, p_1), (e_2, p_2), ..., (e_i, p_i), ..., (e_M, p_M)\}$ stands for the pass-through relationship between nodes, p_i represents the number of connections, and e_i defines the edges. Besides, we record the number of time slots that the node acts as a sender or receiver and it participates in a communication. This is used to analyze the number of network nodes and wireless network topology. Based on the above description, the procedure of the whole network topology discovery can be presented in Algorithm 1.

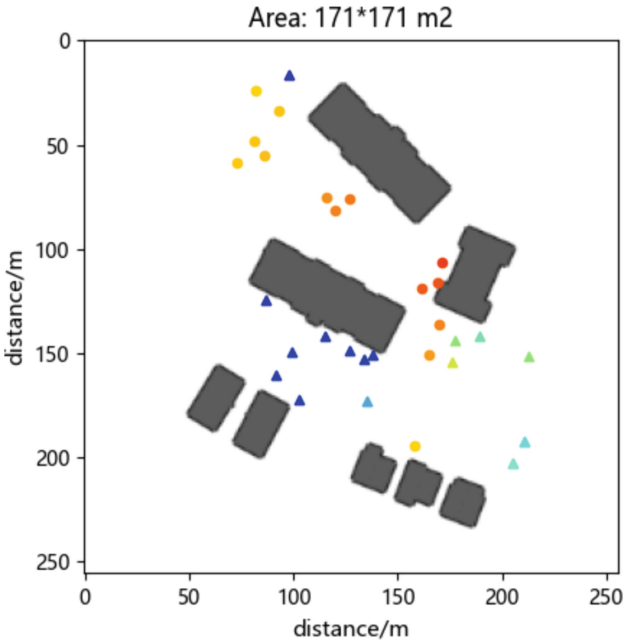


Fig. 3. Node location map.

4.2 VGG16 Model

VGG16 is a well-known CNN model whose name is derived from the initials of the Visual Geometry Group, Oxford University. The model is greatly adapted to the classification task. VGG16 uses small convolutional kernels and small pooling kernels. The model architecture is concise because the convolutional kernels focus on expanding the channel count. The pooling kernels focus on reducing the width and height so that the model is more profound and broader while the computation increases more slowly. At the same time, the fully connected layer is replaced with three convolutional layers in the training phase of the network. The test of the convolutional network has no limitation of full connection, so the input can be an image of any size. The structure of VGG16 model is illustrated in Fig. 4.

The communication relationship can be obtained as there is similarity in spectrum data features. To obtain the communication relationship, we first use CNN to extract the feature vectors of the spectrum data. We do not rely on the physical characteristics of the spectrum data, but the features exhibited by the image of spectrum.

4.3 Hierarchical Clustering

Hierarchical clustering methods put the nearest sample points into one class by calculating the distance between samples and then merges the nearest classes into a larger class. As a branch, split hierarchical clustering method initially put all samples in the same class. By continuously excluding dissimilarities, the samples are finally grouped

3.2 Feature Representation

Assuming that $\boldsymbol{N} = \{1, 2, ..., N\}$ represents the index of network nodes. According to the spectrum data from the monitoring station, the spectrum data at time slot t is $\boldsymbol{U}^t = \left\{\boldsymbol{u}_1^t, \boldsymbol{u}_2^t, ...\boldsymbol{u}_i^t, ..., \boldsymbol{u}_N^t\right\}$, where $\boldsymbol{u}_i^t = \left\{p_i^t, \theta_i^t, t_i^t, f_i^t\right\}$, $i \in \boldsymbol{N}$ represents the set of signal power, signal orientation, monitoring time, and carrier frequency of the node i at time slot t. The prediction of the node's location relative to the monitoring station is based on the spectrum data features. The deviation is found to be small when compared with the actual location. This provides the location of a specific node for network topology mining.

4 Communication Relationship and Network Topology Discovery

4.1 Communication Relationship Mining

The key point of communication relationship discovery is to extract the features of the spectrum data and classify them according to the features. The location of a node is predicted by the power and orientation information as the communication network nodes. After that, according to the communication relationship between the classifications, each network node is connected to form a communication network to complete the mining of the network topology.

To determine the communication relationship, we need to obtain the location of each node first. The data set $\boldsymbol{Z} = \{z_1, z_2, ..., z_i, ..., z_N\}$ represents the orientation and power information in the spectrum data, where $z_i = \{p_i, \theta_i\}$, $\boldsymbol{p}_i = \{p_r, p_t\}$, $i \in \boldsymbol{N}$. We assume that the signal propagates in free space, and the signal power decreases gradually with increasing distance. The node distance from the monitoring station is d, the transmit power is p_r, and the received power is p_t. Therefore, the ratio of the transmit power and received power is defined as

$$\frac{p_t}{p_r} = \left[\frac{\sqrt{G_l \gamma}}{4\pi d}\right]^2, \tag{1}$$

where $\sqrt{G_l}$ represents the product of the transmitting antenna gain and the receiving antenna gain. The location of each node can be uniquely determined by combining its orientation and distance, as shown in Fig. 3.

Communication relationships may vary from time to time. In order to study the evolving communication relationships, we can divide the data set by time slots. In this way, we can further intuitively explore the communication sequences, network connectivity path, and communication directions and lay the foundation for the subsequent construction of the network topology. The predicted location of each node is denoted as (d_i, θ_i).

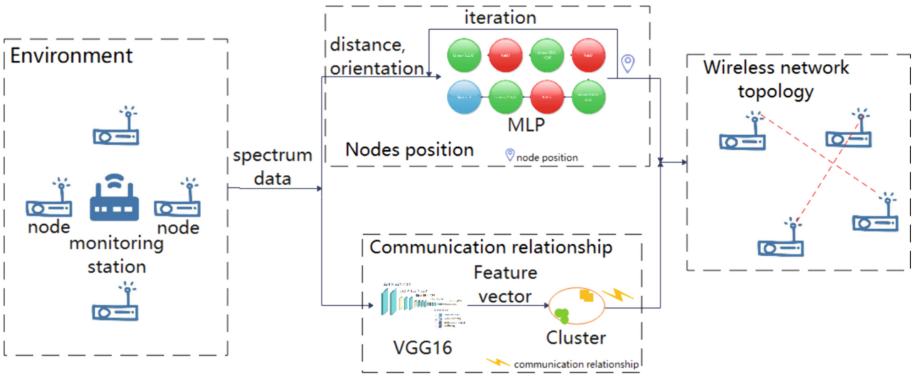


Fig. 2. Overall framework of wireless network topology discovery.

3 Data Pre-processing

3.1 Feature Selection

To mine the communication relationship and topology in wireless nodes from spectrum data, we need to first select the spectrum data features. In fixed-frequency communication, the communication relationship between different nodes can be distinguished by carrier frequency. Then, clustering is then performed based on the obtained spectrum data features and different clustering sets can be obtained. Consequently, each clustering set can represent a communication association and communication relationships are determined according to the carrier frequency characteristics. Here are the features that we select for further feature representation.

- **Signal power:** It is a unit of measurement of signal energy in the communication process. Due to the effect of large-scale signal propagation and fading, smaller power is measured with the increase of distance, and the signal strength is also stable if the node's location keeps unchanged. If a node moves within a short time, we assume that the received signal power remains constant.
- **Carrier frequency:** In long-distance transmission, the signal is not transmitted directly but moved to a fixed high frequency for transmission to improve the propagation distance. This high-frequency signal is called the carrier, and the frequency is called the carrier frequency, also known as the fundamental frequency.
- **Signal orientation:** The node's orientation information is related to the spectrum monitoring station. When the node's location is fixed, the orientation of the received signal is also stable. If the node moves in a relative short time, we consider that the orientation data remains the same.

The features are extracted from the spectrum data by CNN to predict the relative location of the nodes (relative to the spectrum monitoring station), and by using the carrier frequency, the communication association relationships can be uniquely determined.

introduces the simulation platform and discuss the experiment results. Finally, Section 6 summarizes the work of the full paper.

2 System Model

Wireless communication always happens in more than one node, and communication relationship can reflect the wireless network topology. As shown in Fig. 1, it is a scenario of wireless network topology discovery based on spectrum data. The work we do is to obtain the communication relationship between communication nodes and further mine the network topology.

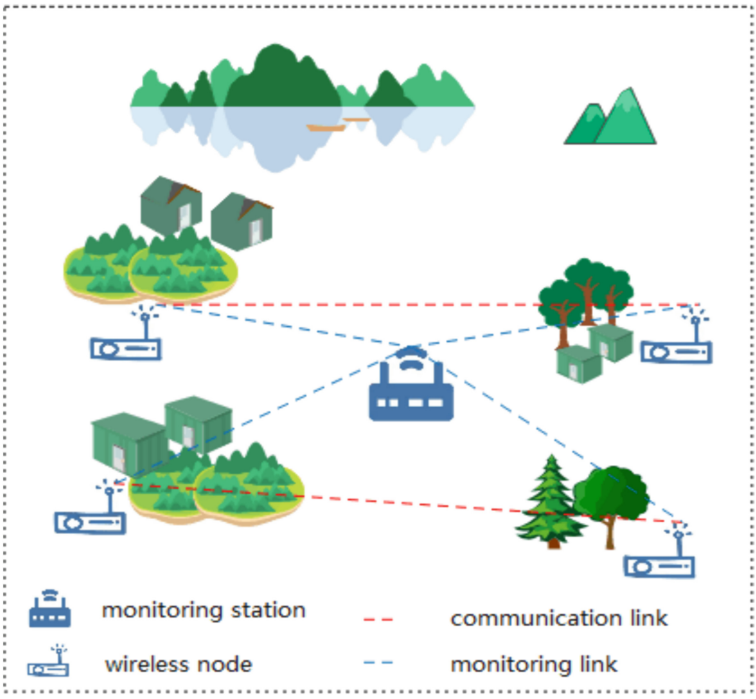


Fig. 1. Scenario of wireless network topology discovery based on spectrum data.

To achieve this goal, we first preprocess the spectrum data and use two variables, i.e., distance and orientation, to train the prediction of node locations. Then, CNN is used to extract image feature vectors, and hierarchical clustering is adopted for classification based on similarity to obtain communication relationships. Finally, the wireless network topology is mined by combining node locations and communication relationships. The overall framework is shown in Fig. 2.

there are numerous works that investigate network topology by mining the statistical laws of spectrum data. The research based on physical characteristics of spectrum data [9–12] needs to obtain a large amount of physical information data before clustering, such as power and frequency. In recent years, machine learning is increasingly being applied to analyze communication relationships with spectrum data. Wu et al. [14] proposed a method to identify different automatic link establishment (ALE) behaviors with an improved DenseNet. The method requires a highly labelled sample size, and the ALE signal strength significantly impacts network performance. Cheng et al. [15] presented a data-enhanced communication behavior recognition scheme to cope with insufficient spectrum data samples. However, the communication scenarios constructed by this method have strict hierarchical relationships and weak generalization ability. Cheng et al. [16] studied squeeze-excitation based communication relationship, but there is an apparent hierarchical relationship between the two communication nodes. Zhang et al. [17, 18] investigated the identification of communication relationship based on convolutional neural networks, but did not perform the network topology. Instead of relying on the communication contents and many physical characteristics, this paper extracts the spectrum data feature vectors combined with hierarchical clustering to mine communication relationships and network topology.

In this paper, by pre-processing the spectrum data and extracting the spectrum data features, we first determine the location of nodes relative to the monitoring station based on the power and orientation information of the spectrum data and then use the image feature vector extracted by convolutional neural network (CNN) to group the ones with high similarity into a category based on the similarity of the features using hierarchical clustering method. Since the spectrum data characteristics of the two nodes are very similar in fixed-frequency communication, it is possible to determine the through-connection relationship between nodes based on spectrum characteristics. Finally, the entire wireless network topology is formed according to the through-linkage association of all nodes. The experimental results prove that the method is simple to implement, the code is concise and has good clustering and adaptability to the spectrum data. The contributions of this paper are as follows:

- We transform the problem of communication relationship into spectrum data image classification problem to recognize communication behavior.
- We study the wireless network topology without relying on the communication content and a large number of physical characteristics, which are obtained by mining the spectrum data features and using the spectrum features similarity to get the association relationships.
- We use a CNN model VGG16 to get feature vectors and then use an unsupervised classification method (clustering) for classification. Our clustering method can achieve better performance than other clustering methods in the Mutual Information (MI) and Normalized Mutual Information (NMI) evaluation index.

The rest of this paper is organized as follows. Section 2 describes the wireless communication network scenario, and also introduces the overall framework for wireless network topology discovery. Section 3 introduces data pre-processing including feature selection and feature representation. Section 4 uses VGG16 model and the hierarchical clustering to mine the communication relationship and network topology. Section 5



Wireless Network Topology Discovery Based on Spectrum Data by Convolutional Neural Network

Xinfeng Deng[1], Zhihui Xie[2], and Li Zhou[1(✉)]

[1] College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China
zhouli2035@nudt.edu.cn

[2] Songjiang Power Supply Company of State Grid Shanghai Electric Power Company, Shanghai, China

Abstract. Wireless network topology can reflect the communication relationships among network node. Since there are significant challenges and difficulties in deciphering the communication contents, spectrum data is adopted to discover communication relationships and network topology of a wireless network. In this paper, we propose a wireless network topology discovery method based on spectrum data to determine the communication relationships of nodes. Since the spectrum data features of nodes are correlated during the communication process, we construct the wireless network topology by mining the communication behaviors of nodes from the spectrum data features based on maximum similarity and hierarchical clustering. Simulation results demonstrate that the proposed method can achieve a better performance of hierarchical clustering than the existing methods.

Keywords: Spectrum Data · Communication Relationship · Network Topology

1 Introduction

With the development of spectrum monitoring and cognitive radio networks, the analysis and mining of spectrum data are receiving more and more attentions [1, 2]. As a reflection of user activities, the study of spectrum data can provide many essential pieces of information and intelligence, especially in anti-terrorism, military communication, and network security fields. Currently, the applications of spectrum data are mainly focused on spectrum situation awareness [3], signal classification [4], and signal feature extraction [5]. The physical characteristics of spectrum data and the statistical laws presented by these features also reflect the through-connection relationships and relevant information. However, mining the connections between the massive amount of spectrum data and the communication relationships among network nodes have yet to be further investigated.

There is a large amount of literature on communication relationship discovery or topology discovery. On one hand, the research of communication relationship discovery can be deciphered with the contents of spectrum data [6–8]. On the other hand,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 398–410, 2024.
https://doi.org/10.1007/978-3-031-65126-7_34

12. Curtmola, R., Garay, J.A., Kamara, S., Ostrovsky, R.: Searchable symmetric encryption: improved definitions and efficient constructions. In: Conference on Computer and Communications Security, Alexandria, VA, USA, 30 October–3 November 2006, pp. 79–88 (2006)
13. Kamara, S., Papamanthou, C., et al.: Dynamic searchable symmetric encryption. In: Conference on Computer and Communications Security, Raleigh, NC, USA, 16–18 October 2012, pp. 965–976 (2012)

6 Conclusion

Medical cloud storage is growing popularity and the demand for cloud security continues to increase. Thus, we proposed a novel forward secure scheme for high-dimension medical data in cloud server. We prove it is adaptive secure, and the experimental results demonstrate it is efficient in practice.

Acknowledgment. This paper is supported by the National Science Foundation of China under grant No. 61871064, 61501080, and 61771090 the Fundamental Research Funds for the Central Universities under No. DUT19JC08, and the China Postdoctoral Science Foundation under grant No. 2019M661097.

References

1. Song, D.X., Wagner, D.A., Perrig, A.: Practical techniques for searches on encrypted data. In: IEEE Symposium on Security and Privacy, Berkeley, California, USA, 14–17 May 2000, pp. 44–55 (2000)
2. Yang, Y., Ma, M.: Conjunctive keyword search with designated tester and timing enabled proxy re-encryption function for E-health clouds. IEEE Trans. Inf. Forensics Secur. **11**, 746–759 (2016)
3. Li, H., Yang, Y., Dai, Y., et al.: Achieving secure and efficient dynamic searchable symmetric encryption over medical cloud data. IEEE Trans. Cloud Comput. (2017). https://doi.org/10.1109/TCC.2017.2769645
4. Zhang, Y., Xu, C., Li, H., et al.: HealthDep: an efficient and secure deduplication scheme for cloud-assisted e-health systems. IEEE Trans. Industr. Inf. **14**, 4101–4112 (2018)
5. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, 23–26 May 1998, pp. 604–613 (1998)
6. Chang, Y.-C., Mitzenmacher, M.: Privacy preserving keyword searches on remote encrypted data. In: Ioannidis, J., Keromytis, A., Yung, M. (eds.) ACNS 2005. LNCS, vol. 3531, pp. 442–455. Springer, Heidelberg (2005). https://doi.org/10.1007/11496137_30
7. Bost, R.: $\sum o\varphi o\varsigma$: forward secure searchable encryption. In: Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016, pp. 1143–1154 (2016)
8. Kim, K.S., Kim, M., Lee, D., et al.: Forward secure dynamic searchable symmetric encryption with efficient updates. In: Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–03 November 2017, pp. 1449–1463 (2017)
9. Etemad, M., et al.: Efficient dynamic searchable encryption with forward privacy. In: Proceedings on Privacy Enhancing Technologies, vol. 2018, pp. 5–20 (2018)
10. Ghareh Chamani, J., Papadopoulos, D., Papamanthou, C., Jalili, R.: New constructions for forward and backward private symmetric searchable encryption. In: Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018, pp. 1038–1055 (2018)
11. Zhang, Y., Katz, J., Papamanthou, C.: All your queries are belong to us: the power of file-injection attacks on searchable encryption. In: 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, 10–12 August 2016, pp. 707–720 (2016)

Table 1. The characteristics of datasets.

Dataset	Dimension	Number of data objects	Size of hash tables	Size of lists
Epileptic	178	11500	32 KB	2752 KB
HCV	29	1385	32 KB	332 KB
Parkinson	22	5875	32 KB	1406 KB

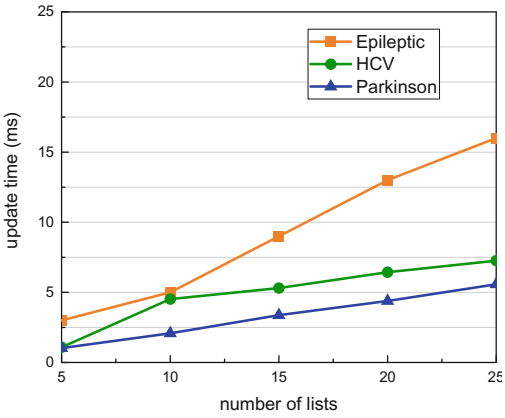


Fig. 4. Update performance.

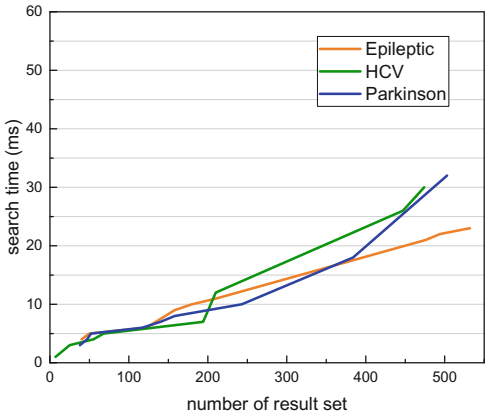


Fig. 5. Search performance.

Game$_{\text{Update}}$: $\mathcal{A}$ Generates random strings as simulated token $\widetilde{t_u}$ when the update object q is sent. Then a random oracle H makes some modifications to the index $\tilde{I}$ based on $\widetilde{t_u}$. Since secure symmetric encryption is semantically secure, $\widetilde{t_u}$ is not computationally indistinguishable from t_u, the index modified based on $\widetilde{t_u}$ and the index modified based on t_u cannot be distinguished.

Game$_{\text{Search}}$: $\mathcal{A}$ Generates random strings as simulated token $\widetilde{t_s}$ when the search object q is sent. Then a random oracle H gets the results $\widetilde{r_s}$ based on $\widetilde{t_s}$. The term $\widetilde{r_s}$ is identical to the real result r_s indicated in $\mathcal{L}_{Search}$. Because secure symmetric encryption is semantically secure, $\widetilde{t_s}$ is not computationally indistinguishable from t_s. Similarly, the results derived from $\widetilde{t_s}$ are the same as the real results. Thus, $\mathcal{A}$ cannot differentiate between simulated $\widetilde{t_s}$ generated by $\mathcal{S}$ and real t_s or between simulated t_s and real $\widetilde{r_s}$.

Conclusion: Summarizing the above game, we can say that for all probabilistic polynomial time adversaries $\mathcal{A}$, there exists a probabilistic polynomial time simulator $\mathcal{S}$, such that:

$$|\Pr\left(\mathbf{Real}_{\mathcal{A}}^{\Omega}(\lambda) = 1\right) - \Pr(\mathbf{Ideal}_{\mathcal{A}}^{\Omega}(\lambda) = 1)| \leq \mathbf{negl}(\lambda)$$

Thus, our proposed scheme is adaptive secure.

Forward Privacy. As described in Sect. 4, we split the hash table and lists and merge all small lists into a large list. Moreover, we store the hash table locally and store the large list in the cloud. Since the search trapdoor t_s for q is generated from the bucket that q mapped to and this bucket updates once a newly added data object p that also maps to this bucket is added to the encrypted dataset in the cloud, the server cannot know the search trapdoor for p until the next query that finds it appears. In other words, the server cannot observe which query request is associated with the update object. Therefore, our proposed scheme is forward secure.

5.2 Performance Analysis

We implement our scheme in Python. The experiments run on a machine with an Intel Core i7-8700U, 3.20 GHz processor, 8 GB RAM, and a 240 GB SSD disk.

The characteristics of the three datasets we used are summarized in Table 1. Moreover, Table 1 shows the index size of each dataset. Specifically, we show the size of the hash table and the size of the lists.

Figure 4 shows the update time. The update time is taken as the average of the update times of all data objects in the dataset. As we can see, the update time is linear with the number of hash tables. Further, we analyze the update time complexity from Algorithm 2 as $O(l)$, where l is the number of hash tables.

Figure 5 shows the search time. We can analyze the search time complexity is $O(r)$, where r is the number of data points in the query result set. This is consistent with the experimental result of Fig. 5.

Algorithm 3 Search

Input: $(K, \sigma, q); (I, C)$
Output: $r_s; \perp$
On the client side:
$i \leftarrow 0$
$t_s \leftarrow \emptyset$
Repeat:
  $Tidx \leftarrow g_i(q)$
  $(key_i, Lidx_i) \leftarrow T_i[Tidx]$
  $t_s \leftarrow t_s \cup (key_i, Lidx_i)$
  $i \leftarrow i + 1$
Until $i = l$
Send t_s to the cloud server
On the cloud server side:
Parse t_s into $(key_1, Lidx_1), \ldots, (key_l, Lidx_l)$
$i \leftarrow 0$
$r_s \leftarrow \emptyset$
Repeat:
  $key \leftarrow key_i$
  $Lidx \leftarrow Lidx_i$
  Repeat:
    $e \leftarrow L_i(Lidx)$
    $mask \leftarrow H(key, Lidx)$
    $(qid||op||Lidx_{pre}||key_{pre}) \leftarrow e \oplus mask$
    $r_s \leftarrow r_s \cup qid$
    $key \leftarrow key_{pre}$
    $Lidx \leftarrow Lidx_{pre}$
  Until $Lidx = \perp$
  $i \leftarrow i + 1$
Until $i = l$
Send r_s to the client

5 Analysis

5.1 Security Analysis

Theorem 1 (Adaptive security)**:** Our scheme with leakage functions $\{\mathcal{L}_{Setup}, \mathcal{L}_{Update}, \mathcal{L}_{search}\}$ is adaptive secure, where $\mathcal{L}_{Setup} = \mathcal{L}'(l)$, $\mathcal{L}_{Update} = \mathcal{L}'(qid, l)$, $\mathcal{L}_{Search} = \mathcal{L}'(r_s)$.

Proof: We give the proof of Theorem 1 through games below.

Game$_{\text{Setup}}$: The simulator $\mathcal{S}$ generates a randomized index $\tilde{I}$ based on $\mathcal{L}_{Setup}$ and $\mathcal{L}_{Update}$, which is the same size as the real encrypted index I. $\tilde{I}$ consists of l lists, and each element in $\tilde{I}$ is a random string with the same length as the real encrypted element. Adversary $\mathcal{A}$ cannot differentiate $\tilde{I}$ from I because of the semantic security of secure symmetric encryption.

Algorithm 2 Update

Input: $(K, \sigma, q, op); (I, C)$
Output: $(\sigma', r_u); (I', C')$
On the client side:
$i \leftarrow 0$
$t_u \leftarrow \emptyset$
$entrynum \leftarrow entrynum + 1$
Repeat:
  $Tidx \leftarrow g_i(q)$
  $(key_{pre}, Lidx_{pre}) \leftarrow T_i[Tidx]$
  key $\xleftarrow{\$} (0,1)^{\lambda}$
  $Lidx \leftarrow entrynum$
  $T_i[Tidx] \leftarrow (key, Lidx)$
  $e_i \leftarrow H(key, Lidx) \oplus \left(qid||op||Lidx_{pre}||key_{pre}\right)$
  $t_u \leftarrow t_u \cup e_i$
  $i \leftarrow i + 1$
Until $i = l$
$q_c \leftarrow Enc(key_q, q)$
$t_u \leftarrow t_u||q_c$
Send t_u to the cloud server
On the cloud server side:
Parse t_u into $(q_c, (e_1, e_2, \ldots, e_l))$
$C \leftarrow C \cup q_c$
$i \leftarrow 0$
Repeat:
  L_i.add(e_i)
  $i \leftarrow i + 1$
Until $i = l$
$r_u \leftarrow$ "$success$"
Send r_u to the client

Search. In Algorithm 3, the search phase is presented. To search similar data objects for data object q, for $\forall i \in [1, l]$, q is mapped to a bucket of T_i through g_i. The search trapdoor t_s is generated with the tuple pairs in the buckets being mapped. Then t_s is sent to the cloud server. After receiving the search trapdoor t_s, the server can locate an element based on the index value $Lidx_i (1 \leq i \leq l)$. The data object and another index information ($Lidx_{pre}$, key_{pre}) hidden in this element are taken out. The server continues the locating operation based on ($Lidx_{pre}$, key_{pre}) until the extracted location information is null tuple pair ($\perp$, $\perp$). In this way, the server collects similar data objects scattered in large lists and returns them to the client as a result of the search operation.

our update operation is to add an encrypted element directly to the end of the large list, the server cannot associate the update object with any previous search. Thus, the storage structure we design can provide forward privacy.

4.3 Our Construction

Setup. Algorithm 1 gives the description of setup phase. It takes as input a security parameter λ and LSH family, selects the symmetric encryption key key_q used to encrypt data objects, determines the concatenation LSH functions $(g_1, g_2, \ldots, g_l)$, and initializes l hash tables $(T_1, T_2, \ldots, T_l)$, l large lists $(L_1, L_2, \ldots, L_l)$ and a variable *entrynum* that represents the number of elements in each large list. Each bucket of the newly initialized hash tables contains a null tuple pair $(\perp, \perp)$. For $\forall i \in [1, l]$, a one-to-one mapping exists between g_i and T_i, and also exists between T_i and L_i. The encryption key key_q and client state $\sigma = \{(g_1, g_2, \ldots, g_l), (T_1, T_2, \ldots, T_l), entrynum\}$ are stored locally, The cloud index $I = (L_1, L_2, \ldots, L_l)$ and the data set C used to hold encrypted data objects are sent to the cloud server. We consider σ and I as the index of our scheme, as shown in Fig. 3.

Algorithm 1 Setup

Input: $(\lambda, \text{LSH family}); \perp$
Output: $(\sigma, key_q); (I, C)$

$i \leftarrow 0$
Repeat:
$j \leftarrow 0$
  Repeat:
    $h \xleftarrow{\$} \text{LSH family}$
    $g_i \leftarrow g_i \cup h$
  Until $j = k$
  $T_i \leftarrow$ empty hash table
  $L_i \leftarrow$ empty list
  $i \leftarrow i + 1$
$entrynum \leftarrow 0$
$\sigma = \{(g_1, g_2, \ldots, g_l), (T_1, T_2, \ldots, T_l), entrynum\}$
$key_q \leftarrow (0,1)^{\lambda}$
$I = (L_1, L_2, \ldots, L_l)$
$C \leftarrow \emptyset$
Send (I, C) to the cloud server

Update. A detailed description of update phase is provided in Algorithm 2. To update a data object q with identiffier qid, for $\forall i \in [1, l]$, q is mapped to a bucket of T_i through g_i. The tuple pair in each bucket is a secret key *key* of the keyed hash function H and an index value *Lidx* of the large list in the cloud. Then the update trapdoor t_u is generated according to the tuple pairs in the buckets being mapped and is sent to the cloud server. Notice that $(e_1, e_2, \ldots, e_l)$ in t_u are elements to be added to the large lists. Therefore, after receiving the update trapdoor t_u, the server adds these elements to the end of l large lists in turn.

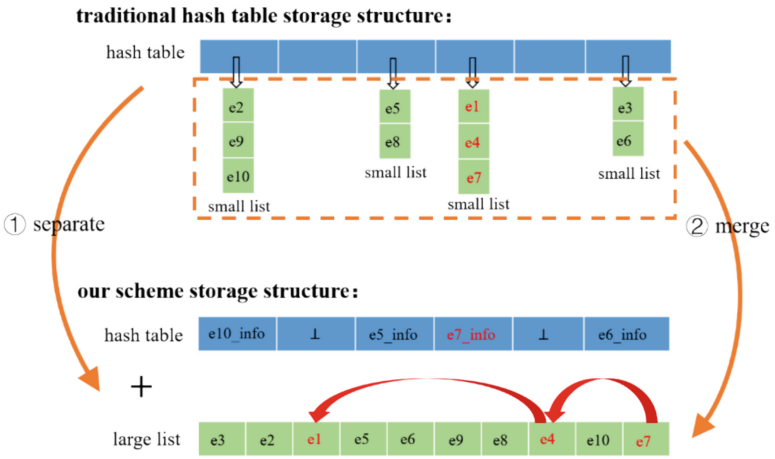


Fig. 2. An example of our storage structure.

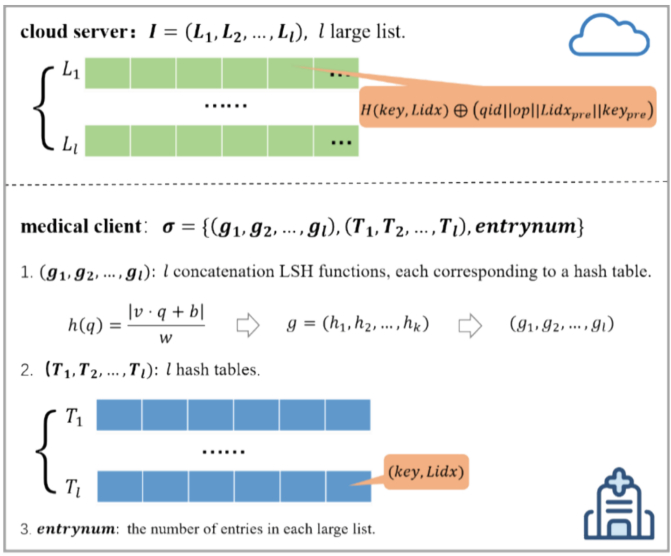


Fig. 3. Index of our scheme.

modified, and a new large list element is generated for q according to the old and new information stored in the bucket. Then this element is sent to the cloud and added to the end of the large list. When performing the search operation, the search object q is locally mapped to a bucket of the hash table by using LSH, and the search token t_s for q is generated based on the information stored in this bucket. Then t_s is sent to the cloud server. The server locates an element of the large list according to the t_s and then locates the next element according to the value of this element until the locating cannot be continued, and then it returns the all data objects traversed as the search result. Because

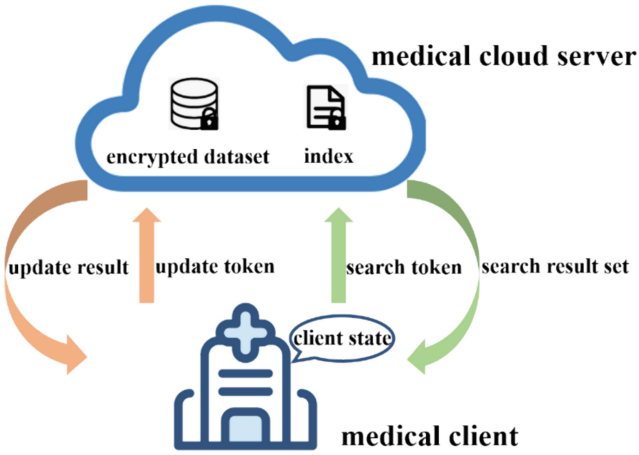


Fig. 1. System model of our scheme.

4.2 Storage Structure

In this paper, the core innovation is the storage structure of the index we designed. Therefore, before we introduce the complete execution flow of the system, we give a detailed introduction to the storage structure of the index. Essentially, we improve the structure of the traditional hash table to design a novel index that can guarantee forward privacy. Thus, we first analyze the security flaws of traditional hash tables and then introduce our index storage structure.

The traditional hash table storage structure is to store a list in each bucket. In previous SSE schemes for high-dimension data, the hash table is always stored in the cloud server as an index. When adding a data object q,q is mapped to one of the buckets by LSH function, and then the ID of q, qid, is added to the list in this bucket. When searching similar objects of a data object q, we map q with LSH function, map it to a bucket, and then the data objects in this bucket's list are returned as similar objects of q. With such storage structure, both query and update operations are mapped to a bucket first, so it is easy for the server to associate the newly added data object with search objects that were previously mapped to the same bucket for searching. Therefore, such a storage structure cannot provide forward privacy guarantee.

We designed a new storage structure that can provide forward privacy. Our storage structure is transformed from the traditional hash table storage structure as shown in Fig. 2. We transform by two steps: separate and merge. First, we separate the hash table from the list in each bucket, which means that the list is no longer stored in each bucket and only a state value information is stored. Then, we merge the lists that existed in each bucket. If we consider the list in each bucket as a small list and the merged list as a large list, then the large list contains all the elements stored in all the original small lists. In our scheme, we store the hash table locally as part of the client state value σ and store the large list as the index I in the cloud.

When performing the update operation, the data object q to be updated is locally mapped to a bucket of the hash table by using LSH, the information in the bucket is

Definition 4 (Adaptive Secure SSE scheme)**:** Let Ω be an SSE scheme with $\mathcal{L} = \{\mathcal{L}_{Setup}, \mathcal{L}_{Update}, \mathcal{L}_{search}\}$, $\mathcal{A}$ be an adversary, S be a simulator, two games are defined as follows:

$\mathbf{Real}_{\mathrm{A}}^{\Omega}$: The adversary a is given C generated by $Setup((\lambda, \text{LSH family}); \perp)$. Then, $\mathcal{A}$ performs adaptive update u or query q. The challenger answers by running $Update((K, \sigma, q, op); (I, C))$ for u. The challenger answers by running $Search((K, \sigma, q); (I, C))$ for q. Finally, $\mathcal{A}$ outputs a bit $b \in \{0, 1\}$.

$\mathbf{Ideal}_{\mathrm{A}}^{\Omega}$: An encrypted index I and an encrypted dataset C are generated by simulator $\mathcal{S}$ through running $\mathcal{S}(Setup)$ based on the leakage function $\mathcal{L}_{Setup}$ and then given to $\mathcal{A}$. Then, $\mathcal{A}$ performs adaptive update u or query q. For u, $\mathcal{S}$ is given $\mathcal{L}_{Update}(q)$, and answers it by running $\mathcal{S}(Update)$. For q, $\mathcal{S}$ is given $\mathcal{L}_{search}(q)$, and answers it by running $\mathcal{S}(Search)$. Finally, $\mathcal{A}$ outputs a bit $b \in \{0, 1\}$.

Ω is adaptively secure with leakage functions $\mathcal{L}$, if for any polynomial-time adversary $\mathcal{A}$, there exists an efficient polynomial-time simulator $\mathcal{S}$ such that:

$$\Pr\left(\mathbf{Real}_{\mathcal{A}}^{\Omega}(\lambda) = 1\right) - \Pr(\mathbf{Ideal}_{\mathcal{A}}^{\Omega}(\lambda) = 1)| \leq \mathbf{negl}(\lambda) \tag{2}$$

where $\mathbf{negl}(\lambda)$ is a negligible function in λ.

Informally, forward privacy requires that an update operation does not reveal whether the newly added data object relates to previous search objects. In this paper, we propose a definition of forward privacy for high-dimension data by extending the definition in [7].

Definition 5 (Forward privacy): An SSE scheme for high-dimension data is forward private if the update leakage function $\mathcal{L}_{Update}$ can be formalized as:

$$\mathcal{L}_{Update} = \mathcal{L}'(qid, l) \tag{3}$$

where qid denotes the identifier of the data object, l denotes the number of hash tables and $\mathcal{L}'$ is stateless.

4 The Proposed Scheme

4.1 System Model

Figure 1 shows our proposed system model composed of two main participants, i.e., the medical client and the medical cloud server. Initially, on the medical client side, the client state σ and the cloud server index I are generated, then I is sent to the cloud server. Then the update and search operations are performed separately by sending update token and search token. Specifically, when the medical client wants to perform an update operation, an update trapdoor t_u is generated based on the client state σ, then t_u is sent to the medical cloud server. The medical cloud server updates index I and the encrypted dataset C according to the received trapdoor t_u. Similarly, when the medical client wants to perform an update operation, an update trapdoor t_s is generated based on the client state σ, then t_s is sent to the medical cloud server. The medical cloud server gets the search result according to t_s and index I, then returns the result to the medical client.

where k is the number of functions in each concatenation function, and $h_{i1}, \ldots, h_{ik}$ are randomly chosen from the H.

We use these concatenation LSH functions to construct hash tables. Therefore:

- If $D(p, q) \leq R$, then $\Pr[g(p) = g(q)] \geq p_1^k$.
- If $D(p, q) > cR$, then $\Pr[g(p) = g(q)] < p_2^k$

To increase the recall, l concatenation LSH functions typically are used to build l hash tables.

3.2 Cryptographic Primitive

A private-key encryption scheme is a tuple that is composed of three algorithms *(Gen, Enc, Dec). Gen* is the probabilistic key generation algorithm that takes a security parameter, λ, as input and returns a secret key, K, where $|K| > \lambda$. *Enc* is the probabilistic encryption algorithm that takes a key, K, and a plaintext, $M \in \{0, 1\}^*$, as input and returns a ciphertext $C \in \{0, 1\}^*$. *Dec* is the deterministic decryption algorithm that takes K and $C \in \{0, 1\}^*$ as inputs and returns $M \in \{0, 1\}^*$.

3.3 SSE Scheme for High-Dimension Data

Definition 3: An SSE scheme for high-dimension data that enables the similarity search, and the dynamic update consists of three algorithms:

- $((\sigma, key_q); (I, C)) \leftarrow Setup((\lambda, \text{LSH family}); \perp)$: This is a probabilistic algorithm run by the data owner. It takes a security parameter, λ, and LSH family as input. It returns the secret key, key_q, and the client's state, σ, to the data owner and returns the encrypted index, I, and encrypted dataset C to the server.
- $((\sigma', r_u); (I\prime, C\prime)) \leftarrow Update((K, \sigma, q, op); (I, C))$: This is a deterministic algorithm run by the client and the server. On the client side, it takes the secret key, K, the client's state, σ, the update object, q, and operation op (*add* or*del*) as input. On the server side, it takes the encrypted index, I, and encrypted dataset,C, as input. It returns the updated client state, σ', and the update result,r_u, to the client, and returns the updated encrypted index, I', and updated dataset, C', to the server.
- $(r_s; \perp) \leftarrow Search((K, \sigma, q); (I, C))$: This is a deterministic algorithm run by the client and the server. On the client side, it takes the secret key, K, the client's state, σ, and a search object, q, as input. On the server side, it takes the encrypted index, I, and encrypted dataset, C, as input. It returns a result set, r_s, to the client.

3.4 Privacy Definition

The security definition in this paper follows the simulation-based model [12]. It is parameterized by a collection of leakage functions, $\mathcal{L} = \{\mathcal{L}_{Setup}, \mathcal{L}_{Update}, \mathcal{L}_{search}\}$. These functions describe the information that the scheme leaks to the adversary.

definitions for SSE schemes. The security notions known as SSE were improved, and a protocol was given that was compatible with the definitions. Later, Kamara et al. [13] constructed the first dynamic and sublinear scheme, although the dynamic SSE had been studied earlier.

Zhang et al. [11] introduced file injection attacks on SSE that highlighted the importance of forward privacy and promoted extensive research on forward privacy. Since then, forward privacy has almost become an essential property for dynamic SSE schemes. Bost et al. [7] presented the forward-privacy scheme known as Sophos using trapdoor permutations. The efficiencies of communication and computation have been improved significantly compared with the ORAM-based schemes. Even so, the computational efficiency has been degraded because the trapdoor permutations are based on a public key algorithm. Later, some forward secure SSE schemes based on symmetric cryptographic primitives were been proposed [8–10].

The emergence of SSE provides a secure and effective technical solution foundation for many cloud storage applications, especially for the medical systems with large amounts of sensitive and private medical data. To ensure the security of medical cloud data, many schemes based on SSE have been proposed [2–4]. Most of these existing schemes retrieve documents by keywords. However, this process does not apply to scenarios where high-dimension data is retrieved by high-dimension data, e.g., searching similar genes for a particular gene or searching similar images for a certain CT image. In this paper, we propose a novel forward secure scheme for high-dimension medical data in cloud server.

3 Preliminaries

3.1 Locality Sensitive Hash (LSH)

LSH is an effective similarity search algorithm in high-dimension space [5]. The basic idea is to hash similar objects into the same bucket with a greater probability than dissimilar objects. When a query operation is performed, first, the query point is hashed into buckets in hash tables, and then all the points in these hit buckets are returned as a result of the query.

Definition 1 (LSH family, H)**:** S is a set of n points, and U is a set of hash values, $S \subset R^d$. A family $H = \{h : S \to U\}$ of functions is called (R, cR, p_1, p_2)-sensitive if $\forall p, q \in R^d$:

- If $D(p, q) \leq R$, then $\Pr[h(p) = h(q)] \geq p_1$.
- If $D(p, q) > cR$, then $\Pr[h(p) = h(q)] < p_2$.

where $c > 1$ and $p_1 > p_2$ for similarity search.

Definition 2 (Concatenation LSH Functions, G)**:** $G = \{g : R^d \to U^k\}$ is a set of concatenation LSH functions, where $\forall g_i \in G$ is composed of k hash functions randomly extracted from G. Formally,

$$g_i(p) = (h_{i1}(p), \ldots, h_{ik}(p)) \tag{1}$$

of medical cloud data, many schemes based on SSE have been proposed [2–4]. These schemes based on SSE technology aim at protecting sensitive medical data and further expand from multiple keywords, security strength and access control. In addition, most of these existing schemes use the form of retrieving documents by keywords. However, such a form does not apply to scenarios where high-dimension data is retrieved by high-dimension data, e.g., searching similar genes for a particular gene or searching similar images for a certain CT image.

In this paper, we propose a novel forward secure scheme for high-dimension medical data in cloud servers. Our scheme uses a combination of SSE and locality sensitive hashing (LSH) [5] as a framework to support searching similar high-dimension data by high-dimension data. LSH is a technology that maps the closer points in high-dimension space to the same hash bucket with a higher probability. In the framework of the combination of SSE and LSH, searching similar high-dimension data by high-dimension data is transformed into searching similar high-dimension data by the hash value of the high-dimension data. In particular, our proposed scheme can achieve forward privacy of SSE by modifying the hash table structure. Forward privacy [6] that requires that the newly added data to not be related to a previous query request is the goal that current dynamic SSE schemes are pursuing [7–10], and it has almost become an essential property since the file-inject attacks [11]. The traditional hash table stores a list of data IDs in each bucket. The data mapped to the same bucket through the LSH function can be regarded as similar data, so the data in the list of the same bucket can be regarded as similar data. Such a structure can easily expose the data's bucket location, so it can reveal the association between the newly added data and the previous query request. In this paper, we separate the hash table and the list, and merge several small lists in all hash buckets into a large list. We store the hash table locally and the large list in the cloud server. The new index makes our scheme forward secure. We summarize the main features of our proposed approach as follows:

Applicability to High-Dimension Medical Cloud Data. We implement searching high-dimension medical data by high-dimension medical data using a combination of SSE and LSH.

Forward Privacy. We modify the traditional hash table structure and design a new index structure allows our scheme to achieve forward privacy.

Practical Implementation and Efficiency. We use two datasets to evaluate our scheme, and the experimental results indicate that our scheme is efficient.

The rest of the paper is organized as follows. Section 2 summarizes related work and background information is given in Sect. 3. Then, we present our scheme in Sect. 4. In Sect. 5, we give the security analysis and experimental results. Finally, Sect. 6 presents conclusions.

2 Related Work

SSE was first proposed by Song et al. [1], and their scheme and their scheme encrypts each word in the document and then allows searching for documents using a keyword. Subsequently, SSE was formalized by Curtmola et al. [12], who developed the security



Forward Secure Searchable Encryption over Medical Cloud Data

Yang Mi, Cheng Guo, Qianqian He(✉), and Xinyu Tang

School of Software Technology, Dalian University of Technology, Dalian, China
he_qianqian@126.com

Abstract. In medical cloud computing, medical data (e.g., electronic health records and diagnostic report) are allowed to be outsourced to the remote cloud server. Since cloud service providers are semi-trusted, it is important to protect the privacy of medical data while ensuring the convenient access of stored data. Several schemes have been proposed for this purpose. In this paper, we propose a novel scheme for high-dimension medical data based on searchable symmetric encryption (SSE) and locality sensitive hashing (LSH), which is able to provide greater security for the stored on the medical cloud—forward privacy, while ensuring efficient search and update performance. Combined with the experimental results, we carried out a detailed security analysis and performance analysis of the proposed scheme. The results show that the server cannot observe which query request is associated with the update object. Therefore, the proposed scheme is forward secure and efficient in practice.

Keywords: Medical cloud computing · searchable encryption · electronic health records · data privacy · forward privacy

1 Introduction

With the increasing development of cloud computing, cloud storage has become a particularly popular data storage solution used by many data owners due to its scalability, high availability, easy managerial advantages and low costs. Like other organizations, health care systems with a large amount of medical data are considering outsourcing their electronic health records to remote cloud servers to reduce storage overhead and management burden. However, cloud storage as a third-party platform is semi-trusted and puts sensitive and private medical data above certain security problems. At the same time, the biggest threat comes from the cloud service provider. Therefore, how to guarantee data security during the process of storage and use of cloud medical data is a significant problem.

To address the above security problem of data in cloud storage, searchable symmetric encryption (SSE) [1] designed to allow users to search over encrypted data has been proposed. Specifically, SSE allows one to encrypt data using a symmetrical encryption algorithm and subsequently to search over the encrypted data. To ensure the security

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 384–397, 2024.
https://doi.org/10.1007/978-3-031-65126-7_33

5. Himmelsbach, M., Müller, A., Lüttel, T., Wünsche, H.J.: LIDAR-based 3D object perception. In: Proceedings of 1st International Workshop on Cognition for Technical Systems, pp. 1–7 (2008)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
7. Wang, D.Z., Ingmar, P.: Voting for voting in online point cloud object detection. Robot. Sci. Syst. **1**(3), 10–15 (2015)
8. Shang, L., Michael, G.: Real-time object recognition in sparse range images using error surface embedding. Int. J. Comput. Vision **89**(2–3), 211–228 (2010)
9. Martinovic, A., Knopp, J., Riemenschneider, H., Van Gool, L.: 3D all the way: semantic segmentation of urban scenes from start to end in 3D. In: Computer Vision & Pattern Recognition, pp. 4456–4465. IEEE (2015)
10. Mirsalar, K., Youngjib, H.: Vision-based volumetric measurements via deep learning-based point cloud segmentation for material management in jobsites. Autom. Constr. **121**, 103430 (2020)
11. Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S.: SEGCloud: semantic segmentation of 3D point clouds. In: 2017 International Conference on 3D Vision (3DV), pp. 537–547. IEEE (2017)
12. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 652–660. IEEE (2017)
13. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: PointCNN: convolution on X -transformed points. In: Neural Information Processing Systems, pp. 1–11 (2018)
14. Wu, B., Wan, A., Yue, X., Keutzer, K.: SqueezeSeg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1887–1893 (2017)
15. Komalpreet, K., Amanpreet, K.: Polarization independent frequency selective surface for marine and air traffic radar applications. Sadhana **47**(2), 81 (2022). https://doi.org/10.1007/s12046-022-01840-3
16. Zhong, Z., Zhang, C., Liu, Y., Wu, Y.: VIASeg: visual information assisted lightweight point cloud segmentation. In: The 26th IEEE International Conference on Image Processing (ICIP 2019), pp. 1500–1504. IEEE (2019)
17. You, H., Li, S., Yifan, X., He, Z., Wang, D.: Tree extraction from airborne laser scanning data in urban areas. Remote Sensing **13**(17), 3428 (2021)
18. Zhou, Y., et al.: End-to-end multi-view fusion for 3d object detection in lidar point clouds. In: Proceedings of the Conference on Robot Learning, PMLR, vol. 100, pp. 923–932 (2020)
19. Liong, V.E., Nguyen, T.N.T., Widjaja, S., Sharma, D., Chong, Z.J.: AMVNET: assertion-based multi-view fusion network for lidar semantic segmentation arXiv:2012.04934 (2020)
20. Zhang, F., Fang, J., Wah, B., Torr, P.: Deep fusionnet for point cloud semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 644–663. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_38
21. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel CNN for efficient 3D deep learning. In: Advances in Neural Information Processing Systems, pp. 1–11 (2019)
22. Ning, Z., et al.: Mobile edge computing and machine learning in the internet of unmanned aerial vehicles: a survey. ACM Comput. Surv. **56**(1), 1–31 (2023)
23. Wang, X., Ning, Z., Guo, L., Guo, S., Gao, X., Wang, G.: Mean-field learning for edge computing in mobile blockchain networks. IEEE Trans. Mob. Comput. **22**(10), 5978–5994 (2022)
24. Ning, Z., et al.: Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE Trans. Mob. Comput. **22**(5), 2628–2644 (2023)

4.3 Ablation Studies

We used the additive fusion method and the attention-based fusion method, respectively, and through the experimental results, we can see that the miou of the additive method reaches 62.7%, and after using the attention-based fusion method, the effectiveness of the feature fusion is greatly improved, and the miou reaches 72.1% (Table 2).

Table 2. Ablation studies on the use of self-attention mechanisms

	Mean IoU	Mean Accuracy
RPNet	62.7	91.7
RPNet + self-attention	72.1	92.2

5 Conclusion

In this paper, by surveying the current practice and literature on point cloud segmentation techniques, we discuss the current research progress related to point cloud segmentation and propose a novel deep learning based lidar point cloud segmentation method. Our method uses range image-based point cloud segmentation to extract the features of point clouds. We also add a point-based branch to directly predict the 3D point cloud data point by point, and make the segmentation results more accurate by fusing the segmentation results of the two branches, which complement each other's features. We evaluate our method on a publicly available lidar point cloud dataset, Semantic KITTI, and demonstrate that it is more effective and robust compared to other fusion algorithms. In future work, we will investigate more accurate segmentation methods by addressing data imbalance in the dataset, optimizing the capability of the algorithm while improving the generalization capability of the model to provide a more accurate reconstruction basis for map reconstruction and other tasks. In addition, we will design lightweight lidar point cloud segmentation method, and facilitate the application of the algorithms in edge computation [22–24].

References

1. Zhao, F., Jiang, H., Liu, Z.: Recent development of automotive LiDAR technology, industry and trends. In: Proceedings of Eleventh International Conference on Digital Image Processing (ICDIP 2019), pp. 1046–1053 (2019)
2. Zhao, Y: Media governance in the new media era. In: Proceedings of 2nd International Symposium on Economic Development and Management Innovation (EDMI 2020), pp. 107–114 (2020))
3. Real-time self-driving vehicle location technology. http://www.taodudu.cc/news/show-4649483.html
4. Sun, P., Sun, C., Wang, R., Zhao, X.: Object detection based on roadside LiDAR for cooperative driving automation: a review. Sensors **22**(23), 9316 (2022)

dataset for semantic segmentation of point clouds, which contains point cloud data collected by LiDAR sensors The Semantic KITTI dataset is an open dataset for point cloud semantic segmentation, which contains point cloud data collected by LiDAR sensors, and semantic labels labeled by human. The Semantic KITTI dataset is divided into a training set and a validation set, where the training set contains 21 sequences and the validation set contains 1 sequence.

We trained 150 epochs on the training set and evaluated the model performance using the validation set. During training, we used learning rate decay to reduce the learning rate to avoid overfitting. A higher learning rate of 10^{-3} is used early in training (first 10 epochs) to converge quickly, and then the learning rate is gradually reduced to enhance model stability. The last few epochs use a smaller learning rate to fine-tune the model and improve accuracy. We also used the early stop method to terminate the training process early to avoid overfitting. Finally, the model that performs best on the validation set is selected for testing.

4.2 Comparison with Other Methods

The experimental results are shown in the following table, while the classical point-based segmentation algorithm, range image-based segmentation algorithm, and with some fusion algorithms are also listed in the table. We mainly choose data from seven main categories of objects to show the result. It can be seen that the proposed method in this paper has improved the value of miou as well as the value of iou for small targets, compared with other methods (Table 1).

Table 1. Experimental results compared to other algorithms

method	car	bicycle	motorcycle	truck	persons	bicyclist	road	Mean IoU	mean accuracy
Pointnet++	53.7	1.9	0.2	0.9	0.9	1.0	72.0	20.1	--
Rangenet++	91.4	25.7	34.4	25.7	38.3	38.8	91.8	52.5	89.0
FusionNet	95.3	47.5	37.7	41.8	59.5	56.8	91.8	61.3	--
SPVNAS	97.2	50.6	50.4	56.6	67.4	67.1	90.2	67.0	--
RPNet	93.5	46.1	48.0	40.9	63.9	54.3	92.7	62.7	91.7
RPNet + self-attention	92.6	52.5	50.7	56.5	62.6	55.7	93.3	72.1	92.2

The experimental results show that our fusion segmentation module can effectively perform the point cloud semantic segmentation task on Semantic KITTI dataset and achieve a good performance. The fusion model has high performance and the final mio value can reach 72.1%.

$$\boldsymbol{f}_{\text{out}} = \sum_{\text{i}=1}^{N} \boldsymbol{w}_i \boldsymbol{f}_i$$

where, $\boldsymbol{f}_i$ denotes the feature vector of the ith point, $\boldsymbol{a}$ denotes a fully connected layer, softmax denotes a softmax function, Z is a normalization constant, $\boldsymbol{w}_i$ denotes the attention weight of the ith point, and $\boldsymbol{f}_{\text{out}}$ denotes the weighted feature vector.

After calculating the attention weights of each point, they are applied to the semantic labels. The attention weights of each point are multiplied by its corresponding semantic label to get the weighted semantic label, and then the weighted semantic labels of all points are summed to get the final semantic label. The weighted semantic labels are calculated by the following equation:

$$\boldsymbol{y}_{\text{out}} = \sum_{i=1}^{N} \boldsymbol{w}_i \boldsymbol{y}_i$$

where, $\boldsymbol{y}_i$ denotes the semantic label of the ith point, and $\boldsymbol{y}_{\text{out}}$ denotes the weighted semantic label.

3.5 Training Pipeline

In this paper, we use the Adam optimization algorithm and cross-entropy loss function to optimize the model. The core idea of the Adam algorithm is to maintain the first-order moment estimates and second-order moment estimates for each parameter and use these estimates to update the parameters. The update of the Adam algorithm The formula is as follows.

In each batch of training, the difference between the output of the model and the true label, i.e., the loss function, needs to be calculated. In this paper we use the cross-entropy loss function to calculate the loss.

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log \hat{y}_i$$

where y is the true label, $\hat{y}$ is the prediction result of the model, and N is the number of categories. The smaller the value of the cross-entropy loss function, the closer the prediction result of the model is to the true label.

After calculating the loss function, the back propagation algorithm is used to calculate the gradient of each parameter and the Adam optimizer is used to update the value of each parameter. The dataset is divided into a training set and a validation set, and the validation set is used to evaluate the performance of the model during the training process.

4 Experiment

4.1 Dataset and Implementation

To implement the deep learning experiments, a GPU-equipped computer with PyTorch framework is required. Dataset preparation: an appropriate point cloud data is selected, here the Semantic KITTI dataset is selected. The Semantic KITTI dataset is an open

3.2 Point-Based Point Cloud Segmentation Network Structure Design

The 3D coordinates of each point and other arbitrary feature vectors are used as input, where the feature vectors include information such as RGB values, normal vectors and curvature. A shared two-layer fully connected layer is used to process the feature vectors of each point, and a maximum pooling layer is used to aggregate the feature vectors of each point. Finally, a fully-connected layer is used to map the aggregated feature vectors to different semantic classes.

3.3 Range Image-Based Point Cloud Segmentation Network Structure Design

Using the projection function $P : R^{N\times(3+C)} \rightarrow R^{(M\times D)}$, the points in 3D are projected into the 2d plane to obtain a representation of the points of the point cloud on the 2D image. A hash map from the point cloud form to the 2D form is created using the build hash function, and the features are passed from the 2D image onto the point-based branches using a bilinear interpolation method with the following equations and partial derivatives:

$$F_p(i) = \Phi(\delta(j), F_R) = \sum_{u\in\delta(j)} \phi(u, j)F_R(u)$$

$$\phi(u, j) = (1 - \lfloor j_x - u_x \rfloor)\left(1 - \left\lfloor j_y - u_y \right\rfloor\right)$$

An encoder-decoder hourglass type architecture is used, inspired by the Darknet53 backbone. The encoder allows to encode contextual information and the decoder up-samples features extracted by the convolutional backbone encoder to the original image resolution. A total of three times of point cloud feature propagation and feature fusion are performed in downsampling and upsampling to complement each other's feature information.

3.4 Attention-Based Feature Fusion Method

The attention mechanism can weight the different local features of the point cloud data, which makes the features of each point more accurately represented. In contrast, other fusion methods such as voting fusion or averaging fusion simply count the results and do not reflect the local features of the point cloud data well. Also for the phenomenon of data category imbalance, i.e., some categories have more or less data than others. Using the attention mechanism can weight the data of different categories according to their characteristics, so as to better deal with the unbalanced data. Therefore, the attention-based fusion approach is chosen in this paper to better integrate the feature vectors extracted from the two branches.

The self-attention mechanism is selected to calculate the attention weight of each point by the following equation:

$$\boldsymbol{w}_i = \frac{1}{Z}\text{softmax}(\boldsymbol{a}(\boldsymbol{f}_i))$$

combining point information from bird's-eye and distance maps, which are then fed into a subsequent network. AMVNet [19] designs a late fusion method by computing the uncertainty of the outputs of different views and using an additional network to refine the results. FusionNet [20] proposes a point-voxel interaction MLP that aggregates features between adjacent voxels and corresponding points, reduces the time required for adjacency search, and achieves acceptable accuracy for large point clouds. In particular, PVCNN [21] proposed an efficient point-voxel fusion method. In this method, voxels provide coarse-grained local features, while points provide fine-grained geometric features by performing simple MLPs on each point.

3 Method

3.1 Network Architecture

Aiming at the problems such as semantic loss caused by lidar point cloud segmentation projecting 3D point cloud into 2D range image, inspired by the method of rangenet++, we design a two-branch segmentation model, which takes RANGE-BASE and POINT as inputs, and through the feature fusion method based on the mechanism of self-attention, fuses the information of different modalities and supplements the information lost due to the change of dimensionality, so as to obtain a more accurate segmentation output (Fig. 1).

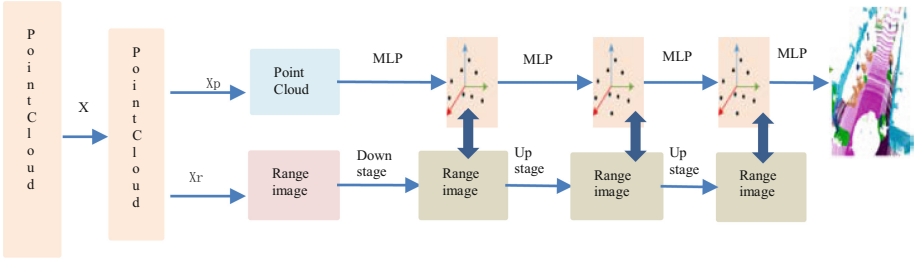


Fig. 1. Shows the network architecture of the method in this paper.

The method we proposed, named RPNET, contains two branches, the R branch and the P branch. First we preprocess the point cloud data X to obtain the R branch input X_r and the input X_p of the P branch. The R branch transforms the points in the three-dimensional space by the spherical coordinate to the point represented by the range image representation in 2D space, uses darknet as baseline for semantic segmentation, maps the fused features to the semantic label space using the fully connected layer, and outputs the semantic labels of each point to obtain the feature F_a. The P branch contains multiple MLP structures, and directly performs point-by-point feature extraction on the 3D point cloud to obtain the feature F_b. Attention mechanism is utilized to dynamically adjust the weight of the features obtained from the two branches in order to improve the fusion effect. The fused features are centralized in the P branch, and through the MLP structure, the point cloud reconstruction is performed to regenerate the point cloud after semantic segmentation.

manual features called feature descriptors or descriptors. These features are coupled to the point cloud density [10] and it is time consuming to find these relevant features from a large number of point clouds due to the presence of noise and inhomogeneous density. Although some acceleration algorithms have made improvements to the extraction time, they are changes on small-scale data scenarios, which are difficult to generalize to large-scale complex scenarios. Moreover, the multi-stage processing may bring the accumulation of errors. Most importantly, the time consumed by manual feature extraction is unacceptable for real-time applications, while autonomous driving requires more and more real-time and robustness of algorithms, and traditional methods are not adapted to such scenarios.

In addition to traditional methods for point cloud segmentation, there is also deep learning-based semantic segmentation of point clouds. Deep learning [11] has the unique advantage of being able to automatically extract data features using convolutional networks and enables end-to-end training. According to the different data representations of point cloud input networks, there are three main categories of voxel-based and 3D convolutional methods, disordered point cloud-based methods, and point cloud 2D processing-based methods for semantic segmentation.

Deep learning methods based on disordered point clouds. This category of methods is to input the point cloud directly into the deep network for processing. When processing point cloud data, this class of methods has to solve the problem of disorderly input of point clouds as well as to ensure that the spatial transformation of point clouds remains invariant. Pointnet proposed by Qi et al. [12] pioneered the method of using deep learning network processing directly on point cloud data, but Pointnet focuses more on global features and does not learn enough local features of point clouds, which causes it to lose the ability to capture The ability to capture spatial relationships between features is lost, which limits its applicability to complex scenarios. The learning of local features is crucial to the segmentation task, and many research methods have made improvements to address this problem. For example, the PointCNN framework proposed by U et al. [13] designs an X-Conv algorithm to learn local region features. It not only improves the accuracy but also reduces the network complexity than Pointnet.

In 2018, SqueezeSeg [14] developed by Wu et al. used spherical projection for point clouds to support the use of 2D convolution and proposed a real-time fully convolutional semantic segmentation method. The point cloud 2Dization method, although it loses dimensional information and is not a leader in terms of accuracy, it relies on the maturity of 2D deep learning algorithms and is cost-effective in terms of real-time and spatial complexity, and can be used in many small or specific scenarios such as road scenes [15]. The earliest deep voxel network is the VoxNet network architecture proposed by Maturana and Scherer [16], which is used for point cloud target detection and classification tasks. With the application and development of fully convolutional neural networks in images, Tchapmi et al. [17] were inspired to propose a three-dimensional fully convolutional network architecture, SEGCloud.

2.2 Fusion Segmentation Methods

Since single views are more or less problematic, some recent approaches try to fuse two or more different views. For example, the approach in [18] performs early fusion by

dataset, Semantic Kitti, and demonstrate that it is more effective and robust compared to other fusion algorithms. The contributions of this paper are as follows:

1. For the large-scale point cloud segmentation problem, we designed an end-to-end dual-branching model, which directly extracts the point cloud data branches to obtain the semantic information of each specific point, and at the same time, uses the point cloud projected as a branch of the range image to obtain the local semantic information, which solves the problem of achieving a better trade-off between the running speed and the segmentation performance in the point cloud semantic segmentation algorithm.
2. An attention-based inter-branch fusion approach is adopted, and a post-processing algorithm is used to further enhance the semantic segmentation results of the full point cloud as the range image segmented by the 2D semantic segmentation method produces a fuzzy output.
3. we conducted semantic segmentation experiments on Semantic KITTI dataset and proved the effectiveness of this method, the miou achieved by this method is 72.1%, which is 52% higher than the basic pointnet method, 19.6% higher than the basic range image based segmentation method and higher than other fusion method.

The rest of this paper is organized as follows. Section 2 reviews the related work on lidar point cloud segmentation, including both traditional and deep learning-based approaches. Section 3 describes the proposed method in detail, including the network structure, loss function, and training strategy. Section 4 describes the experimental setup and presents the experimental results. Section 5 concludes the paper and discusses future work.

2 Related Work

2.1 Point Cloud Segmentation Method

The point cloud obtained from laser scanning contains most of the ground points, and this high redundancy can cause problems for subsequent processing such as detection and classification of the target point cloud, so traditional point cloud semantic segmentation methods generally include multiple stages: first, the ground points need to be segmented off, then the remaining point clouds are individually formed into blocks, and then operations such as classification are performed based on the features extracted from each point cloud block [5].

Semantic segmentation of point clouds based on traditional methods can be divided into two categories: methods based on purely mathematical models and geometric inference techniques and methods based on machine learning. For example, the methods based on triangular mesh surface for region growth and the model fitting methods based on RANSAC [6] proposed in the literature, etc. These methods can achieve point cloud segmentation faster by fitting linear and nonlinear models to point cloud data using the basic features and geometry of point clouds. Machine learning-based methods use typical supervised learning algorithms including support vector machines (SVMs) [7, 8], random forests [9], etc., and have achieved more successful results in some tasks such as 3D model detection and segmentation. However, such methods usually rely on a set of

one of the indispensable sensing sensors for autonomous driving systems [2]. The perception module of autonomous driving relies on various sensors, such as cameras, lidar, positioning devices, etc., to collect information about the surrounding environment and the vehicle's own position. The perception module then processes and analyzes the collected information with perception algorithms [3]. The decision making and control of the autonomous driving system depend on the information provided by the perception module, so it is very important to perform accurate perception in all aspects. Lidar has the advantages of being independent of light conditions, fast frequency, high accuracy, and rich information collection [4]. By scanning the surrounding environment, it can provide the sensing system with point cloud data that contain high-precision environmental information.

The processing of lidar point clouds involves many tasks, such as filtering, segmentation, and object recognition. Among these tasks, point cloud segmentation is a key step, aiming at segmenting the point cloud into different semantic parts and delivering this information to the autonomous driving system to help the system realize real-time, robust and accurate perception of the surrounding environment. Exploring the semantic segmentation technology of point cloud and realizing the real-time accurate segmentation of point cloud using deep learning methods is a challenging and realistic task, which is of great significance for the development of autonomous driving technology and intelligent transportation as well as the promotion of its application in other large-scale point cloud semantic segmentation tasks.

Lidar point cloud segmentation requires accurate allegorical segmentation of the scene, which increases the difficulty of point cloud semantic segmentation due to the complex characteristics of the point cloud itself, such as sparsity, large volume, and high noise, and the limitation of the equipment's computational capability. Traditional segmentation methods, such as geometric feature-based point cloud segmentation, use geometric models to fit point cloud images, which are difficult to meet the real-time and generalization requirements of complex scenes. Deep learning has been gradually applied to 3D point cloud data processing and become a mainstream research method due to its advantages such as automatic feature extraction. Although large-scale semantic segmentation systems require multi-sensors, working together and assisting each other, in this paper, we hope to take the point cloud as an independent information source, and further improve the accuracy of point cloud segmentation under the deep learning method by researching the segmentation model, fusion method and other parts.

In this paper, we propose a novel lidar point cloud segmentation method based on deep learning. Our method uses range image-based point cloud segmentation, which takes a 3D point cloud, projects it to obtain a 2D range image, and performs a convolutional segmentation of the image, which is based on Darknet's network architecture and aims to capture both local and global features of the point cloud. To complement the features of the point cloud, point-based branches are also added to directly project the 3D point cloud data point by point, and the segmentation results of the two branches are fused to make the segmentation results more accurate. In order to address the efficiency of fusion and the fusion effectiveness based on both considerations, we also investigate novel fusion methods. We evaluate our method on a publicly available lidar point cloud



A Novel Method for Semantic Segmentation on Lidar Point Clouds

Fei Wang[1], Liangtian Wan[1(✉)], Yan Zhu[2,3], Lu Sun[4], Xiaowei Zhao[1], Jianbo Zheng[2,3(✉)], and Xianpeng Wang[5]

[1] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China
wangfei2021@mail.dlut.edu.cn, {wanliangtian, xiaowei.zhao}@dlut.edu.cn

[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{1120210386,jianbo.zheng}@smbu.edu.cn

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

[4] Department of Communication Engineering, Institute of Information Science Technology, Dalian Maritime University, Dalian 116026, China
sunlu@dlmu.edu.cn

[5] School of Information and Communication Engineering, Hainan University, Haikou 570228, China
wxpeng2016@hainanu.edu.cn

Abstract. Autonomous driving relies on multiple sensors, such as lidar and cameras, to perceive the surrounding environment and the vehicle's own position. Among them, lidar point cloud segmentation is a crucial and challenging task for 3D scene understanding. In this paper, we propose a novel deep learning method RPNet for lidar point cloud segmentation that combines range image-based segmentation and point based segmentation. Our method extracts point cloud features from range images and predicts 3D point cloud labels from point clouds. The segmentation results of both branches are fused to improve accuracy. We evaluate our method on the Semantic KITTI dataset and show that it outperforms other fusion algorithms in terms of effectiveness and robustness.

Keywords: Semantic Segmentation · Lidar Point Clouds · Deep Learning

1 Introduction

Lidar is a remote sensing system that uses a pulsed laser beam to measure the position, velocity, and other characteristics of target objects. It can sense the surrounding environment and infer high-precision three-dimensional information [1], LiDAR has become

This work is supported by National Natural Science Foundation of China (62101088, 61801076), National Natural Science Foundation of Liaoning Province (2022-MS-157, 2023-MS-108, Fundamental Research Funds for the Central Universities (3132023250), the Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 374–383, 2024.
https://doi.org/10.1007/978-3-031-65126-7_32

References

1. Yang, L., Li, C., Ding, Q., et al.: Combining lexical and semantic features for short text classification. Procedia Comput. Sci. **22**, 78–86 (2013)
2. Pradhan, R., Sharma, D.K.: A hierarchical topic modelling approach for short text clustering. Int. J. Inf. Commun. Technol. **4**, 20 (2022)
3. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. Adv. Neural. Inf. Process. Syst. **26**, 3111–3119 (2013)
4. Reynolds, A.P., Richards, G., Rayward-Smith, V.J.: The application of K-Medoids and PAM to the clustering of rules. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. Lecture Notes in Computer Science, vol. 3177, pp. 173–178. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28651-6_25
5. Aggarwal, C.C.: An introduction to uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng. **21**(5), 609–623 (2009)
6. Lu, M., Feng, G., Fan, M., et al.: New clustering algorithms for large data processing. Syst. Eng. Electron. **5**, 1010–1015 (2014)
7. Gullo, F., Ponti, G., Tagarelli, A.: Clustering uncertain data via k-medoids. In: Greco, S., Lukasiewicz, T. (eds.) Scalable Uncertainty Management, pp. 229–242. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87993-0_19
8. Xie, L., Li, L.: Research on multi-attribute group decision under interval number information. Comput. Eng. **40**(10), 210–213 (2014)
9. Zhou, B., Xu, Y., Tang, Q.: New method for determining optimal number of clusters in K-means clustering algorithm. Comput. Eng. Appl. **46**(16), 27–31 (2010)
10. Cheng, X., Yan, X., Lan, Y., et al.: BTM: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)
11. Gu, B., Sung, Y.: Enhanced reinforcement learning method combining one-hot encoding-based vectors for CNN-based alternative high-level decisions. Appl. Sci. **11**(3), 1291 (2021)
12. Bengio, Y., Schwenk, H., Senécal, J.S., Morin, F., Gauvain, J.L.: Neural probabilistic language models. In: Holmes, D.E., Jain, L.C. (eds.) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol. 194, pp. 137–186. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-33486-6_6
13. Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: 24th International Proceedings on Machine Learning, pp. 641–648, ACM, Oregon (2007)
14. Blei, D.M., Ng, A., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
15. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
16. Hall, P.: The SMART retrieval system: experiments in automatic document processing. Inf. Storage Retrieval **9**(3), 199 (1971)

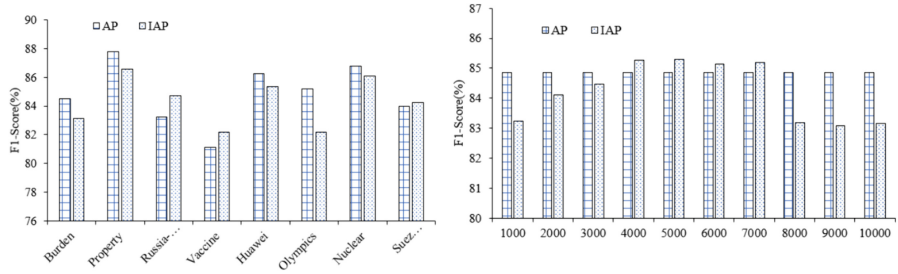


Fig. 5. Comparison for different clustering algorithm

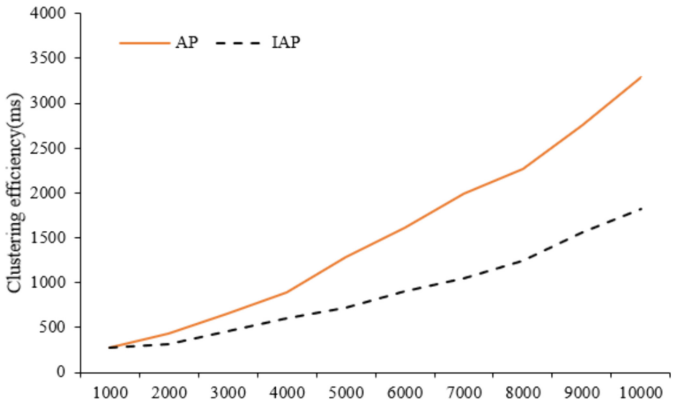


Fig. 6. Comparison for cluster efficiency

6 Conclusions

This article proposes a short text vector representation model that integrates weights and topic features. At the same time, it combines the improved incremental AP clustering algorithm to cluster the short text data of Sina Weibo and discover hot topics that the public is concerned for a period of time. HVRM mines short text data features from local and global perspectives, which solves the problem of feature information loss caused by sparse short text data. The improved incremental AP clustering solves the problem of asynchronous updating processes of responsibility and availability matrices and improves the AP clustering efficiency. Experimental results show that the short text vector representation model proposed in this paper can effectively improve clustering performance.

In future work, other feature factors of short text will be considered, such as the impact of the timeliness of short text on clustering algorithms to discover hot issues.

Acknowledgments. This work is partially supported by the Liaoning Social Science Planning Fund Project under Grant L14AGL002, L13AGL002 and the Liaoning Soft Science Foundation Project under Grant 22022JH4/10100052.

influence of feature words through TF-IDF method. At the same time, the integration of BTM does not significantly increase the feature dimensions of short text, so it does not reduce the efficiency of the algorithm.

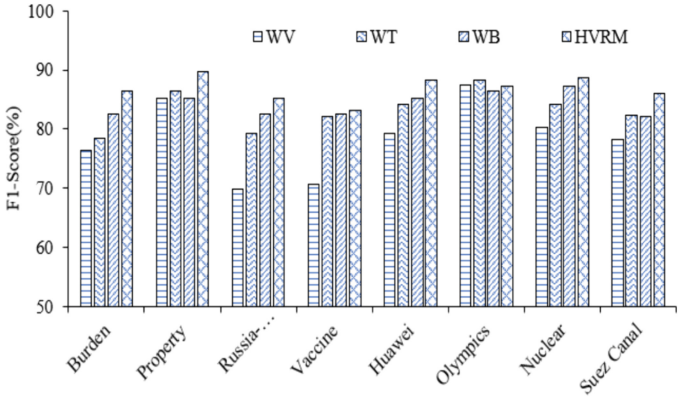


Fig. 4. Comparison for different feature vector models

5.3 Comparison AP and IAP

To verify the performance of IAP, we compare F1-Score values of each clustering after clustering the overall dataset using static AP and IAP. Set the whole dataset as C and randomly divide C into 10 subsets. When performing IAP, add one subset at a time. Figure 5 shows the experimental results.

We can see that the average F1-Score of clustering is 84.85%, while the average F1-Score value of IAP is 84.30%, with little difference between the two algorithms. When gradually adding subsets, their clustering accuracy is lower than the overall clustering accuracy at first, due to the uneven random extraction of subsets, resulting in lower accuracy than AP. But as the data increases, its clustering accuracy will increase, and the overall F1-Score of two algorithms are similar, indicating that IAP is effective. However, as shown in Fig. 6, the IAP efficiency is higher than AP.

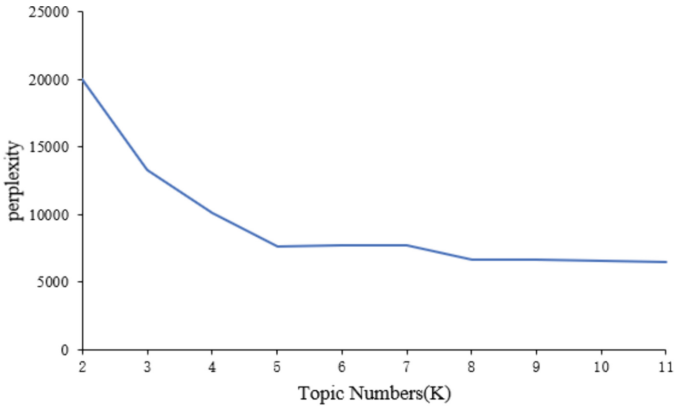


Fig. 3. Perplexity of BTM

each topic. As shown in Fig. 3, the perplexity curve tends to be stable when K = 8, which is consistent with the real topic category of the dataset.

In addition, we compare the top10 keywords of BTM and LDA under each topic. Through these keywords, it is possible to understand the content expressed by each topic. However, due to the wide coverage of some topics, such as the "Tokyo Olympics" includes multiple sub topics. Therefore, the topic center is not prominent. Comparing the keywords determined by BTM and LDA, we find that some keywords determined by LDA are not related to the topic, such as the keywords "OPPO" in the topic "Huawei, Chip", and the keywords "flu" in the topic "Vaccine". This further indicates that BTM is more suitable for short text mining.

5.2 Clustering Accuracy Analysis

In the experiment, the dimension of word vector is set to 100, and the window size is set to 4. According to the perplexity of BTM, the document theme dimension is set to 8, the hyper-parameters $\alpha = 0.5, \beta = 0.01$. The number of iterations in the Gibbs sampling process is set to 1000.

We select four models: Word2Vec(WV), Word2Vec+TF-IDF(WT), Word2Vec+BTM(WB), and Word2Vec+TF-IDF+BTM(HVRM) to obtain short text vector. Then, AP algorithm is used to cluster the short text and get F1-score value for each clustering. Figure 4 shows the results of short text clustering under various feature vector models. It can be seen HVRM proposed in this paper has the best clustering effect. Except for the topic of "Tokyo Olympic", the overall accuracy is higher than other models. The reasons mainly include: Word2Vec focuses on local relationships between words to reduce the dimension of text vectors, while ignoring the global semantic information and the contribution of different words to the topic. So, there is a certain gap in clustering accuracy compared to HVRM. To some extent the feature weights or topic features of words enhances the topic features of word vectors, so the clustering effect is significantly higher than Word2Vec. HVRM not only overcomes the global semantic loss, but also enhances the

Given the responsibility matrix R_t and availability matrix A_t, for the new data $ID_{i'}(t + 1 \leq i' \leq t + p)$, calculate its similarity with the original data, and obtain K-nearest neighbors of the new data, that is, $KNN(I_{i'}) = \{D_{q1}, D_{q2}, \ldots D_{qk}\}$, where $1 \leq q_1, q_2, \ldots q_k \leq t$. The extended responsibility matrix is shown in formula (11):

$$R_{t+p}(i,j) = \begin{cases} R_t(i,j) & i \leq t, j \leq t \\ \frac{1}{k}\sum_{m \in KNN(i)} R_t(m,j) & i > t, j \leq t \\ \frac{1}{k}\sum_{m \in KNN(j)} R_t(i,m) & i \leq t, j > t \\ 0 & i > t, j > t \end{cases} \tag{11}$$

Similarly, the availability matrix A_{t+p} is defined as follows:

$$A_{t+p}(i,j) = \begin{cases} A_t(i,j) & i \leq t, j \leq t \\ \frac{1}{k}\sum_{m \in KNN(i)} A_t(m,j) & i > t, j \leq t \\ \frac{1}{k}\sum_{m \in KNN(j)} A_t(i,m) & i \leq t, j > t \\ 0 & i > t, j > t \end{cases} \tag{12}$$

The process of IAPC algorithm is:

For the original data, the responsibility information of each data in the similarity matrix is updated, and the availability information is calculated;

Update the availability information and calculate the availability of each data;

Sum up the responsibility and availability information of each data to make a decision. If the preset number of iterations is reached, or the clustering center no longer changes, or the decision of the sample data within a sub region does not change after several iterations, the iteration can be terminated.

Calculate the K-nearest neighbors of each new data at regular intervals, construct the responsibility and availability matrices, and continue iteration until convergence.

5 Experiments

The dataset comes from the Sina Weibo in March 2021, including a total of 10125 pieces of data. The data have 8 topics which are "Burden Reduction", "Property Market", "Russia-Ukraine Conflict", "Vaccine", "Huawei Chip", "Tokyo Olympics", "Japanese nuclear wastewater", "Suez Canal ship grounding". In the stage of pre-processing, the data is denoised, including removing stop words and repeated text. Finally, we obtained 9832 valid data. The experimental environment is Intel(R) Core(TM) i5-12600KF 3.70 GHz, 16.0GB memory, 500GB hard disk, Windows 11 operating system, and Python programming language.

5.1 Determination of the Number of Topics

In order to verify the feasibility of BTM in the short text clustering process, first we calculate the perplexity of different topic numbers and determine the topic category K through the perplexity curve. At the same time, we extract the top10 keywords under

this paper selects BTM to find the topic distribution characteristics of the text through the probability distribution of lexical co-occurrence information and documents, topics, words. BTM first assigns a unique serial number i-1 to each word w_i in the corpus dictionary W. At the same time, it constructs the structured short text $D_i = (w_{i1}, w_{i2}, w_{i3},\ldots, w_{in})$ after pre-processing, where w_{in} represents the serial number corresponding to the n_{th} word in short text D_i. Input the structured short text of the corpus and the short text set into BTM, BTM forms word pair set B by extracting all co-occurrence word $b = (w_i, w_j)$, and then randomly assigns topic z to the co-occurrence word pair. After Gibbs sampling, the document topic distribution matrix θ is obtained. Each line of θ represents the probability distribution of short text under each topic, which constitutes a document topic vector.

3.3 Connection of Short Text Vector

BTM and Word2Vec have their own emphasis when vectoring short text. To highlight their respective characteristics, we connect the text weighted word vector with BTM's document topic vector to form a document feature vector R, which makes up for the shortcomings of BTM and Word2Vec and enriches the semantic information of short text vector.

$$R_i = \delta_i \oplus \theta_i \tag{10}$$

4 Incremental AP Clustering

The topic of short text has the characteristics of wide range, strong timeliness, and fast update. The static clustering algorithms cannot meet the demand for real-time discovery of hot topic. In order to quickly discover relevant hot topics in massive changing data, this paper proposes an incremental AP clustering algorithm(IAPC). Assuming that the original short text dataset $D = \{D_1, D_2, \ldots, D_t\}$, the new short text dataset is $I = \{ID_1, ID_2,\ldots ID_p\}$, $D \cap I = \varnothing$, and $t + p = n$. IAP first clusters the original short text dataset to obtain the responsibility and availability matrices. Secondly, we use KNN to obtain the extended responsibility and availability matrices of the newly added data, and then continue iteratively passing the message until convergence.

After clustering the original dataset, the data points in the dataset have accumulated a lot of support. That is, through message transmission, the responsibility and availability matrices are non-zero. For the new data, the responsibility and availability between the new data and other data are zero. If we continue with the previous message transmission, the differences in the responsibility and availability make it difficult for the new data to become the new clustering center. If the data are similar, they not only have a high probability of belonging to the same clustering, but also should have similar responsibility and availability matrices. Based on this, we use KNN to construct the responsibility and availability matrices of the new data, where the corresponding values in the responsibility and availability matrices of the new data are replaced by the values of the K-nearest neighbors of the data points.

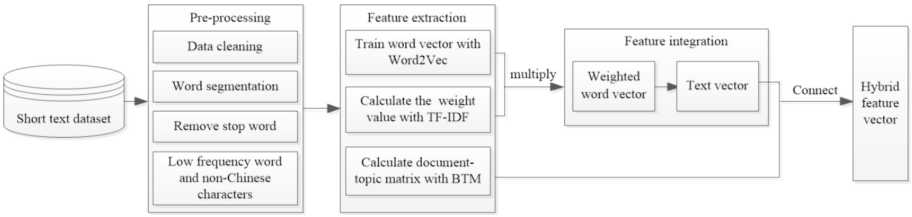


Fig. 2. Hybird vector representation model

3.2 Representation of Document Weighted Vector

Given a short text set $D = \{D_1, D_2, D_3, \ldots, D_m\}$, each short text is divided into several words. We use Word2Vec to obtain the word vector. The vector of w_i is represented as $w_i = (w_{i1}, w_{i2}, \ldots, w_{id})$ in the word list, where d represents the dimension of each word vector, and the size of word list is n.

Word2Vec reflects the semantic association between short text vocabulary, but Word2Vec only focuses on a certain range of vocabulary relationships, which can easily lead to the loss of global information. For the whole short text set, the contribution of each word to the topic of a certain short text is different. Therefore, this paper uses TF-IDF to calculate the weight value of w_i in the short text D_j. As shown in formula (5), where $tf(w_i, D_j)$ represents the frequency of word w_i appearing in a single short text D_j, $idf(w_i)$ represents the weight of word w_i in set D, $N(w_i, D_j)$ represents the frequency of word w_i appearing in D_j, $N(D_j)$ represents the total number of word in set D, and $N(w_i)$ is the number of short text where word w_i appears.

$$K(wi, Dj) = \frac{tf(w_i, D_j) \times idf(w_i)}{\sqrt{\sum_{w_i \in D_j} [tf(w_i, D_j) \times idf(w_i)]^2}} \tag{5}$$

$$tf(w_i, D_j) = \frac{N(w_i, D_j)}{N(D_j)} \tag{6}$$

$$idf(wi) = \log\left(\frac{M}{N(w_i)} + 0.01\right) \tag{7}$$

Assuming the short text $D_j = (w_1, w_2, \ldots, w_t)$, which w_i is a word that corresponds to a vector. We multiply the word vector of w_i and the weight of w_i in the short text to obtain the weighted word vector.

The vector representation of each short text in set D is shown in formula (9)

$$y_i = w_i * K(w_i, D_j) \tag{8}$$

$$\delta_i = \frac{\sum_{j=1}^{t} y_j}{|D_i|} \tag{9}$$

Word2Vec reflects the semantic association between lexical sequences and ignores the global semantics of the text. In order to make up for the lack of global information,

samples by calculating similarity. Give any two data objects i and j, $S(i, j)$ represents the suitability of data point i as the clustering center of data point j. The diagonal of the similarity matrix indicates the data point i as the reference of the clustering center. The larger the value, the more likely it will become the clustering center.

In order to obtain a suitable clustering center, AP clustering needs to continuously transmit information during the iteration process. We use R and A to represent the responsibility and availability matrix in message passing. $R(i, k)$ indicates that data i sends information to data k, reflecting the degree to which data k serves as the clustering center of data i. $A(i, k)$ indicates that data j sends information to data i, reflecting the suitability of data point k as the clustering center for data i. Initially, the availability matrix A is initialized to zero. Then calculate the responsibility matrix R using the following formula:

$$R_{t+1}(i,k) = (1-\lambda)R_{t+1}(i,k) + \lambda R_t(i,k) \tag{1}$$

$$R_{t+1}(i,k) = \begin{cases} S(i,k) - \max\limits_{j \neq k}\{A_t(i,j) + R_t(i,j)\},\ i \neq k \\ S(i,k) - \max\limits_{j \neq k}\{S(i,j)\},\ i = k \end{cases} \tag{2}$$

The iterative formula for the availability matrix is as follows:

$$R_{t+1}(i,k) = (1-\lambda)R_{t+1}(i,k) + \lambda R_t(i,k) \tag{3}$$

$$A_{t+1}(i,k) = \begin{cases} min\{0, R_{t+1}(k,k) + \sum_{j \notin \{i,k\}} max(0, R_{t+1}(j,k)\},\ i \neq k \\ \sum_{j \neq k} max\{0, R_{t+1}(j,k)\},\ i = k \end{cases} \tag{4}$$

The availability matrix represents the sum of self attraction $R(k, k)$ and positive attraction obtained by other points, i.e. $R(k, k) > 0$, because only positive attraction supports k as the clustering center. During the implementation of the algorithm, a damping coefficient λ with a value of [0.5,1] is used to prevent numerical oscillations during the update process.

3 Hybrid Vector Representation Model

3.1 Construction of Hybrid Vector Representation Model

HVRM combines weights and subject features. Figure 2 depicts the process of generating short text vectors. The work mainly includes word segmentation, removing stop word, low frequency word and word of non-Chinese characters, filtering out low-quality and repeated text. Then, for the stage of processed short text, Word2Vec training is used to obtain the word vector representation of each word segment. TF-IDF algorithm [16] is used to calculate the weight value of each word in the short text, and the word vector is multiplied with the weight of the word in the short text to obtain the weighted word vector. BTM is used to obtain the document topic distribution matrix. Finally, we connect the weighted vector with BTM document topic vector to form a short text feature vector.

2.2 BiTerm Topic Model

The topic model is an unsupervised learning algorithm that automatically organizes, searches, and understands a lot of documents. Such models can find topics that span many documents together. Latent Dirichlet Allocation (LDA) builds a model by calculating the importance of terms in documents [14]. When analyzing hot topics, it is difficult to determine the importance of words due to the short text, which leads to data sparsity. To address this defect, BiTerm Topic Model (BTM) obtains a pair of unordered words to model and learn the whole data corpus. This method avoids the sparse short text data in topic modeling.

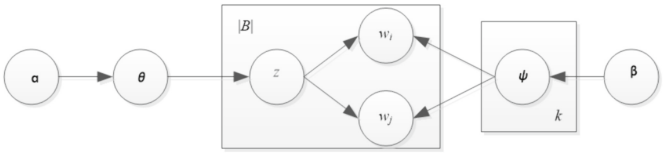


Fig. 1. BTM model

The structure of BTM is shown in Fig. 1, where α, β is the Dirichlet prior parameter, the matrix θ represents the probability distribution of the document topic, and k represents the number of topics. Each line of the matrix θ represents the probability distribution of each document under each topic, such as $\theta_i = (z_{i1}, z_{i2}, ..., z_{ik})$ is the topic vector of the document D_i.

The matrix φ is the probability distribution of topic words. Each column of the matrix φ represents the probability distribution of each word under each topic, such as $\varphi_i = (z_{1i}, z_{2i}, ..., z_{ki})$ is the topic vector of the word w_i in the vocabulary.

The number of word pairs is |B| in corpus. For the word pairs in the corpus, the modeling process of BTM is described as follows:

(1) For the corpus, there is a top distribution $\theta \sim Dir(\alpha)$.
(2) For each topic z, the word distribution under that topic is $\varphi_z \sim Dir(\beta)$.
(3) For each word in the set of B: $b = (w_i, w_j)$:
   a. Randomly select a topic z from the distribution of topics θ in the corpus, and then get z~*Multi*(θ).
   b. Randomly select two different words $w_i$andw_j that make up the word pair b from the extracted topic z, and there are w_i, w_j ~*Multi* (φ_Z).

2.3 Affinity Propagation Clustering

Affinity Propagation (AP) clustering is a new clustering algorithm first proposed in the article "Clustering by Passing Messages Between Data Points" in 2007 [15]. AP takes the similarity matrix of data points as the input. In the initial stage of clustering, AP does not require setting the clustering numbers and centers. The algorithm considers all data as the centers of potential clustering, and then continuously searches for suitable data points through iterative calculations, automatically identifies the clustering centers between data points and determines the clustering numbers. Data points find potential

Word2Vec model in 2013 for calculating word vectors. Word2Vec model utilizes the contextual information of words to transform a word into a low dimensional real vector space, and the more similar words are, the closer they are in the vector space. The application of word vectors in natural language processing has been very successful and has been widely applied in fields such as Chinese word segmentation [4, 5], sentiment classification [6], and syntactic analysis [7–9]. Word2Vec model reflects the contextual associations, which only focuses on a certain range of neighboring vocabulary relationships, which can easily lead to the loss of global information. BiTerm Topic Model (BTM) [10] discovers the topic distribution characteristics of a text through the co-occurrence of vocabulary information and the probability distribution between documents, topics, and vocabulary, thereby discovering the global semantic information and feature expression of the text. This paper proposes a hybrid vector representation model(HVRM) that integrates weight and topic features to address the problems of current single text feature representation models. HVRM is based on local and global features of short text to vectorize the short text. Then, an incremental AP clustering(IAPC) algorithm is proposed to cluster short text. Finally, the effectiveness of the proposed model in short text topic discovery is verified by comparing it with traditional methods.

2 Preliminaries

2.1 Vectorization of Words

Vectorization is to map words to a vector space and expresses a word mathematically. There are usually two ways to vectorize words.

(1) One-hot Encoding

One-hot encoding is a vectorized representation of words with a vector length equal to the size of the dictionary [11]. The components of a vector are composed of 0 and 1, where the position of component 1 is the corresponding position of the word in the dictionary, and the rest of the bits are all 0. Although one-hot encoding can clearly represent a word, it cannot express its semantics. Furthermore, when the dictionary size increases, its dimensional features are sparse, and each sparse column has a linear relationship, which is prone to collinearity problems.

(2) Word2Vec

Hinton first proposed mapping a word into a low dimensional, dense real vector space. If the meaning of two words is closer, the distance between them in the vector space is closer. Obviously, using this representation can better distinguish the similarity between words. Mikolov proposes Word2Vec by drawing inspiration from Bengio's Neural Network Language Model [12] and Hinton's Log_Linear model [13]. Word2Vec expresses words in vector through optimized training models based on a fixed corpus. There are two models for Word2Vec, namely CBOW and Skip-gram model. CBOW model uses k words before and after the current predicted word to predict the current word, while Skip-gram model uses the current word to predict its k words before and after.



Short Text Data Mining Based on Incremental AP Clustering

Fuyu Lu, Ying Guo, Peiyi Qu, and Yonglin Leng(✉)

College of Information Science and Technology, Bohai University, Jinzhou 121000, China
lengyonglin@qq.com

Abstract. The rapid development of mobile internet technology generates many short text data, which contains many hot topics. By clustering short text data, we can identify many hot topics in time. This information is crucial for discovering public opinion and analyzing user emotions. This paper proposes a hybrid vector representation model (HVRM) that combines weight and topic features to address the feature information loss caused by a single short text vector representation model and short text sparsity. Firstly, HVRM mines the local features using Word2Vec and TF-IDF to get the weighted vector of short text. Next, use BTM to obtain global feature vectors. And then connect the two feature vectors to form short text vectors. Finally, we use KNN to initialize the responsibility and availability matrices of incremental AP clustering (IAPC). The experimental results show that the hybrid vector representation model proposed in this paper can effectively improve the clustering effect.

Keywords: Short Text · Vector Representation Model · Incremental AP Clustering

1 Introduction

The rapid development of mobile internet technology allows people to express themselves online at any time. Weibo, WeChat and others have become the main ways for people to communicate. These platforms generate a large amount of text data, especially short text. These data contain lots of available information. Mining these data are beneficial for discovering hot topics, emotional tendencies that users are concerned about, and strengthening public opinion monitoring [1]. How to extract valuable information from short text is a hot research topic for scholars. Clustering technology is an unsupervised data classification method. It classifies data with similar features into one class by analyzing the similarity between data. Because the features of text data are not obvious, it is difficult for traditional processing methods to calculate the similarity between data accurately [2]. An effective method is to use deep learning to map short text to a low dimensional vector space and use vectors to describe text features. The self-feature extraction based on neural network is increasingly receiving attention from the industrial and academic communities. Based on previous research, Mikolov et al. [3] proposed the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 363–373, 2024.
https://doi.org/10.1007/978-3-031-65126-7_31

9. Li, X., Gao, R.X., Wong, B.C., Zhao, Y.: Deep learning-based fault diagnosis of rotating machinery using time-domain vibration signals. Mech. Syst. Signal Process. **155**, 107618 (2022)
10. Zhang, S., Wang, C., Wu, H.: Bearing fault diagnosis based on hybrid features of variational mode decomposition and deep belief network. IEEE Trans. Instrum. Meas. **70**, 1–11 (2021)

Table 3. (*continued*)

test data	Target output	Actual output	Accuracy
		0 0 0 0 1	
Total accuracy			90%

speed, especially when the input data of the neural network is large, the advantages of neural networks with principal component analysis will be very obvious.

6 Conclusions

This article introduces a cascaded inverter fault detection method that integrates digital signal processing technology and neural networks. This method determines the location and type of faults by monitoring and analyzing the output voltage of the inverter. Fast Fourier transform (FFT) is used to analyze the frequency spectrum of output voltage signal and extract fault features. A fault diagnosis system based on feed forward artificial neural network has been established, using the amplitude values of each harmonic wave of the fault voltage as the training input data of the neural network, and using the fault type as the output of the neural network. Simplify input data through principal component analysis (PCA), improve neural network structure, and implement inverter fault data feature recognition technology based on neural network to improve training speed and generalization accuracy, as well as timeliness and effectiveness.

References

1. Zhang, Z., Yan, R., Yang, Y.: Fault diagnosis of inverter based on deep learning. J. Eng. **2022**(18), 4272–4276 (2022)
2. Li, M., Wei, X., Ding, S.X., Han, Z.: Deep neural networks for fault diagnosis using time-frequency image-based features of rolling element bearings. Mech. Syst. Signal Process. **116**, 145–162 (2019)
3. Liu, F., Chen, J., Fang, L., Jiang, N.: A novel method for inverter fault diagnosis based on random forest optimized by particle swarm optimization. IEEE Access **8**, 84388–84398 (2020)
4. Ma, W., Wang, Y., Cai, G., Luo, X.: Motor bearing fault feature extraction based on convolutional neural network. IEEE Access **8**, 128717–128725 (2020)
5. Zhao, H., Jiang, D., Yu, J.: Robust fault diagnosis of wind turbine gearbox based on stacked sparse autoencoders. Renew. Energy **152**, 62–73 (2020)
6. Zhang, T., Chen, C., Mu, N., Chai, T.: Fault feature extraction of rolling bearings using time–frequency scattering convolutional network. J. Sound Vib. **484**, 115516 (2020)
7. Li, X., Huang, C., Yang, J.: Intelligent fault diagnosis for rolling bearings using a deep cascaded network. Measurement **173**, 108644 (2021)
8. Tao, M., Hu, W., Tian, Z.: A novel fault diagnosis method for rotating machinery based on deep convolutional neural networks and variational mode decomposition. Neurocomputing **429**, 14–25 (2021)

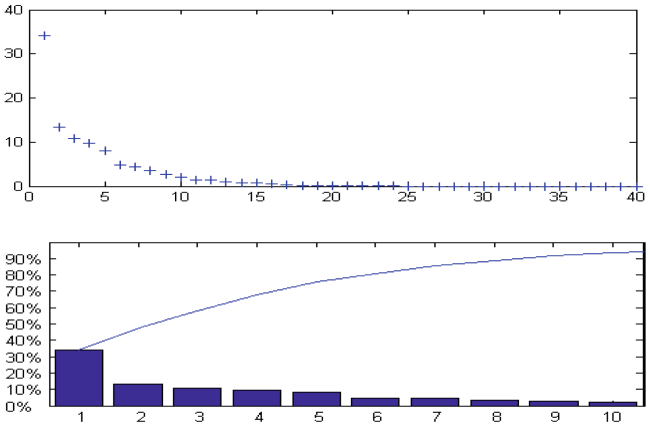


Fig. 9. Contribution rate and cumulative contribution rate of each principal component

Table 3. Generalization results of PCA neural networks.

test data	Target output	Actual output	Accuracy
M = 0.6, 0.7, 0.8, 0.9	1 0 0 0 0	1 0 0 0 0	100%
		1 0 0 0 0	
		1 0 0 0 0	
		1 0 0 0 0	
	0 1 0 0 0	0 1 0 0 0	100%
		0 1 0 0 0	
		0 1 0 0 0	
		0 1 0 0 0	
	0 0 1 0 0	0 1 0 0 0	100%
		0 0 1 0 0	
		0 1 0 0 0	
		0 0 1 0 0	
	0 0 0 1 0	0 0 0 1 0	100%
		0 0 0 1 0	
		0 0 0 1 0	
		0 0 0 1 0	
	0 0 0 0 1	0 0 0 0 1	100%
		0 0 0 0 1	
		0 0 0 0 1	

(*continued*)

After conducting principal component analysis on the original sample data in Fig. 5 and Fig. 6, the variance and cumulative contribution rate of each principal component can be obtained as shown in Fig. 9.

Table 2. Generalization results of neural networks.

test data	Target output	Actual output	Accuracy
M = 0.6, 0.7, 0.8, 0.9	1 0 0 0 0	1 0 0 0 0	100%
		1 0 0 0 0	
		1 0 0 0 0	
		1 0 0 0 0	
	0 1 0 0 0	0 1 0 0 0	100%
		0 1 0 0 0	
		0 1 0 0 0	
		0 1 0 0 0	
	0 0 1 0 0	0 1 0 0 0	50%
		0 0 1 0 0	
		0 1 0 0 0	
		0 0 1 0 0	
	0 0 0 1 0	0 0 0 1 0	100%
		0 0 0 1 0	
		0 0 0 1 0	
		0 0 0 1 0	
	0 0 0 0 1	0 0 0 0 1	100%
		0 0 0 0 1	
		0 0 0 0 1	
		0 0 0 0 1	

It can be seen from Fig. 9 that the cumulative contribution rate of the first seven principal components has reached 85%, including the main information of the original data. In this way, the first seven principal components can be used to replace the original data for neural network training. Reconstruct the fault diagnosis system of the artificial neural network shown in Fig. 8, simplifying the input from the original 30 to Fig. 7. The generalization test of the reconstructed fault discrimination neural network using test data can obtain the results shown in Table 3.

Comparing the generalization results of the neural network shown in Table 2 with the generalization results of the neural network with principal component analysis shown in Table 3, it can be concluded that both proposed neural network structures have good generalization accuracy, and both have an overall accuracy of 90%; However, neural networks with principal component analysis have a simple structure and fast training

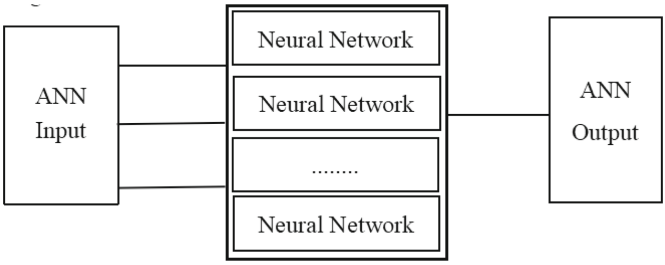


Fig. 8. Fault diagnosis system based on artificial neural network

The first 30 frequency component amplitudes of the fault voltage waveform corresponding to different modulation ratios are used as input data for the artificial neural network. The meanings of the 5 output data of the artificial neural network are shown in Table 1.

Table 1. Means of network output binary code

fault type	Network output				
	Network 1	Network 2	Network 3	Network 4	Network 5
System is normal	One	0	0	0	0
S_{11} open circuit	0	One	0	0	0
S_{12} open circuit	0	0	One	0	0
$\overline{S_{11}}$ open circuit	0	0	0	One	0
$\overline{S_{12}}$ open circuit	0s	0	0s	0s	One

The neural network is trained using the Levenberg Marquardt paradigm trainlm. Input the test data (corresponding harmonic amplitudes of the fault voltage at modulation ratios M = 0.65, 0.75, 0.85, and 0.95) into the trained neural network for prediction, and obtain the prediction results shown in Table 2. It can be seen from the table that the S_{12} accuracy of the fault discrimination network in judging whether the whole system operates normally, S_{11} faults, $\overline{S_{11}}$ faults and $\overline{S_{12}}$ faults is 100%, the accuracy in judging faults is 50%, and the overall accuracy of the system can reach 90%.

5 Artificial Neural Network Fault Diagnosis Using Principal Component Analysis

Although the data in Fig. 5 and Fig. 6 can be directly used to train neural networks for fault classification, these data still have high correlation and similarity. Therefore, this paper adopts a statistical method - Principal Component Analysis (PCA) to first simplify the input data, thereby improving the training speed and generalization accuracy of the neural network.

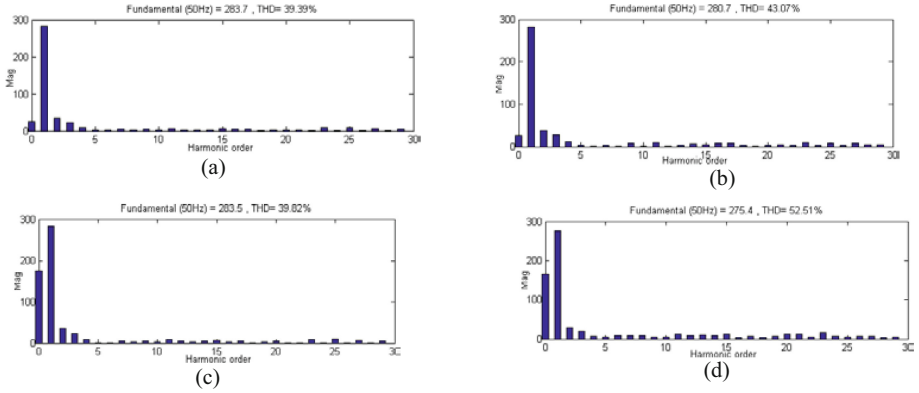


Fig. 5. The FFT of the output voltage signal under single circuit disconnection (a) S_{11} (b) S_{12} (c) $\overline{S_{11}}$ (d) $\overline{S_{12}}$

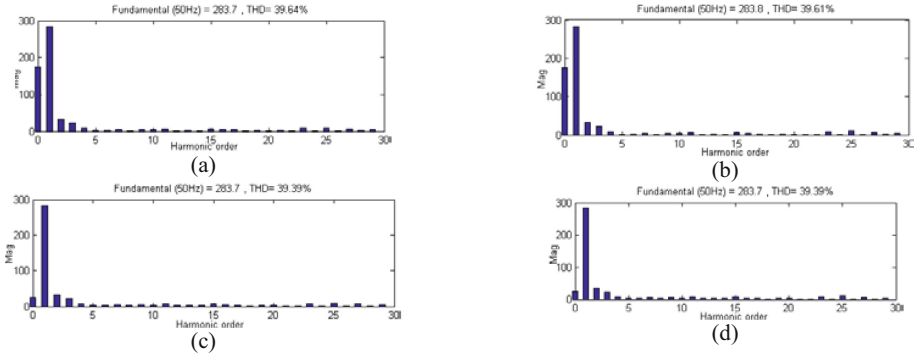


Fig. 6. The FFT of the output voltage signal under the condition of single tube short circuit (a) S_{11} (b) S_{12} (c) $\overline{S_{11}}$ (d) $\overline{S_{12}}$

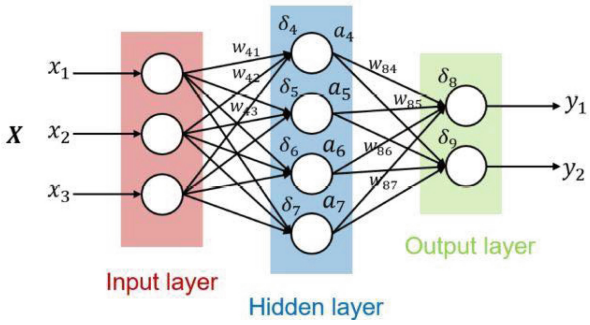


Fig. 7. The structure of artificial neural network

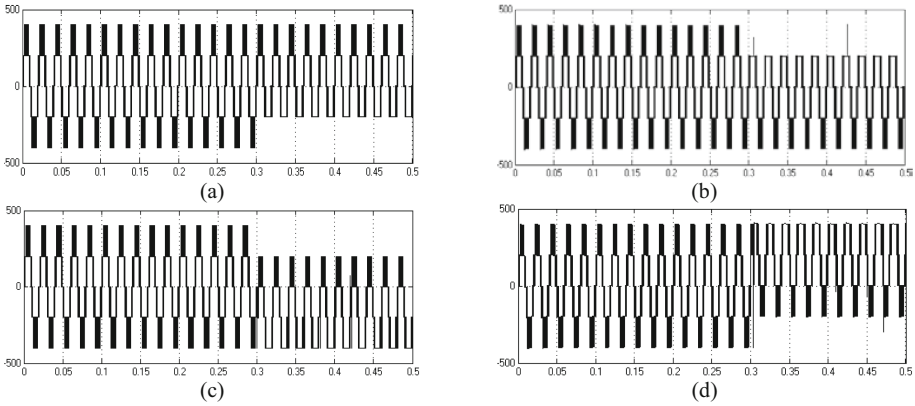


Fig. 3. Output voltage of different switching off circuits in A phase H Bridge1 (a) S_{12} S_{11} (b) $\overline{S_{11}}$ (c) (d) $\overline{S_{12}}$

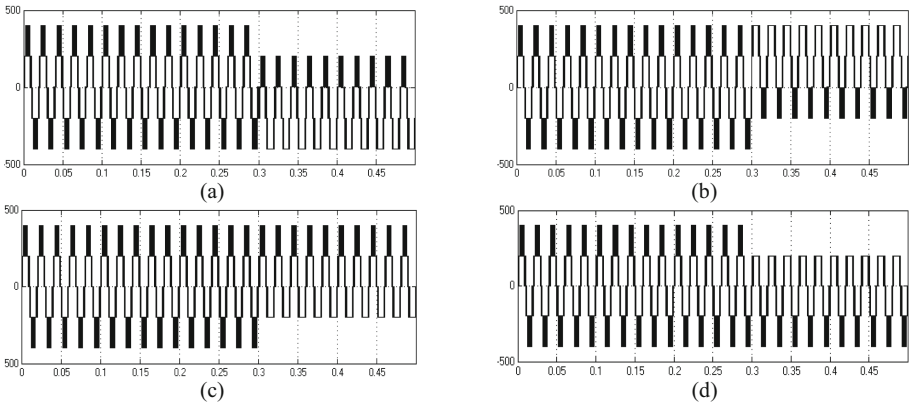
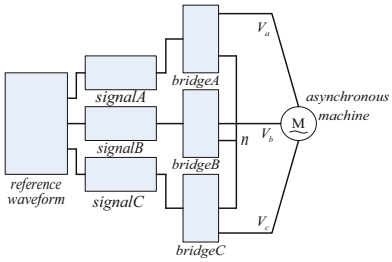


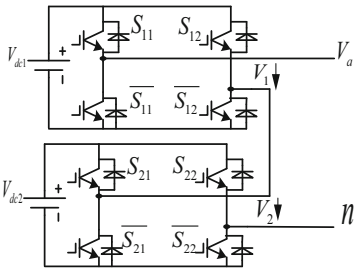
Fig. 4. Output voltage of short circuit with different switching tubes in A phase H Bridge1 (a) S_{11} (b) S_{12} (c) $\overline{S_{11}}$ (d) $\overline{S_{12}}$

4 Fault State Recognition Based on Feedforward Artificial Neural Network

In order to achieve fault diagnosis of cascaded inverters by applying the spectra shown in Fig. 5 and Fig. 6, this paper constructs a BP feedforward artificial neural network as shown in Fig. 7, which includes a hidden layer, 30 input nodes corresponding to the harmonic order, and an output node. The Activation function used by the neural network is sigmod function [9, 10]: tansig is used for the hidden layer, and log sig is used for the output layer. The output of the neural network is 0 or 1. The proposed fault diagnosis system is shown in Fig. 8, which consists of five neural networks with the same structure. The output of the system is a 5-bit binary code, which corresponds to different faults [5, 7].



(a) Block diagram of induction motor drive system with inverter



(b) A phase schematic of inverter

Fig. 1. Induction motor system drive by five levels three phases cascaded H bridge inverter

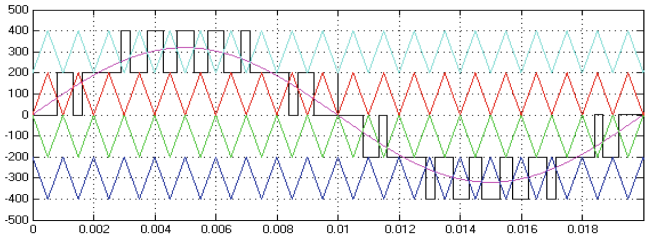


Fig. 2. The distribution of carrier wave and modulation wave and its output voltage when M = 0.8

Due to the three-phase symmetry of the inverter, the following analysis will take phase A as an example. To accurately diagnose faults, it is necessary to study the output characteristics of inverters under different faults. When any switch tube of the actual higher-level connected inverter malfunctions, it will affect the output voltage of the inverter. Figure 3 and Fig. 4 respectively show the corresponding output voltage waveforms for different switch tube open circuit and short circuit faults in phase A H Bridge1 [6].

By analyzing the output voltage in Fig. 3 and Fig. 4, it can be seen from the time domain that different fault states correspond to different output voltage waveforms.

3 FFT Analysis of 3 Fault Voltage Waveforms

In order to enable computers to accurately identify fault states by outputting voltage waveforms, this paper studies the application of artificial neural networks to analyze, classify, and distinguish fault waveforms. In the training of neural networks, it is necessary to provide characteristic training data [7, 8]. However, the output voltages in Fig. 3 and Fig. 4 are not suitable as training data directly, as there is a high similarity between the output voltage data and the feature is not obvious [9]. Therefore, this paper first carried out Fourier analysis on the output voltage, as shown in Fig. 5 and Fig. 6, and extracted the spectrum amplitude representing the first 30 harmonic components of the fault waveform as the training data of the neural network.

Research on Feature Extraction and Recognition of Inverter Fault Data Based on Neural Networks

Jingpeng Hu[1,2(✉)] and Zhiguo Xiong[3]

[1] School of Computing, Beijing Institute of Technology, Zhuhai 519000, Zhuhai, China
02092@bitzh.edu.cn
[2] Software Engineering Technology Research Center, Zhuhai 519000, Zhuhai, China
[3] School of Aviation, Beijing Institute of Technology, Zhuhai 519000, Zhuhai, China

Abstract. This article proposes a fault detection method for cascaded inverters that combines digital signal processing technology and neural networks. This method determines the location and type of faults by detecting and analyzing the output voltage of the inverter. Fast Fourier transform (FFT) is used to analyze the frequency spectrum of output voltage signal to extract fault characteristics. By simplifying input data through Principal Component Analysis (PCA), the structure of neural networks can be improved. The feature recognition technology of inverter fault data based on neural networks can timely and effectively improve training speed and generalization accuracy.

Keywords: Neural network · Inverter failure · Data feature recognition

1 Introduction

High power cascaded multi-level converters without power frequency transformers belong to complex power electronic systems [1]. High reliability and ease of maintenance are the key factors for the practical application of such a complex power electronic system in industry. One of the effective methods to improve system reliability is to use fault-tolerant technology and redundancy technology [2]. However, the core of such technology is the timely and accurate detection of faults, Timely and accurate detection and diagnosis of faults [3].

2 Inverter Output Voltage Waveform and Fault Status

The above method is explained in detail through an example of an open circuit fault in a three-phase cascaded five level inverter modulated by the POD (Phase Opposition Disposition) strategy [4]. Figure 1(a) shows the structural diagram of an asynchronous motor system powered by a three-phase cascaded inverter, and Fig. 1(b) shows the schematic diagram of each phase of the three-phase cascaded inverter. Figure 2 shows the carrier wave, modulation wave, and corresponding output voltage waveform of the inverter using POD modulation [5].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 354–362, 2024.
https://doi.org/10.1007/978-3-031-65126-7_30

20. Horowitz, D.: Essay examination prompts and the teaching of academic writing. Engl. Specif. Purp. **5**(2), 107–120 (1986)
21. Kim, A.Y.: Investigating second language reading components: Reading for different types of meaning (2009)
22. Lumley, T.: The notion of subskills in reading comprehension tests: an EAP example. Lang. Test. **10**(3), 211–234 (1993)
23. Munby, J.: A problem-solving approach to the development of reading comprehension skills. In: Presentation at the University Teachers of English in Israel (UTELI) Regional Meeting, Jerusalem, January, vol. 25 (1978)
24. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet Rasch measurement: Part I. J. Appl. Meas. **4**(4), 386–422 (2003)
25. O'hare, F.: Sentence Combining: Improving Student Writing without Formal Grammar Instruction. NCTE Committee on Research Report Series, No. 15 (1973)
26. Plakans, L., Gebril, A.: A close investigation into source use in integrated second language writing tasks. Assess. Writ. **17**(1), 18–34 (2012)
27. Plakans, L., Gebril, A.: Using multiple texts in an integrated writing assessment: source text use as a predictor of score. J. Second. Lang. Writ. **22**(3), 217–230 (2013)
28. Purpura, J.E.: The development and construct validation of an instrument designed to investigate the cognitive background characteristics of test-takers. In: Kunnan, A.J. (ed.), Validation in Language Assessment, pp. 111–139. Mahwah, NY: Lawrence Erlbaum Associates, Inc (1998)
29. Purpura, J.: Assessing grammar. Cambridge University Press, Cambridge, UK (2004)
30. Purpura, J.E.: A Rationale for Using a SBA to Measure Competency-Based, Situated S_FL Proficiency. Jan 3 final version manuscript (2021)
31. Purpura, J.E.: Class notes from Second Language Assessment (2022)
32. Raimes, A.: Out of the woods: emerging traditions in the teaching of writing. TESOL Q. **25**(3), 407–430 (1991)
33. Schoonen, R.: How language ability is assessed. Handb. Res. Second Lang. Teach. Learn. **2**, 701–716 (2011)
34. Schoonen, R.: Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. Read. Writ. **32**(3), 511–535 (2019)
35. Selinker, L., Todd-Trimble, M., Trimble, L.: Rhetorical function-shifts in EST discourse. TESOL Q. **12**, 311–320 (1978)
36. Selzer, J.: The composing processes of an engineer. Coll. Compos. Commun. **34**(2), 178–187 (1983)
37. Shanahan, L.E.: Reading and writing multimodal texts through information and communication technologies (2006)
38. Weigle, S.C.: Assessing Writing. Cambridge University Press, Cambridge (2002)
39. Weigle, S.C.: Integrating reading and writing in a competency test for non-native speakers of English. Assess. Writ. **9**(1), 27–55 (2004)
40. Wiersma, W., & Jurs, S. G.: Research Methods in Education: An Introduction (9th ed.) (2009)
41. White, R.V.: New Ways in Teaching Writing. New Ways in TESOL Series: Innovative Classroom Techniques (1995)
42. Zamel, V.: Re-evaluating sentence-combining practice. Tesol Q. 81–90 (1980)
43. Carr, N.: Designing and Analyzing Language Tests. Oxford University Press, Oxford, UK (2011)

emphasizes the significance of the learning process itself. However, it is essential to acknowledge the limitation of small sample sizes in some studies, which may impact the generalizability of findings. Future research should aim to address this limitation and incorporate larger and more diverse samples to enhance the validity and reliability of the assessment outcomes. Overall, the learning-oriented test model holds promise in promoting effective language learning and personalized educational practices, contributing to the advancement of language assessment in the educational context.

References

1. Abbott, R.D., Berninger, V.W., Fayol, M.: Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. J. Educ. Psychol. **102**(2), 281–298 (2010)
2. Alderson, C.: Assessing Reading. Cambridge University Press, Cambridge (2000)
3. Bachman, L.F., Palmer, A.S.: Language Testing in Practice. Oxford University Press, Oxford (1996)
4. Bazerman, C.: A relationship between reading and writing: the conversational model. Coll. Engl. **41**(6), 656–661 (1980)
5. Bennett, R.E., Deane, P., van Rijn, W., P.: From cognitive-domain theory to assessment practice. Educ. Psychol. **51**(1), 82–107 (2016)
6. Berninger, V.W., Abbott, R.D., Abbott, S.P., Graham, S., Richards, T.: Writing and reading: connections between language by hand and language by eye. J. Learn. Disabil. **35**(1), 39–56 (2002)
7. Borsboom, D., Mellenbergh, G.J., Van Heerden, J.: The concept of validity. Psychol. Rev. **111**(4), 1061 (2004)
8. Bridgeman, B., Carlson, S.: Survey of academic writing tasks required of graduate and undergraduate foreign students. ETS Res. Rep. Ser. **1983**(1), i–38 (1983)
9. Camp, R.: The place of portfolios in our changing views of writing assessment. In Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment, pp. 183–212 (1993)
10. Cao, Y., Chen, J., Zhang, M., Li, C.: Examining the writing processes in scenario-based assessment using regression trees. ETS Res. Rep. Ser. **2020**(1), 1–16 (2020)
11. Carson, J.E., Carrell, P.L., Silberstein, S., Kroll, B., Kuehn, P.A.: Reading-writing relationships in first and second language. TESOL Q. **24**(2), 245–266 (1990)
12. Cumming, A., Grant, L., Mulcahy-Ernt, P., Powers, D.E.: A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. Lang. Test. **21**(2), 107–145 (2004)
13. Davis, F.B.: Research in comprehension in reading. Read. Res. Q. **3**, 499–545 (1968)
14. Emig, J.: The composing processes of twelfth graders (1971)
15. Gebril, A., Plakans, L.: Investigating source use, discourse features, and process in integrated writing tests. In: Spaan Fellow Working Papers in Second or Foreign Language Assessment, , vol.7, no. 1, pp. 47–84 (2009)
16. Grabe, W.: Current developments in second language reading research. TESOL Q. **25**(3), 375–406 (1991)
17. Grabe, W., Zhang, C.: Reading-writing relationships in first and second language academic literacy development. Lang. Teach. **49**(3), 339–355 (2016)
18. Hamid, M.O., Hardy, I., Reyes, V.: Test-takers' perspectives on a global test of English: questions of fairness, justice and validity. Lang Test Asia **9**, 16 (2019)
19. Hayes, J. R.: Understanding Cognition and Affect in Writing. Perspectives on writing: Research, theory, and practice, vol. 6 (2000)

Table 18. Pearson Correlation Matrix for writing average: Observed Variables (N = 8)

	Language Cont Ave	Contt Accu Ave	Rhet Con Ave	Cont Elab Ave	
Person Correlation	0.61	.934**	.895**		**Cont Elab Ave**
Sig. (2-tailed)	0.108	<.001	0.003		
Person Correlation	0.667	.882*			**Rhet Con Ave**
Sig. (2-tailed)	0.071	0.012			
Person Correlation	0.348				**Contt Accu Ave**
Sig. (2-tailed)	0.398				
Person Correlation					**Language Cont Ave**

*. *p* < .05, 2-tailed
**. *p* < .01, 2-tailed

As the results in Table 18 show, Content Elaboration and Content Accuracy, among all the relationships between variables, showed a very high positive correlation at .934. It is also encouraging to see that the correlation was found to be statistically significant at the .01 level, which means there is a 99% chance that this correlation is not due to chance. Similarly, we observed that Content Accuracy and Rhetoric Control had a high correlation at .822, and Content Elaboration and Rhetoric Control had a high correlation at .895. Both correlations were found to be statistically significant, at the level of .05 and .01, respectively. It was also observed that Language Control displayed a positive correlational relationship with Content Accuracy, Rhetoric Control, and Content Elaboration, at .348, .667, and .610, respectively. However, none of the three correlations were statistically significant, which means the correlations could have occurred due to chance. Nevertheless, it is encouraging to see all the correlations are positive, and three of them were found to be statistically significant, suggesting that some of the variables were measuring the same underlying construct to a high degree, and some to a certain degree. This supports our previous argument based on the literature review that writing ability in this test consists of four components: language control, content accuracy, content elaboration, and rhetoric control.

4 Conclusion

In conclusion, the learning-oriented test model offers a valuable approach to assessing students' language proficiency, focusing on their learning progress and individual growth. Adopting a dynamic and student-centered perspective, this assessment method

Table 16. Inter-Rater Reliability Correlation Matrix: Observed Variables (N = 8)

Language control R1, R2	.577
Content accuracy R1, R2	.600
Rhetorical control R1, R2	.667
Content elaboration R1, R2	.775*

*. *p < .05, 2-tailed*

Table 17. Inter-Rater Reliability Correlation Matrix: Writing Average Scores (N = 8)

	Writing Average Rater 2
Writing Average Rater 1	.883**

**. *p < .01, 2-tailed*

The correlation coefficients between the four individual variables ranged from moderately positive (e.g., 0.577) to highly positive (e.g., 0.775). The inter-rater reliability was .577 for language control, .600 for content accuracy, and .667 for rhetoric control, none of which, according to SPSS, was found to be statistically significant at the .05 level, meaning that there is not a significant linear correlation between rater 1 and rater 2 in the sample. The correlation coefficient for content elaboration, however, was .775 and was found to be statistically significant at the .05 level, meaning that there is a 95% chance the correlation is not a chance phenomenon. Turning to the measurements of Inter-rater reliability computed using the composite averages by rater, Table 18 below provides the Pearson correlation yielded from this analysis.

The correlation for the Writing Average Rater 1 and Writing Average Rater 2 was .883 at the .01 level, meaning there is a 99% chance that the correlation is not a chance phenomenon. Considering that the correlation coefficient range is −1 to 1, we consider the $r = .833$ to be a high correlation coefficient, meaning that the two raters were highly congruent in their overall rating for writing performance of students.

The high inter-rater reliability on writing average score of the R1 and R2 might be because the two raters worked together, or 'norming', through multiple rounds of discussions and revisions on the test design, scoring rubrics, and scoring descriptors to reach an agreement on what should be tested and to reconcile their definitions and standards for the four components under the writing construct. However, the two raters are still different in years of teaching and learning experience, personality, grading leniency, and this was their first time working together as a team, all of which can be the possible reasons why not all individual variables have high and statistically significant correlations. In this section, we examined the level of construct validity for the writing task. There were four components, or variables, for the writing tasks: Language Control, Content Accuracy, Rhetoric Control, and Content Elaboration. The participants' average scores on each of the observed variables were imported into SPSS for correlation analysis. We chose the Person Product-Moment formula because those scores were composite in nature.

responses of students. Two big reasons that can potentially explain the high reliability are a). The two raters both have spent a considerable amount of time studying, living, or working in the United States with frequent exposure to English as a second language so they might have a very similar overall judgment of the writing of a student; *and* b). The two raters had numerous discussions and rounds of revisions of the scoring rubrics and descriptors to make sure they had reached an agreement on how they would assess the writing response of a test-taker.

The standard error of measurement (*SEM*) was also calculated in this section. The formula is 1 minus r_{xx} (Reliability Estimate) and then calculating the square root of the result. Then the square root is multiplied by S (the Standard Deviation). Below is the formula.

$$SEM = \mathbf{SEM} = S\sqrt{1 - r'_{xx}} \tag{4}$$

$$S = \mathbf{standard\ deviation}(S = .504) \tag{5}$$

$$r'_{xx} = \mathbf{reliability}(r'_{xx} = .888) \tag{6}$$

Based on the descriptive statistics in previous sections, we can know that the standard deviation is .504 and the calculated SEM is 0.169. Considering this is a low-stakes test, we use a 68% confidence interval ($\pm 1SEM$), which means if one of the test takers were to take the test again, we could say with 68% confidence that his or her score is most likely to fall within $\pm 1SEM$ of their current score on the writing task. The cut-score for passing is 70%, and it means we can be 68% confident that test-takers who received below 3.331 did not receive the passing grade for the task. $+1SEM$ is 3.669 or more and this means that we can be 68% confident that those who scored above 3.669 received a passing grade for the task (i.e., 3.5 out of 5 = 70%). For those test-takers who scored between -1 SEM (3.331) and $+1SEM$ (3.669), we cannot say with 68% confidence that they either passed or didn't pass. A 68% confidence level ($\pm 1SEM$; ± 0.169) with 70% as the cut-score would mean a student needs to get at least 3.33 to pass the exam (Tables 15, 16 and 17).

Table 15. Writing scores for all the test-takers

ID	1	2	3	4	5	6	7	8
SCORE	4.375	3.635	4.625	4.000	4.875	3.75	4.75	4.875

Based on the 68% confidence interval, eight test-takers (IDs 1, 2, 3, 4, 5, 6, 7, and 8) received passing grades with scores above 3.669 out of 5. However, one test-taker (ID2) received a score of 3.635, falling within the interval (3.331 to 3.669), making it uncertain to confidently conclude a passing grade within the 68% confidence level. To assess inter-rater reliability, correlations were calculated between two raters' scores for four components (language control, content accuracy, rhetoric control, and content elaboration) using Spearman rank-order and Pearson product-moment.

Table 13. Stem and Leaf Plot of results for Writing task

Score	3.75	4.00	4.25	4.50	4.75	5.00
Frequency	XX	X	X		XX	XX

Stem Width: .25.
Each Leaf: 1 case(s).

Table 14. Reliability of the writing task (N = 8)

Cronbach's Alpha (α)	Number of components of rubric
.888	4

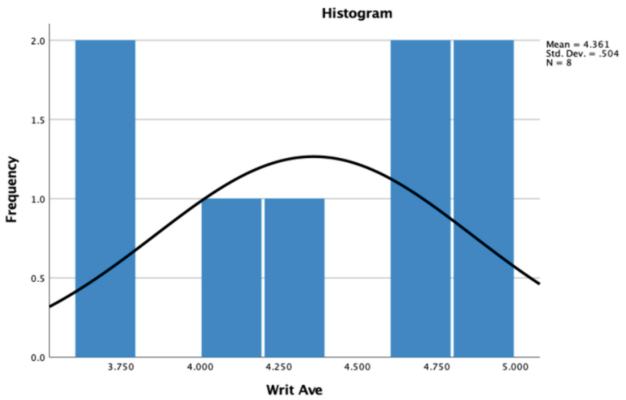


Fig. 1. Histogram of the results of the writing task with a normal distribution

Control, Content Accuracy Control, Rhetoric Control, and Content Elaboration Control), and the calculations were conducted using interval variables. Table 15 shows the alpha was estimated to be .888.

A .888 reliability can be interpreted as that only 11.2% of the observed variance can be attributed to measurement errors or other unaccounted factors but 88.8% of the score variance is attributed to true score variance. Internal consistency reliability for high-stakes standardized tests like TOEFL usually needs to be extremely high. TOEFL claims to have an overall reliability estimate of 0.95 and a 4.26 SEM (standard errors of measurements) (ETS, 2018), which can be seen as high consistency and reliability since the smaller the value of SEM, the higher the measurement quality, and the more precise the test scores would be. For low-stakes classroom-based assessments like our test, .70 is considered an acceptable threshold (Carr, 2011), and thus we can consider .888 to be a very high reliability in this case. In CTT, an alpha of .888 is interpreted as 88.8% of variance attributable to true test-taker ability, and 11.2% of variance attributable to error (Carr, 2011, pp. 108–109). It means the writing section can reliably assess writing

support the claimed validity of the test. Also, the sample size of the study (N = 8) was extremely small, and the total reading score range of students was very small (9 was the lowest and 12 was the highest), meaning the data was negatively skewed. On the bright side, if all the reading scores of test-takers were negatively skewed, it means most of them did very well in the test. This could be likely attributed to the fact that some of the questions reflected the content taught in class. Good and bad email examples of writing were analyzed and discussed in the class, so maybe by the time the test was administered, the students already had a good understanding of the dos and don'ts of writing emails to professors.

3.2 Results for the Writing Task

Table 12 showed writing performance of test-takers in language control, content accuracy, content elaboration, and rhetorical control, four of which were scored from 1 to 5, respectively. Within a total possible score of 5, the mean was 4.36, the median was 4.5), and the mode was 4.875. The standard deviation was .504. The skewness was −.456, with a standard error of skewness of .752. We can also read that the kurtosis was −1.714 and the standard error of kurtosis was 1.481. Since the minimum for the writing was 3.635 and the maximum was 4.875, the range was 1.24.

Table 12. Descriptive Statistics for the Writing

	Central Tendency						Dispersion		Frequency	Distribution
Write Ave	N	Mean	Mode	Median	Min	Max	Range	SD	Skewness	Kurtosis
	8	4.36	4.875	4.5	3.635	4.875	1.24	0.504	−.456	−1.714

The negative skewness of −.456 indicates that higher scores occurred fairly more frequently than lower scores. A negative skewness is desirable in this case since the test was designed to be an achievement test so we wanted to see more high scores. It suggests that most of the test-takers did a fairly good job in successfully completing the email writing task. Since the two parts of the test were designed in a learning-oriented way where test-takers needed to incorporate the reading input into writing tasks, this negative skewness could also be interpreted that a big proportion of the test-takers well comprehended the reading materials and successfully incorporated the understanding into their writing. The mean score of 4.361 and the median score of 4.50 (out of 5) also suggested that the test-takers did a good job in writing overall. The standard deviation of 0.504 and kurtosis of −1.714 suggest a standard distribution of scores. To put it in a different way, scores of test-takers showed some moderate variation in writing (refer to Table 13 and 14). (Fig. 1)

Internal Consistency Reliability for the Writing Task was calculated using Cronbach's Alpha and was done in a similar way when Internal Consistency was calculated for the Reading MC items. The difference is unlike the Reading MC scores which were dichotomous, Writing Task's Internal Consistency was estimated based on the average composite variables (between rater 1 and 2) for each of the four components (Language

As shown in Table 9, all the test-takers (all the high- and low-performing group and the people in the middle) chose the right answer D, making the Item Facility 1.00, which indicates this question might be too easy. But considering the nature of the test is a low-stakes classroom-based language learning achievement test, we decided to keep the item. The D-value of the item is 0, which means it cannot discriminate between low performing test-takers from high performing ones. No test-takers chose any of the other three distractors, indicating that the distractors may need to be revised had we decided to revise this item. We propose to revise the items in the following way had we decided to: A, and B, and C need to be more distracting so they can properly discriminate between the high and low performing groups. Again, as previously stated, the purpose of revision is to make sure the item is actually assessing what it claims it measures, which is the endophoric implied reading ability of the student (Table 10).

Table 10. Stem and Leaf Plot of results for Reading task

Score	9.00	10.00	11.00	12.00
Frequency	XXX	XXX	X	X

Stem Width: 1.00.
Each Leaf: 1 case(s).

The reading part of the test consists of two variables: endophoric literal and endophoric implied. The 12 multiple-choice items were evenly split, with six for each variable. We used the Pearson Product-Moment formula to analyze the scores and assess the relationship between the two item types. Positive correlation between them is expected, as they measure the same reading ability. A correlation coefficient above .75 is high, .5–.74 is moderate, .25–.49 is low, and 0 means no correlation.

Table 11. Pearson Correlation Matrix for reading total: Observed Variables (N = 8)

Endo Lit Total		
Endo Implied Total	Person Correlation	−.114
	Sig. (2-tailed)	.788

The result from Table 11 showed a slightly negative correlation between total endophoric literal and total endophoric implied scores for each student. This means that the better a test taker did in the endophoric literal questions, the poorer he or she did in the endophoric implied questions, and vice versa. But this correlation was not found to be statistically significant, which means this could be a chance phenomenon. Based on the data, we may conclude that there is not sufficient evidence to suggest the two types of questions were measuring the same ability as they were supposed to. A few reasons could explain this. One reason could be that the reading test per se had low reliability, and the items need to be improved to serve as a sufficient condition to

a classroom-based language teaching context. D-values range from −.314 to .051, suggesting limited discrimination ability, likely influenced by the small sample size (N = 9) and narrow range of total reading scores (9–11). Two items (endoliteral 1 and endoimplied 2) with negative D-values will be revised, while four items with low D-values will be kept, considering the small sample size and low-stakes nature of the test. The negative calculated alpha indicates issues with the reliability model assumptions, so decisions are not solely based on alpha. Distractor Analysis was conducted to examine why some items failed to differentiate between high and low-performing test-takers. Representative items (endoliteral 1 and endoimplied 9) were selected for further investigation. Endoliteral 1, with a negative D-value, will be removed to improve Cronbach's Alpha, and endoimplied 9, with an IF of 1 and D-value of 0, was found to be non-discriminating (Table 8).

Table 8. Distractor Analysis for endoliteral 1 (total test-takers N = 6, high and low-performing group K = 6)

Answer Choice	Frequency	High (n = 3)	Low (n = 3)	Item Facility (Upper)	Item Facility (Lower)	Item Discrimination
A, B	0	0	0	0	0	0
C	2	1	0	.333	0	.111
D*	4	2	3	.667	1	−.111

*. *Key*

The IF of .63 shows 63% answered correctly, but the Dd-value of −.228 reveals an issue, favoring low-performers. Deleting the item increases Cronbach's Alpha to .079 (−.170 originally due to small sample size). Table 13 indicates negative discrimination for Choice D (key) with only two high-performers selecting it, while all three low-performers did. Choices A and B were poor distractors with no selections. In conclusion, all choices need revision. A and B should be more distracting, and C and D must better discriminate between high and low performers, measuring endophoric implied understanding accurately.

Table 9. Distractor Analysis for endoimplied 9 (total test taker N = 6, high and low-performing group K = 6)

Answer Choice	Frequency	High (n = 3)	Low (n = 3)	Item Facility (Upper)	Item Facility (Lower)	Item Discrimination
A, B, C	0	0	0	0	0	0
D*	6	3	3	1	1	0

*. *Key*

for low-stakes classroom assessments. Due to a small sample size, normal distribution of scores might be compromised. As alpha cannot be below 0, we consider it as .00. Another measure of internal reliability, the standard error of measurement (*SEM*), accounts for sample variability and measurement error. It is calculated using the square root of "1" minus Cronbach's alpha, multiplied by the standard deviation (*S*) for the task.

$$SEM = \mathbf{SEM} = S\sqrt{1 - r'_{xx}} \tag{1}$$

$$S = \mathbf{standard\ deviation}(S = 1.069) \tag{2}$$

$$r'_{xx} = \mathbf{reliability}(r'_{xx} = 0.00) \tag{3}$$

The SEM value based on the standard deviation (*S*) of 1.069 is also 1.069. Using a 68% confidence interval, the passing cut-score for the reading part is 8.4 out of 12. Scores above 9.469 ($1SEM$) can be 68% confident of passing, while scores below 7.331 ($-1SEM$) can be 68% confident of not passing. Due to the low alpha, we cannot confidently determine pass or fail for scores between $-1SEM$ and $+1SEM$ (Carr, 2011). For the internal consistency analysis of the 12 multiple-choice reading questions, we examined Item Facility (IF), Discrimination Index (DI), and Alpha if Deleted. The ideal IF range is .3-.7. Positively discriminating items (DI $\geq$ .4) are usually desirable, while negatively discriminating items should be revised or deleted (Carr, 2011) (Table 7).

Table 7. Item analysis of the Multiple-Choice Questions

Item	Item Facility	Discrimination	Cronbach's alpha if removed	Decision
Endoliteral 1	.63	−.228	.079	Revise
Endoimplied 2	.38	.038	−.319[a]	Keep
Endoliteral 3, 7	1	0	−.319[a]	Keep
Endoimplied 4, 5, 6				
Endoimplied 8	.88	−.314	.07	Revise
Endoimplied 9	1	0	.07	Keep
Endoliterate 10	.63	.038	−.319[a]	Keep
Endoimplied 11	.88	.051	−.273[a]	Keep
Endoimplied 12	.63	.038	−.319[a]	Keep

Note. ([a])The value is negative due to a negative average covariance among items. This violates reliability model

The items in the test show a wide range of difficulty, with Item Facility (IF) ranging from .38 to 1.00. Six items were considered "no variance" as they were answered correctly by all test-takers (IF = 1), indicating they might be too easy. Despite this, we decided to keep all non-discriminating items due to the low-stakes nature of the test in

3 Data Analysis

3.1 Results for the Reading Task

As the results in Table 5 showed, the reading section had 12 multiple choice questions, for a full score of 12. The mean was 10.00, the median was 10.00, the mode was 9, and the standard deviation was 1.069. The skewness was .935. The standard error of skewness was .752, and the kurtosis was .350, with a standard error of kurtosis of 1.481.

Table 5. Descriptive Statistics for the Reading

	Central Tendency						Dispersion		Frequency	Distribution
Read Tot	N	Mean	Mode	Median	Min	Max	Range	SD	Skewness	Kurtosis
	8	10	9	10	9	12	3	1.069	.935	.35

The measures of distribution, kurtosis and skewness were included to measure comparability to a normal distribution, based on the assumption that a large enough group was being tested. However, it is important to remember that our sample size (N = 8) is very small so the distribution and characteristics were only situated within this study's context. The positive skewness of .935 for the reading total indicates that there were more lower scores than higher scores. If we only look at the skewness, we might conclude that having more lower scores is not ideal since the test was designed to be an achievement test, so we wanted more people to have a high score than otherwise.

However, a mean and median score of 10 out of 12 indicates that most of the test-takers did a fairly good job. It suggests although some test-takers performed better, the test-taker did a good job of reading overall. The standard deviation of 1.069 and kurtosis of 1.481 were both within the normal range, suggesting a moderate to fairly narrow distribution of scores. This means that most of the reading scores of the test-takers were similar. The internal consistency reliability and standard error of measurement (*SEM*) were calculated for all variables. Taking the composite variables for each component of the data, the internal consistency reliability was measured by Cronbach's alpha. Alpha is calculated by using the individual item-level data for all the items (Table 6).

Table 6. Reliability of the reading (N = 8)

Cronbach's alpha (α)	Number of MC Items
− 0.170	12

Note. The alpha(α) is negative due to a negative average covariance among items. This violates reliability model assumptions. The reliability will thus be considered .00

Cronbach's alpha for our test is −.170, indicating a negative average covariance among items and violating reliability assumptions. Ideally, alpha should be .70 or higher

Table 4. A sample score report for the test

Student M
Score: 22/24 for reading, 19/20 for writing
The email was well-written with a careful selection of politeness hedges (e.g., would you mind, would you please, I would like to…, etc.). The email exhibits a clear and appropriate explanation of the situation and suggested reasonable alternative solutions. The email also shows a good amount of consideration of the reading input. There are some minor grammatical and lexical errors, but they did not interfere with understanding. Please see your Google Doc. for detailed edits.
Please see below for the scoring rubrics.

Score	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
	Your score: 4	**Your score: 5**	**Your score: 5**	**Your score: 5**
5	The language of email is clear, cogent and smooth, and displays high accuracy in grammatical and lexical choices from a wide range. Consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some very minor lexical or grammatical errors which do not interfere with understanding.	The email information is accurate and relevant to the task. The form of the email is complete and shows an accurate understanding of the reading input.	The email includes clear and logical explanations of the student's situation. The relationships between ideas are supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are persuasive.	The email is fully developed with considerations and examples from the reading input. The email may also contain considerable extra background knowledge of filing a complaint or writing a polite email in a professional setting.
4	The language of email is generally clear, cogent and smooth, and displays general accuracy in grammatical and lexical choices from a wide range. Fairly consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some minor lexical or grammatical errors, which may or may not interfere with understanding.	The email information is generally accurate and relevant to the task and shows a generally good understanding of the reading input. The form of the email is fairly complete, maybe with minor missing parts.	The email includes fairly clear and logical explanations of the student's situation. The relationships between ideas are fairly supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are fairly persuasive.	The email is fairly developed with considerations and examples from the reading input. The email might also contain some extra background knowledge of filing a complaint or writing a polite email in a professional setting.
3	The language of email is moderately clear and cogent, and displays moderate accuracy in grammatical and lexical choices from a moderately wide range. Somewhat consistent pragmatic appropriateness and politeness in tone, register, and stance, but may contain obvious inconsistency. The email has a noticeable amount of grammatical or lexical errors, which slightly interfere with understanding.	The email information is moderately accurate and relevant to the task, and shows some understanding of the reading input. The form of the email is moderately complete, but a few parts are missing.	The email includes some clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are somewhat persuasive.	The email is somewhat developed with details or considerations from the reading input. The email might also contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting.
2	The language of email is clear or cogent at times but exhibits problems in being consistent. The email displays an inappropriate tone, register, or stance. The language either shows limits in control of vocabulary and grammar, or has a large amount of errors, which significantly interfere with understanding.	The email information is somewhat relevant to the task but leaves most of it unattended. The response fails to show understanding or consideration of the reading input. The form of the email is incomplete.	The email barely includes any clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are barely persuasive.	The email lacks important considerations and examples from the reading input. The email might sometimes contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting.
1	The language of email is very limited in coherence or clarity without any pragmatic appropriateness considerations. Language only has discreet words or phrases that barely connect, or is full of lexical and grammatical errors, which severely interfere with understanding.	The email information is not relevant to the task. The response fails to show understanding or consideration of the reading input. At this level, the form of the email is either missing important parts or only has few parts.	The email has poor and illogical explanations of the student's situation without any logical connectors or/and cohesive devices. The explanations and requests made in the email are not persuasive.	The email lacks considerations and examples from the reading input. The email fails to contain extra background knowledge of filing a complaint or writing a polite email in a professional setting.

Table 3. Item Coding and Keys for the Reading Section.

Observed Variable	Item Number	Answer Key
Endophoric Literal	1	**D**
	3	**C**
	4	**B**
	6	**C**
	7	**B**
	10	**D**
Endophoric Implied	2	**A**
	5	**D**
	8	**B**
	9	**D**
	11	**C**
	12	**A**

The test was designed into 2 parts, reading comprehension, and extended polite email writing. Once the test-takers completed the first part, they received detailed feedback (explanation for different options) for each multiple-choice question that was delivered in Google Forms. After the test-takers finished reading the feedback, they proceeded to the second part to write a formal email to the professor based on what they had learned from part 1. Test-takers can write their response in either the given Google forms or in a Google doc.

2.2 Feedback System

Multiple types of feedback were given to the test-takers in the test, in both summative and formative ways. The first feedback was given in between the two parts, where students could utilize the feedback for their reading section performance to pave the way for the email writing in the second part. After the two raters finished scoring all the responses, they gave each test-taker a score report. The score report included the overall score of the student, their score for each component, explanations on how each construct was rated, edits of their email, together with individualized detailed feedback on how they could possibly improve. The reports were also sent to the class teachers to help them better understand where the students were and how to close their learning gap.

Based on the scores, most of the test-takers showed good comprehension of the information from the reading passage, and most of them successfully composed an email that showed at least moderate understanding or consideration of the reading passage. This test also critically and effectively evaluated how much CLP students comprehensibly learned the grammar focus and the socio-pragmatic knowledge focus from Unit 10 on how to properly initiate a complaint.

Below is an example score report (Table 4).

Table 2. (*continued*)

Score (Level of Control)	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
2	The language of email is clear or cogent at times but exhibits problems in being consistent. The email displays an inappropriate tone, register, or stance. The language either shows limits in control of vocabulary and grammar, or has a large amount of errors, which significantly interfere with understanding	The email information is somewhat relevant to the task but leaves most of it unattended. The response fails to show understanding or consideration of the reading input. The form of the email is incomplete	The email barely includes any clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are barely persuasive	The email lacks important considerations and examples from the reading input. The email might sometimes contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting
1	The language of email is very limited in coherence or clarity without any pragmatic appropriateness considerations. Language only has discreet words or phrases that barely connect, or is full of lexical and grammatical errors, which severely interfere with understanding	The email information is not relevant to the task. The response fails to show understanding or consideration of the reading input. At this level, the form of the email is either missing important parts or only has few parts	The email has poor and illogical explanations of the student's situation without any logical connectors or/and cohesive devices. The explanations and requests made in the email are not persuasive	The email lacks considerations and examples from the reading input. The email fails to contain extra background knowledge of filing a complaint or writing a polite email in a professional setting

Google forms, an online survey platform, was used to administer the test, allowing users to create customized and complex surveys and receive statistical reports after the data had been collected. In the preceding lecture, students were informed of the test structure and the format.

Eight students in community language program (CLP) Upper-intermediate 3 took the exam. They were given the link to the Google forms test in the Zoom chat box after the test instructions. While taking a test, students were required to remain in the Zoom conference and keep the camera on. Students were also told that the test time was 60 min for both sections of the test, during which they could go back to different sections to check their responses. All students were told to submit their tests when the time was up, and the class then moved on to regular class activities.

Table 2. (*continued*)

Score (Level of Control)	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
4	The language of email is generally clear, cogent and smooth, and displays general accuracy in grammatical and lexical choices from a wide range. Fairly consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some minor lexical or grammatical errors, which may or may not interfere with understanding	The email information is generally accurate and relevant to the task and shows a generally good understanding of the reading input. The form of the email is fairly complete, maybe with minor missing parts	The email includes fairly clear and logical explanations of the student's situation. The relationships between ideas are fairly supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are fairly persuasive	The email is fairly developed with considerations and examples from the reading input. The email might also contain some extra background knowledge of filing a complaint or writing a polite email in a professional setting
3	The language of email is moderately clear and cogent, and displays moderate accuracy in grammatical and lexical choices from a moderately wide range. Somewhat consistent pragmatic appropriateness and politeness in tone, register, and stance, but may contain obvious inconsistency. The email has a noticeable amount of grammatical or lexical errors, which slightly interfere with understanding	The email information is moderately accurate and relevant to the task, and shows some understanding of the reading input. The form of the email is moderately complete, but a few parts are missing	The email includes some clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are somewhat persuasive	The email is somewhat developed with details or considerations from the reading input. The email might also contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting

(*continued*)

Table 1. The Test Structure for Unit 10

Test Component	Task type	Number of Items/ Tasks	Time	Scoring
Reading: -Endophric literal (e.g., summarizing the gist, identifying details) -Endophoric implied (e.g., infering)	Selected response: -MC items	12 items	60 min in total	- Dichotomous (0/1) - 12 points in total
Writing: -Language Control -Content Accuracy -Rhetorical Control -Content Elaboration Theme: Polite email in the academic domain	Constructed-response task/ extended production: -Integrated reading-for-writing task with scaffolding			Analytic Scoring with a rubric
				-Rating scale ranging from 1–5 criteria
				- 5 points for each
				- 20 points in total
				- 2 raters

Table 2. The Writing Scoring Rubrics used on this test

Score (Level of Control)	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
5	The language of email is clear, cogent and smooth, and displays high accuracy in grammatical and lexical choices from a wide range. Consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some very minor lexical or grammatical errors which do not interfere with understanding	The email information is accurate and relevant to the task. The form of the email is complete and shows an accurate understanding of the reading input	The email includes clear and logical explanations of the student's situation. The relationships between ideas are supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are persuasive	The email is fully developed with considerations and examples from the reading input. The email may also contain considerable extra background knowledge of filing a complaint or writing a polite email in a professional setting

(*continued*)

3. To what extent were the raters consistent when rating writing ability in this test?
4. What is the nature of the relationship between the test-takers' self-reported length of studying English, and their performance on the test?

1.2 Organization

The organization of this work is as follows. Section 2 will first present procedure of the theoretical conceptualizations for the constructs of the test is measuring, including online big-data collection and feedback system used for a learning-oriented approach, assessing reading abilities and providing detailed feedback to improve writing. In Section 3 , this paper will elaborate on the statistical analysis of the results of the assessment. Section 4 of this work will conclude with a discussion of the main findings, the major limitations, as well as recommendations for future tests or research directions.

2 Online Big-Data Collection and Feedback System

2.1 Online Big-Data Collection: Platform and Procedures

With a learning-oriented approach in mind, twelve multiple-choice questions for the reading passage were designed in a way that they could also serve as the input for the writing assignment. The reading passage was about how to write a polite email to the professor and the test-takers will be assessed in terms of their endophoric literal and endophoric implied reading abilities. The reading passage with salient discourse markers also acts as a socio-transactional and linguistic resource for test-takers for the writing section.

To complete the writing assignment, students had to go through three stages: read, reflect, and write. To begin, students were given a reading passage on how to politely write an email to professors where they need to identify and highlight essential features of the input to finish the reading questions. During the reflection phase, they received feedback for each question once they completed all the multiple-choice items. The first two steps of the process were meant to teach students how to construct a professional and polite email through reading and reflection, as well as to provide them with background knowledge and resources to write their own emails in the second part of the test (Table 1).

A writing scoring rubric based on four components was used to evaluate the students for their writing section. The components took into consideration the language control, the content's accuracy and elaboration, and the rhetorical control. The sociolinguistic appropriateness like tone and formality were incorporated under language control. The scoring results were provided to the students together with feedback on various parts they should improve on. For the actual scoring part, both raters first independently rated the four constructs for each response, and then they averaged their scores on each construct and the total added-up score as well.

Each test-taker received a score report that has both summative scores (i.e., each construct's score and an overall score) and formative feedback (i.e., how their response was scored, a complete version of the rubrics for their reference, interpretations of each construct's score, global comment, as well as suggestions on how their writing could be improved) (Tables 2 and 3).

The motivation behind this paper lies in the development and implementation of a learning-oriented achievement test for ESL students at the Upper Intermediate Level. The test aims to assess reading comprehension and writing skills in a professional context, focusing on the theme of "The Art of Complaining" from Unit 10 of the coursebook. In order to ensure the relevance and effectiveness of the test, it is essential to explore the nature of reading and writing abilities on this test and investigate the relationship between these two constructs.

The literature review conducted for this study revealed valuable insights into the relationship between reading and writing abilities in language learning. Studies by Bazerman (1980) and Cumming et al. (2004) highlighted the interconnectedness of reading1 and writing, showcasing how these skills mutually reinforce each other [4, 12]. Additionally, the works of Berninger et al. (2002) and Grabe (1991) contributed to a deeper understanding of how language by hand (writing) and language by eye (reading) are closely intertwined in the learning process [5, 16].

By drawing on the insights from these literature reviews, the test design was strategically developed to create an integrated learning-oriented assessment. The counterpart not only provides students with summative scores but also offers formative feedback, aiding their improvement in both reading comprehension and writing abilities. The design of test aligns with the established theoretical frameworks and aims to support the development of real-world language and communicative competencies.

Through the utilization of an online platform, the test was administered using a big-data driven approach, following the footsteps of Gebril and Plakans (2009) and Plakans & Gebril (2012) [15, 26]. This approach harnessed the benefits of data analysis from student responses, providing valuable insights into student performance and test effectiveness. Furthermore, the statistical evaluation conducted in this study adheres to the principles of validity and reliability highlighted by Borsboom et al. (2004) and Myford & Wolfe (2003) [7, 24].

This paper aims to achieve three primary objectives:

1. To provide a comprehensive overview of the learning-oriented test, including its design process, the test constructs it measures, the delivery methods employed, and the feedback mechanisms in place.
2. To conduct a statistical evaluation of the reliability and validity of the test, meticulously assessing whether the test aligns with its intended claims.
3. To collect the data, we utilized an online platform, adopting a truly online big-data driven approach. This allowed us to benefit from an extensive array of student responses and conduct thorough data analysis, providing valuable insights into student performance and the effectiveness of the test.

The paper will provide detailed information about the test design and present a comprehensive statistical analysis of the results. The discussion section will present our main findings, address any significant limitations, and offer recommendations for future research and test design enhancements.

The following research questions will be addressed in this paper:

1. What is the nature of reading and writing ability on this test?
2. What is the nature of the relationship between reading and writing ability on this test?



An Online Big-Data Driven Design of Reading and Writing Test

Yuwei Sun[1], Yongcheng Wen[2(✉)], and Yazhen Zhu[3]

[1] Columbia University, New York, NY 10027, USA
[2] Shenzhen MSU-BIT University, Shenzhen, Guangdong 518172, People's Republic of China
1120200244@smbu.edu.cn
[3] Royal College of Art, London SW7 2EU, UK

Abstract. This paper presents an online big-data driven design of reading and writing tests, incorporating empirical data analysis. The study aims to investigate the nature of reading and writing abilities, their corresponding relationship, and the impact of background variables on learning-oriented test performance. The counterpart was administered through an online platform, where data are collected for assessing the performance of students. The objectives of our work are to provide insights into the test design, delivery, and feedback mechanisms, and to conduct a statistical evaluation of the test's reliability, validity, and correlations. The findings contribute to our understanding of reading and writing assessment in an online context, while also highlighting the implications of background variables on test performance.

Keywords: Learning-oriented test · Online big-data driven task · Reliability & Validity of the test

1 Introduction

1.1 Motivation

Language proficiency assessment plays a crucial role in language learning and teaching. However, traditional assessment methods often focus solely on measuring students' outcomes without considering their learning progress and individual growth. In contrast, a learning-oriented test model offers a valuable approach that emphasizes the learning process itself and provides formative feedback to support students' improvement.

The objective of this work is two-fold:

- To provide information about the learning-oriented test in terms of the process of design, its test constructs, delivery, and feedback.
- To conduct a statistical evaluation of the reliability and validity of the test and examine whether the test supports its claims.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 335–353, 2024.
https://doi.org/10.1007/978-3-031-65126-7_29

Networks and Applications

13. Li, D., Wang, J.: Fedmd: heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581 (2019)
14. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 1–19 (2019)
15. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
16. Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., Qi, H.: Beyond inferring class representatives: user-level privacy leakage from federated learning. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, pp. 2512–2520. IEEE (2019)

5 Conclusion

This paper proposes a personalized federated learning scheme based on differential privacy. On the basis of traditional federated learning, transfer learning and knowledge distillation methods are used to complete the personalized setting, and Gaussian mechanism is introduced for privacy protection. This paper uses Mnist data set to experiment the performance of the algorithm, understand the impact of privacy budget on the performance of the algorithm, and compare with the Laplace mechanism. From the experimental results, compared with the noise-free scheme, the proposed scheme has little accuracy loss, and the Gaussian mechanism is greatly superior to the Laplace mechanism in reducing the accuracy loss.

References

1. Ning, Z., et al.: Mobile edge computing enabled 5G health monitoring for internet of medical things: a decentralized game theoretic approach. IEEE J. Sel. Areas Commun. **39**(2), 463–478 (2020)
2. Wang, X., Ning, Z., Guo, S., Wen, M., Guo, L., Poor, V.: Dynamic UAV deployment for differentiated services: a multi-agent imitation learning based approach. IEEE Trans. Mob. Comput. **22**(4), 2131–2146 (2021)
3. Ning, Z., et al.: Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE Trans. Mob. Comput. **22**(5), 2628–2644 (2021)
4. Konečnỳ, J., McMahan, H.B., Ramage, D., Richtárik, P.: Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527 (2016)
5. Pasdar, A., Lee, Y.C., Hong, S.H.: Mask off: analytic-based malware detection by transfer learning and model personalization. arXiv preprint arXiv:2211.10843 (2022)
6. Chen, Y., Zhang, T., Jiang, X., Chen, Q., Gao, C., Huang, W.: Fedbone: towards large-scale federated multi-task learning. arXiv preprint arXiv:2306.17465 (2023)
7. Vettoruzzo, A., Bouguelia, M.R., Vanschoren, J., Rögnvaldsson, T., Santosh, K.: Advances and challenges in meta-learning: a technical review. arXiv preprint arXiv:2307.04722 (2023)
8. Zhu, Z., Hong, J., Zhou, J.: Data-free knowledge distillation for heterogeneous federated learning. In: International Conference on Machine Learning, pp. 12878–12889. PMLR (2021)
9. Ozkara, K., Girgis, A.M., Data, D., Diggavi, S.: A statistical framework for personalized federated learning and estimation: theory, algorithms, and privacy. In: The Eleventh International Conference on Learning Representations (2022)
10. Yuan, B., Qiu, W.: Personalized federated learning system based on permissioned blockchain. In: 2021 International Conference on Intelligent Computing, Automation and Systems (ICICAS), pp. 95–100. IEEE (2021)
11. Wang, X., Ning, Z., Guo, L., Guo, S., Gao, X., Wang, G.: Mean-field learning for edge computing in mobile blockchain networks. IEEE Trans. Mob. Comput. **22**(10), 5978–5994 (2022)
12. Wu, J., et al.: Hierarchical personalized federated learning for user modeling. In: Proceedings of the Web Conference 2021, pp. 957–968 (2021)

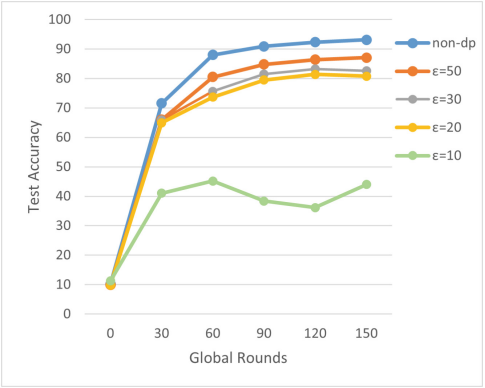


Fig. 3. System accuracy with different privacy budgets. (Color figure online)

the lower the degree of privacy protection, so according to Fig. 3, we can choose $\epsilon = 20$ as the appropriate degree of privacy protection.

Figure 4 compares the system accuracy when Gaussian noise is added with Laplacian noise. It can be seen that under the same privacy budget, adding Gaussian noise to the framework for privacy protection is better than adding Laplacian noise, and the contrast is extremely obvious. The reason is presumably that, unlike the Laplace mechanism, the noise added by the Gaussian mechanism is variable and can be adjusted according to the characteristics of the data and the type of query. This adjustable noise makes the Gaussian mechanism more accurate when dealing with large-scale data, and it is also more suitable for processing application scenarios that require high-precision query results.

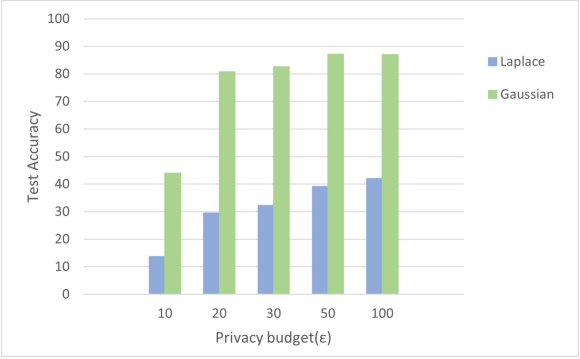


Fig. 4. The comparison of Gaussian noise and Laplacian noise.

Algorithm 2. Personalized Federated Learning based on Differential Privacy.

Input: Amount of participants N, amount of system iterations K, amount of local iterations ep, local optimized sample batch size bs, local optimized learning rate η, privacy budget ϵ.
Output: Participant personalization model W^*.
[Clients]
Initialize the local model $W^{(0)}$.
The local amount of data T_n is sent to the central server.
for $k = 1, 2...K$ **do**
    Receive $W^{(k-1)}$ from the server.
    $W_n^{(k)} \leftarrow M - SGD_n(W_n^{(k-1)}, ep, bs, \eta_n^{(k)})$
    Send $W^{(k)}$ to the server.
end for
[Personalized Federated Learning Server]
Random initialization $W^{(0)}$.
The amount of data T_n received from each participant.
Calculate $\beta_n = \frac{T_n}{\sum_{n=1}^{N} T_n}$.
Send $W^{(0)}$ to each client.
for $k = 1, 2...K$ **do**
    Receive W_n^k from each client.
    $W^* \leftarrow \sum_{n=1}^{N} \beta_n W_n^k$
    Send W^* to each client.
end for

to the label size, the algorithm was divided into 200 groups, each group had 250 data, and each user randomly selected two groups, and each group could not be selected repeatedly. After building the local dataset for each user, 90 percent of the local dataset is randomly selected as the local training dataset and the remaining part is used as the local test set.

In this experiment, the number of global iterations is set to 150, the number of iterations is 1, and the learning rate is 0.1.

4.2 Results and Analysis

Figure 3 represents the system accuracy when taking different privacy budgets. The figure shows the accuracy of the system when the privacy budget is 10,20,30, and 50. It can be seen that when $\epsilon = 10$, the accuracy of the system is very low, but when $\epsilon = 20$, the accuracy of the system is greatly improved, and there is little difference between the accuracy of the system when the privacy budget is 20,30, and 50. The blue curve represents the system accuracy when no noise is added. It can be seen that the training accuracy of personalized federated learning is higher when no noise is added. The experimental results show that adding noise reduces the accuracy of the user's local personalized model to a certain extent, and the smaller the privacy budget, the greater the accuracy of the personalized model. However, we know that the larger the privacy budget,

2) Noise Settings: For numerical data, this can be guaranteed using the Gaussian mechanism defined in [16]. According to [16], we can propose the following DP mechanism by adding artificial Gaussian noise.

To ensure that a given noise distribution $n - G(0, \sigma^2)$ preserves (ϵ, δ) - DP, where G represents the Gaussian distribution, we use

$$C = \frac{\sqrt{2 \cdot log(\frac{1.25}{\delta})}}{\epsilon} \tag{7}$$

to calculate the noise parameters.
In this paper, if the privacy budget of the client is ϵ, and the client is selected T times, the budget of the client per noise is ϵ/T.

3) Parameter aggregation: Before training starts, each participant sends its own amount of available training data to the server, and the server computes weights based on the amount of data T_n for each participant

$$\beta_n = \frac{T_n}{\sum_{n=1}^{N} T_n} \tag{8}$$

When the personalized federated learning server aggregates parameters, the base layer parameters of each participant are weighted and averaged according to the weight β_n of each participant

$$W^* = \sum_{n=1}^{N} \beta_n W \tag{9}$$

In summary, the algorithm formulation of personalized federated learning scheme based on differential privacy is as Algorithm 2.

4 Experiment and Analysis

4.1 Experiment Settings

In this experiment, it is assumed that there are 100 clients participating in the training of personalized federated learning, and the data distribution of users is heterogeneous. A common neural network architecture, Convolutional Neural network (CNN), is used for the user local model, and the loss function is the cross-entropy loss function. In this experiment, the training accuracy is used as the performance evaluation index, and the mnist dataset is used for performance testing.

The Mnist dataset is a commonly used database of handwritten digits. It contains 60,000 training images and 10,000 test images, where each image is made up of 28*28 pixels. The labels in the Mnist dataset include ten numbers from 0 to 9 and are one-hot encoded.

In this experiment, 60000 training data of the Mnist dataset are sorted and divided into two parts. One part has 10000 data, which is used as a public data set, and the remaining 50000 data is used as a private data set. Then, according

Table 1. The representative meanings of the different parameters.

Parameters	Description
N	the amount of clients
W_j	the weight of the j th clients
α_j	the activation function of the j vector
x	the output of the n th user
y	the result of each round of forward propagation
L	the common loss function of each participant
W^*	the global optimal solution obtained after optimization
$(x_{n,t}, y_{n,t})$	the t th sample in the training data
T_n	the amount of training data available to the n th participant
δ	Term of relaxation
β	Weights when parameters are aggregated
C	the noise parameters
σ	standard deviation
ϵ	privacy budget
K	amount of system iterations
ep	amount of local iterations
bs	local optimized sample batch size
η	local optimized learning rate

3.2 Algorithm Settings and Description

The algorithm setup mainly includes the following parts: model setup, noise setup, and parameter aggregation.

1) Model setup: Suppose there are N clients participating in the training, then the weight matrix is $(W_N, ..., W_2, W_1)$. If the activation function of the vector is denoted by $(\alpha_N, ..., \alpha_2, \alpha_1)$, then for the output x of the n th user, the result of each round of forward propagation can be denoted by

$$y = \alpha_N(W_N \cdots \alpha_1(W_1 x) \cdots) \tag{5}$$

If let $L(\cdot, \cdot)$ denotes the common loss function of each participant, the learning objective is to minimize the average personalized loss function without adding noise, denoted by

$$W^* = argmin_W(L(W) = \frac{1}{N}\sum_{n=1}^{K}\frac{1}{T_n}\sum_{t=1}^{T_n} l(y_{n,t}, f(x_{n,t}; W))) \tag{6}$$

Where, T_n denotes the amount of training data available to the n th participant, $(x_{n,t}, y_{n,t})$denotes the t th sample in the training data, and $W^*$is the global optimal solution obtained after optimization.

Algorithm 1. The FedMD framework enabling federated learning for heterogeneous models.

Input: Public datasets D_0,private datasets D_k independently designed model $f_k, k = 1...m$,
Output: Trained model f_k
Transfer learning: Each party trains to convergence on the local D_k.
for $j = 1, 2...P$ **do**
    Communicate: Each party computes the class scores $f_k(x_i^0)$ on the datasets, and transmits the result to a central server.
    Aggregate: The server computes an updated consensus, which is an average $\tilde{f}(x_i^0) = \frac{1}{m}\sum_k f_k(x_i^0)$.
    Distribute: Each party downlosds the updated consensus$\tilde{f}(x_i^0)$.
    Digest: Each party trains its model f_k to approach the consensus $\tilde{f}$ on the public datasets D_0.
    Revisit: Each party trains its model f_k on its own private data for a few epochs.
end for

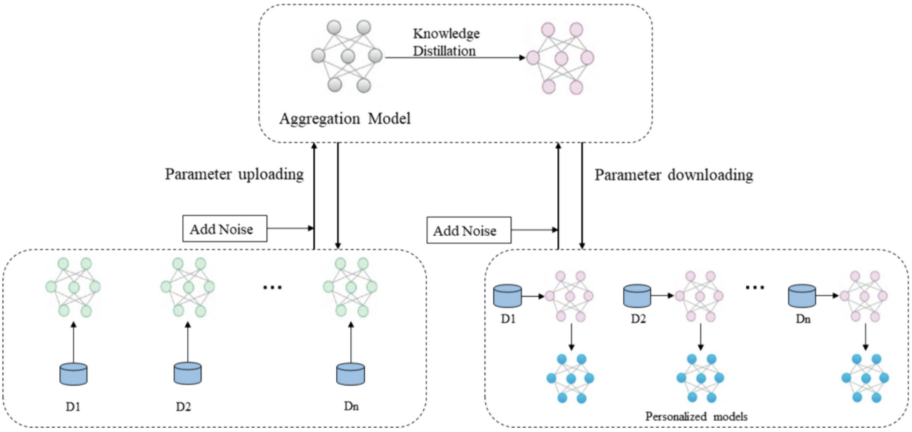


Fig. 2. Illustration of the algorithm framework.

Firstly, the client downloaded the initial model from the central server, and then used the public data set and private data set for local training. After the number of local iterations reached, the client added Gaussian noise to the model parameters, and segmented the perturbed parameter set according to the maximum and minimum values of the original parameters before uploading to the server, so as to prevent privacy leakage caused by the data upload process. Then, the central server performs knowledge distillation to compute the consensus, and the clients approach the consensus through training until convergence. Finally, the client performs local training on the private dataset to obtain a personalized local model. Table 1 lists the meaning of the parameters used in this paper.

$$\Delta_g f = max_D \Delta_l f(D) \tag{3}$$

After introducing the above concepts, the mathematical definition of differential privacy is stated as follows.

Definition 5: ϵ - Differential privacy. For two datasets D and D', where only one record differs, a randomization mechanism m protects ϵ-differential privacy, and for all $S \in Range(M)$have:

$$Pr[M(D) \in S] \leq e^{\epsilon} \cdot Pr[M(D') \in S] \tag{4}$$

Where, ϵ represents the privacy budget, which is used to control the privacy protection degree of the algorithm; the smaller is ϵ, the higher is the privacy protection degree; on the contrary, the lower is the privacy protection degree. From the Definition 5, we can see that when the parameter ϵ is smaller, the distribution of the query results of the random algorithm M on the adjacent data set is closer, and the privacy protection effect is better. When $\epsilon = 0$, the probability distribution of the query result is exactly the same, and the degree of privacy protection reaches the highest. On the contrary, the larger ϵ is, the less privacy is preserved.

2.3 FedMD Framework

In order to solve the negative impact of heterogeneity on model training, researchers have proposed a series of personalized methods. Among them, Li et al. [13] proposed a method FedMD based on transfer learning and knowledge distillation. The steps are as follows:

Step 1: Transfer learning. The client first trains on the public dataset until convergence, and then trains on the private dataset until convergence to obtain the local model.

Step 2: Knowledge Distillation. Clients compute class scores on a public dataset, upload the results to a central server, which computes a consensus and distributes it to clients. Clients train the model until they get close to the consensus. The knowledge of one participant can be understood by other participants, which completes the process of knowledge distillation.

Step 3: Revisit. The client trains the model on its own private dataset until convergence.

To minimize training overhead, this paper enhances the algorithm by conducting training solely on private datasets until convergence is attained in the transfer learning stage. The Algorithm procedure is shown in Algorithm 1.

3 Trusted Personalized Federated Learning

3.1 Framework

The personalized federated learning scheme based on differential privacy consists of a cloud server and N clients, and the system architecture of this scheme is shown in Fig. 2.

Centralized differential privacy: The client uploading the data to the data center, which then privacy-protects the data. However, the drawback of this method is that it requires a trusted third party to collect the data.

Local differential privacy: The client adds noise to the data locally and uploading it, so as to avoid the risk of privacy leakage caused by the data uploading process and the untrusted data collector.

Before describing the specific mathematical definition of differential privacy, it is important to understand the following key concepts:

1) Proximity data

Definition 1. Adjacent Dataset. Has two have the same properties of the structure of the data set D, D', if the data set D, D' differ by at most a record, namely $|D \oplus D'| = 1$, then according to the data set D, D' for data sets.

2) Query functions

Definition 2. Query functions. For a data set D various mapping functions are defined as queries, such as sum, median, range queries, etc., denoted as $f(D)$, and a set of query functions can be expressed as $F = f_1, f_2, ..., f_n$.

3) Global sensitivity and local sensitivity

Definition 3. Global sensitivity. Given a pair of proximity data sets D, D', for the query function $f : D \rightarrow R$, R is the query result returned by the query function, and the global sensitivity of function f is defined as

$$\Delta_g f = max_{D,D'} ||f(D) - f(D')||_1 \tag{1}$$

Among them, $||f(D) - f(D')||_1$ is the Manhattan distance between $f(D)$ and $f(D')$.

From Definition 3, we know that the sensitivity f of a query function is a property of the function itself, independent of the data set. This property represents the maximum variation in the result of the query function when applied to a pair of adjacent data sets. A corresponding concept is the local sensitivity, which mainly shows the change of the query function f on a data set, which is determined by the query function and the given data set. The definition is as follows:

Definition 4: Local sensitivity. Given a dataset D, for a query function f: $D \rightarrow R$, R is the query result returned by the query function, on any data set D' adjacent to the data set D. The local sensitivity of function f is defined as

$$\Delta_l f(D) = max'_D ||f(D) - f(D')||_1 \tag{2}$$

Among them, $||f(D) - f(D')||_1$ is the Manhattan distance between $f(D)$ and $f(D')$.

According to Definition 4, when the given data set D is the same as the data set that achieves the maximum Manhattan distance between $f(D)$ and $f(D')$ in Definition 3, the local sensitivity is the same as the global sensitivity, i.e.

introduction to federated learning. They classify federated learning into three types:

1) Horizontal federated learning: federated learning by sample;
2) Vertical federated learning: feature-wise federated learning;
3) Federated transfer learning: Federated learning with few participants and few features.

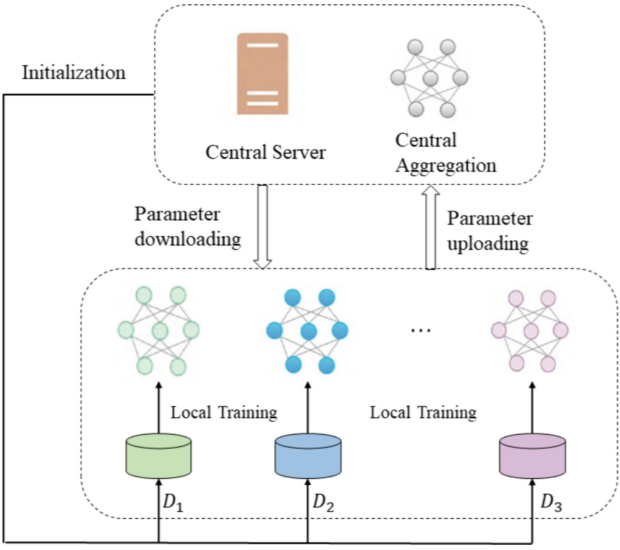


Fig. 1. Federated Learning flow.

The process of federated learning is illustrated in Fig. 1. Federated learning consists of a central server and multiple participants, and trains a global model after multiple iterations. The process is as follows:

1) System initialization. The central server initializes the global model and distributes it to the participants;
2) Local training. Where participants train the model using a local dataset and then upload the updated model parameters to a central server;
3) Central aggregation. Where the central server aggregates the received parameters to obtain the global model.

2.2 Differential Privacy

Dwork et al. [15] proposed differential privacy mechanism in 2006. By introducing random noise, the public output will not be significantly changed by the change of an individual, so as to achieve the effect of hiding the individual. Differential privacy can be divided into centralized differential privacy and localized differential privacy according to the different subjects adding noise.

as a federated learning server. And a personalized federated learning algorithm based on permission blockchain is proposed, which can achieve privacy protection and resist poisoning attacks with high accuracy. Experimental results show that the proposed system has high privacy protection and anti-poisoning attack ability, and can be deployed in edge computing environment. Recent research results such as the HPFL [12] framework also use a similar hierarchical idea to some extent. Participants in HPFL first divide local training data into public and private information, and then divide the model into public and private components based on this information. Deliver a draft of the private component. In the aggregation phase, the server directly weighted the public components with the same attribute as the common part of the global model, and aggregated the private components as the private part of the global model, so as to avoid the impact of data heterogeneity and complete privacy protection.

In general, the number of research results that consider both heterogeneity and privacy issues is small, and most of the existing results are constructed based on a certain kind of personalized federated learning framework, lacking certain scalability. At the same time, the introduction of privacy protection technology still brings the loss of model accuracy or the increase of computation and communication costs.

Inspired by this, we propose a privacy protection mechanism for personalized federated learning in this paper, which uses the local differential privacy mechanism to add Gaussian noise to the parameters uploaded to the central server to achieve more effective privacy protection without affecting the model performance, so as to build a personalized federated learning system with privacy protection ability. The main contributions of this paper are as follows:

1. We introduce the Gaussian mechanism into the modified FedMD [13] framework to achieve privacy protection while ensuring personalization.
2. Before the parameters are uploaded to the central server, the noise obeying Gaussian distribution is added to them, and the central server then performs knowledge distillation and aggregation of the noisy parameters, so as to avoid the original data from being obtained by the attacker to a certain extent.
3. This paper sets up experiments to explore the influence of privacy budget on the accuracy of the system, and compares the effect with the traditional Laplace mechanism to verify the effectiveness of the proposed scheme.

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 presents the system model as well as the algorithmic details. Experimental verification is carried out in Sect. 4. Finally, we conclude the paper in Sect. 5.

2 Related Work

2.1 Federated Learning

Federated learning was proposed by Google in 2016 for updating language prediction models on smartphones. In 2019, Yang et al. [14] gave a systematic

server to train models. However, this kind of centralized machine learning methods will encounter some challenges when solving practical problems. For example, during the transmission of a large amount of data, there will be high communication overhead and there will be the risk of privacy leakage. Therefore, how to train the model while ensuring privacy and low communication overhead has become a problem that artificial intelligence technology needs to consider. [3] The solution is given by the technology of federated learning [4] proposed by Google in 2016. Federated learning is a distributed machine learning framework that can learn by utilizing data available in decentralized devices while maintaining data privacy. Each client has local data to participate in the training. After completing the local training, the client uploading the model parameters or gradients to the server, and then the server aggregates the received information, and the client downloads the aggregated results, thus allowing users to obtain a shared model without compromising data privacy.

Despite the good performance reflected in the model training process, there are still some problems when federated learning is applied to real-world specific problems. A significant challenge arises from the non-uniform distribution of data among devices, each with varying computing, storage, and communication capabilities. Consequently, the model trained with private data may outperform the global model for certain users. Secondly, there are also privacy issues in federated learning. Attackers can easily infer the original data by observing the model parameters or gradient information uploaded by participants, or pass malicious parameters to participants to affect the model training effect, which will lead to different degrees of privacy leakage. In order to make the federated learning model in heterogeneous scenarios more in line with individual preferences, some studies have proposed the concept of personalized federated learning. They use methods based on transfer learning [5], multi-task learning [6], meta-learning [7], knowledge distillation [8] and other methods to solve heterogeneous problems, realize local or global model personalization, and make the model more suitable for each participant. However, most of the research on personalized federated learning does not consider privacy issues.

However, the number of research results for personalized federated learning is small. Recently, Ozkara et al. [9] proposed a statistical framework for personalized joint learning and estimation, and this work started with a statistical framework that unitized several different algorithms and proposed new ones. In this paper, we apply the framework to personalized estimation and connect it with the classical empirical Bayesian method. In this framework, a novel estimation of private personalization is developed. Then, new personalized learning algorithms are proposed, including AdaPeD based on information geometric regularization, which numerically outperforms several known algorithms. This work extends privacy considerations to personalized learning methods, ensuring user-level privacy and composition. Yuan et al. [10] proposed a personalized federated learning system based on permissioned blockchain, which is divided into a four-layer architecture, namely iot device layer, network layer, edge computing layer, blockchain layer and application layer, and uses permissioned blockchain [11]



Trusted Personalized Federated Learning Based on Differential Privacy

Ruixin Liu[1,3], Zhenquan Qin[1], Xi Cheng[1], Rui Zhang[2,3], and Jianbo Zheng[2,3(✉)]

[1] Dalian University of Technology, Dalian 116024, Liaoning, China
[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
jianbo.zheng@smbu.edu.cn
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

Abstract. As an emerging machine learning model, federated learning ensures that user data is stored locally while breaking data siloes, which ensures the privacy of training data. However, in practical applications, the training data across different user terminals are often non-independent and identically distributed. Moreover, federated learning does not provide strict privacy protection for users. Attackers can infer data information based on the model parameters or gradients uploaded or downloaded, and there is a great risk of privacy leakage in the process of information transmission. Therefore, we design a personalized federated learning scheme with privacy protection capability. Considering that non-independent and identically distributed data will have a negative impact on model training, we aim to obtain a personalized model for each client by joining transfer learning and knowledge distillation. Since we first propose a personalized federated learning algorithm with privacy protection ability based on the Gaussian mechanism in differential privacy, and then compared with the traditional Laplace mechanism. Experimental results demonstrate that our proposed method achieves better accuracy and improves the privacy protection ability.

Keywords: Federated learning · Personalized federated learning · Differential privacy

1 Introduction

With the rapid development of computers and the Internet, artificial intelligence technology emerges as The Times require [1,2]. Traditional machine learning methods require a large number of user terminals to upload data to a central

This work was supported in part by the Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 320–332, 2024.
https://doi.org/10.1007/978-3-031-65126-7_28

11. Thing, V.L.L.: IEEE 802.11 network anomaly detection and attack classification: a deep learning approach. In: Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6 (2017). https://doi.org/10.1109/WCNC.2017.7925567
12. Doshi, R., Apthorpe, N., Feamster, N.: Machine learning DDoS detection for consumer Internet of Things devices. In: Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), pp. 29–35 (2018). https://doi.org/10.1109/SPW.2018.00013
13. Cheng, Y., Xu, Y., Zhong, H., Liu, Y.: Leveraging semisupervised hierarchical stacking temporal convolutional network for anomaly detection in IoT communication. IEEE Internet Things J. **8**, 144–155 (2021). https://doi.org/10.1109/JIOT.2020.3000771
14. Ning, Z., Dong, P., Kong, X., Xia, F.: A cooperative partial computation offloading scheme for mobile edge computing enabled Internet of Things. IEEE Internet Things J. **6**, 4804–4814 (2019). https://doi.org/10.1109/JIOT.2018.2868616
15. Wang, X., et al.: Wireless powered mobile edge computing networks: a survey. ACM Comput. Surv. (2023). https://doi.org/10.1145/3579992
16. Nie, L., Ning, Z., Wang, X., Hu, X., Cheng, J., Li, Y.: Data-driven intrusion detection for intelligent internet of vehicles: a deep convolutional neural network-based method. IEEE Trans. Netw. Sci. Eng. **7**, 2219–2230 (2020). https://doi.org/10.1109/TNSE.2020.2990984
17. Nie, L., et al.: Intrusion detection in green internet of things: a deep deterministic policy gradient-based algorithm. IEEE Trans. Green Commun. Networking **5**, 778–788 (2021). https://doi.org/10.1109/TGCN.2021.3073714

7 Conclusion and Future Work

This study aims to achieve universal, accurate, and comprehensive anomaly detection in the IoT environment. From the perspective of device interaction behavior, a method for IoT anomaly traffic detection is proposed. By analyzing the manifestation of device interaction processes in network communication, a comprehensive representation of device interaction behavior is developed based on time and space attributes at four levels: packet, flow, session, and host. Subsequently, specific learning and modeling targeting interaction device pairs are conducted to construct an integrated anomaly detection model, achieving comprehensive anomaly detection. The propose method is evaluated using the BoT-IoT dataset, demonstrating its effectiveness and superiority. Considering the method's real-world application, the challenge of rapidly and automatically modeling for newly introduced devices becomes a focal point for future research.

References

1. Koroniotis, N., Moustafa, N., Sitnikova, E., Turnbull, B.: Towards the development of realistic botnet dataset in the internet of things for network forensic analysis: Bot-IoT dataset. Futur. Gener. Comput. Syst. **100**, 779–796 (2019). https://doi.org/10.1016/j.future.2019.05.041
2. Deorankar, A.V., Thakare, S.S.: Survey on anomaly detection of (IoT)-Internet of Things cyberattacks using machine learning. In: Proceedings of the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 115–117 (2020). https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00023
3. Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W., Wahab, A.: A review of intrusion detection systems using machine and deep learning in Internet of Things: challenges, solutions, and future directions. Electronics **9**, 1177 (2020)
4. Illy, P., Kaddoum, G., Miranda Moreira, C., Kaur, K., Garg, S.: Securing Fog-to-Things environment using intrusion detection system based on ensemble learning. In Proceedings of the 2019 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–7 (2019). https://doi.org/10.1109/WCNC.2019.8885534
5. Qiu, T., Liu, X., Zhou, X., Qu, W., Ning, Z., Chen, C.L.P.: An adaptive social spammer detection model with semi-supervised broad learning. IEEE Trans. Knowl. Data Eng. **34**, 4622–4635 (2022). https://doi.org/10.1109/TKDE.2020.3047857
6. Dutta, V., Chora's, M., Pawlicki, M., Kozik, R.: A deep learning ensemble for network anomaly and cyber-attack detection. Sensors **20** (2020). https://doi.org/10.3390/s20164583
7. Abdelmoumin, G., Rawat, D.B., Rahman, A.: On the performance of machine learning models for anomaly-based intelligent intrusion detection systems for the Internet of Things. IEEE Internet Things J. **9**, 4280–4290 (2022). https://doi.org/10.1109/JIOT.2021.3103829
8. Wang, X., Zhu, H., Ning, Z., Guo, L., Zhang, Y.: Blockchain intelligence for internet of vehicles: challenges and solutions. IEEE Commun. Surv. Tutor. (2023). https://doi.org/10.1109/COMST.2023.3305312
9. Wang, X., et al.: Toward accurate anomaly detection in industrial Internet of Things using hierarchical federated learning. IEEE Internet Things J. **9**, 7110–7119 (2022). https://doi.org/10.1109/JIOT.2021.3074382
10. Liu, Y., Kumar, N., Xiong, Z., Lim, W.Y.B., Kang, J., Niyato, D.: Communication-efficient federated learning for anomaly detection in industrial Internet of Things. In: Proceedings of the GLOBECOM 2020–2020 IEEE Global Communications Conference, pp. 1–6 (2020). https://doi.org/10.1109/GLOBECOM42002.2020.9348249

6.3 Detection Performance

In this study, we initially employed ten-fold cross-validation to validate the model, and the results are presented in Table 3. The outcomes straightforwardly demonstrate the effectiveness and utility of the propose model.

Table 3. Ten-fold cross-validation results.

ID	P	R	Acc	AUC	P-R
1	97.13%	96.18%	95.07%	0.94	0.98
2	97.23%	96.02%	95.05%	0.94	0.98
3	97.35%	95.89%	95.07%	0.94	0.98
4	97.31%	97.43%	96.13%	0.95	0.98
5	97.52%	96.77%	95.84%	0.94	0.98
6	96.43%	97.52%	95.50%	0.94	0.98
7	97.13%	97.32%	95.89%	0.95	0.98
8	97.25%	96.40%	95.40%	0.95	0.98
9	98.12%	97.10%	96.01%	0.96	0.98
10	97.99%	96.87%	95.88%	0.95	0.98

Taking into account the real-world scenario's distribution between normal network traffic and attack traffic, we conducted five rounds of validation by controlling the ratio between training and testing data. We divided all the data into different portions using various values of λ (the ratio of training data to testing data), and the model's detection performance is illustrated in Table 4. According to the detection results, the domain-aware anomaly detection approach proposed in this study can accurately identify all abnormal behaviors that deviate from normal scenario settings.

Table 4. Model effect under different traffic ratios.

λ	P	R	Acc	AUC	P-R
50%:50%	92.10%	99.78%	93.79%	0.89	0.96
60%:40%	92.98%	99.52%	94.00%	0.89	0.96
70%:30%	92.90%	99.44%	94.00%	0.89	0.97
80%:20%	92.18%	99.50%	93.62%	0.89	0.96
90%:10%	93.51%	98.42%	93.83%	0.9	0.97

6.2 Evaluation Metrics

Anomaly detection is a binary classification problem. In this study, we employ five metrics to evaluate the performance of the model: Precision (P), Recall (R), Accuracy (Acc), area under the Precision-Recall (P-R) curve, and Area Under ROC Curve (AUC). The definitions of P, R, and Acc are shown in Eqs. (4), (5), and (6):

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{4}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{5}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

TP, TN, FP, and FN respectively stand for: True Positives, True Negatives, False Positives, and False Negatives. AUC represents the area under the ROC curve. The ROC curve has False Positive Rate (FPR) on the horizontal axis and R on the vertical axis. If the ROC curve of one classifier is entirely "enclosed" by the ROC curve of another, it can be inferred that the latter has better performance than the former. However, situations with crossing curves might also occur. Therefore, a more reasonable approach is to compare the areas under the ROC curves. The computation of FPR is given by Eq. (7):

$$\mathrm{FPR} = \frac{FP}{TN + FP} \tag{7}$$

The P-R curve plots R on the x-axis against P on the y-axis. Similar to AUC, the area under the P-R curve reflects the proportion of a learner's performance that achieves a relatively high balance between Precision and Recall.

Table 2. The type of attack marked by the BoT-IoT dataset.

Attack Types	Subtypes
DoS	DoS HTTP
	DoS TCP
	DoS UDP
DDoS	DDoS HTTP
	DDoS TCP
	DDoS UDP
Scan	OS Scan
	Service Scan
Theft	Data Exfiltration Keylogging

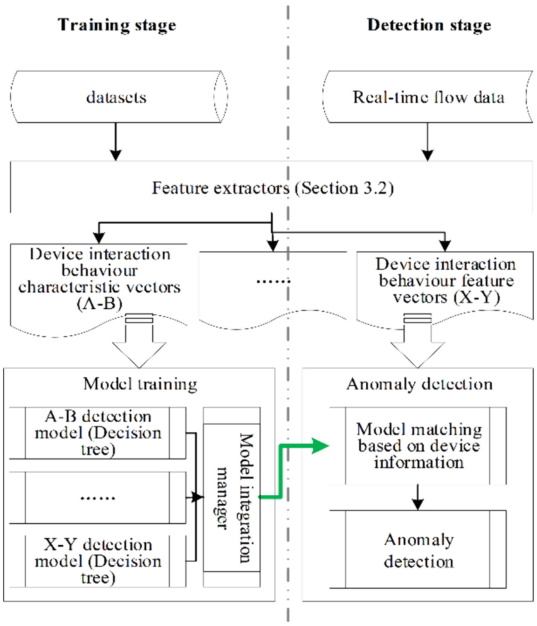


Fig. 5. Anomaly detection model.

Detection Phase: In real-time, the collected flow data undergoes feature extraction using a feature extractor to generate feature vectors for each interacting device pair. By leveraging IP identification information of the interacting devices, their corresponding "baseline" models are selected for anomaly detection. If a feature space deviates from the baseline, it is classified as an anomaly and triggers an alert directly.

6 Experiment Validation

6.1 Dataset

In this study, we conducted experiments and evaluations using the BoT-IoT dataset. The BoT-IoT dataset was created by designing a real-world network environment within the University of New South Wales Canberra Network Range Laboratory. This environment combines normal traffic with zombie network traffic. The dataset's source files are provided in various formats, including original pcap files, generated argus files, and CSV files. These files are segregated based on attack categories and subcategories to facilitate the labeling process. The captured pcap files amount to 69.3 GB in size, encompassing over 72,000,000 records. Extracted CSV-format traffic data constitutes a size of 16.7 GB. The dataset includes attacks such as DDoS, DoS, operating system and service scanning, keystroke logging, and data leakage. Furthermore, DDoS and DoS attacks are further organized based on the protocols employed. Table 2 illustrates the labeled attack types present within the BoT-IoT dataset.

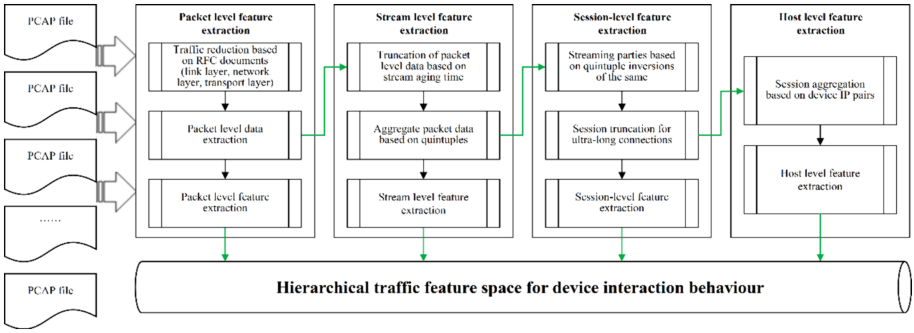


Fig. 4. Feature extraction flow char.

data. The format of flow-level data is as follows: second-level timestamp, source IP, source port, destination IP, destination port, protocol number, packet count, byte count, and flow interval. The timestamp for flow data is taken from the maximum time within that packet group. The features extracted are packet count, byte count, and flow interval.

Extraction of Session-Level Features: Flow data is aggregated based on identical five-tuples (as the sender) or reverse five-tuples (as the receiver). For aggregated data, if the time interval exceeds 30 min, it is truncated and divided into multiple sessions. The format of session-level data is as follows: second-level timestamp, sender IP, sender port, receiver IP, receiver port, protocol number, sender packet count, sender byte count, receiver packet count, receiver byte count, total packet count, total byte count, sender duration, receiver duration, total duration, sender packet count sequence, sender byte count sequence, receiver packet count sequence, receiver byte count sequence, number of flows in session.

Extraction of Host-Level Features: Session-level data is aggregated based on (sender IP, receiver IP) to obtain host-level data in the format: sender IP, receiver IP, sender port sequence, receiver port sequence, session count, total packet length, total packet count. Corresponding features are extracted using the calculation formula for information entropy.

5.3 Detection Framework

Built upon the stability characteristics of device interactions in the IoT ecosystem, the core idea of this research is to model the interactions between device pairs present in the network traffic collected at the IoT egress point. This modeling process is conducted individually for each device pair, enabling precise and comprehensive anomaly detection.

As depicted in Fig. 5, the propose IoT anomaly traffic detection framework comprises two main phases: the training phase and the detection phase.

Training Phase: During this phase, normal network traffic generated during regular device operation is collected as training samples. The interactions between device pairs are aggregated to extract respective feature vectors. To ensure practicality in the detection framework and to model each set of interactions, a simple decision tree is employed for autonomous modeling. This results in the formation of a baseline model ensemble that encompasses various groups of interacting devices.

Table 1. (*continued*)

Feature Hierarchy	Feature Name				
Host Level	Number of Open Ports by Sender	Port Information Entropy of the Sender	Number of Open Ports by Receiver	Port Information Entropy of the Receiver	Total Number of Open Ports Between Hosts
	Port Information Entropy Between Hosts	Number of Connection Sessions Between Hosts	Total Number of Packets in Communication Between Hosts	Total Packet Length in Communication Between Hosts	Packet Length Slope

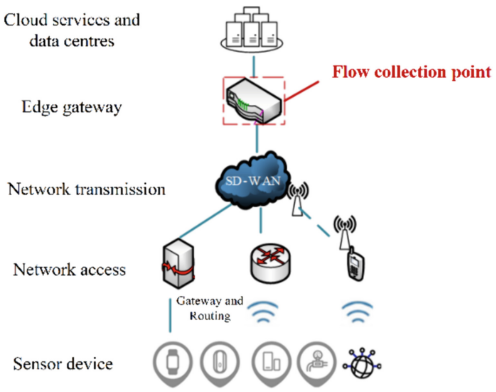


Fig. 3. Typical topology of the IoT.

5.1 Traffic Collection

A typical topology of an IoT network is illustrated in Fig. 3. The traffic generated by devices is initially aggregated at the edge gateway and subsequently funneled through routers before being connected to the Internet. Consequently, at the edge gateway, the incoming and outgoing traffic of devices can be collected in real-time.

5.2 Feature Extraction

According to Sect. 4.2, the feature space for representing interactive behaviors is illustrated in the process diagram shown in Fig. 4.

Extraction of Flow-Level Features: The packet-level data is truncated based on a flow aging time of 15 s, as defined in this study. The data is grouped every 15 s. For each group of data, the packet-level data's five-tuple (source IP, source port, destination IP, destination port, protocol number) is aggregated as key-value pairs to obtain flow-level

Table 1. Interaction behavior features space between different IoT devices.

Feature Hierarchy	Feature Name				
Package Level	Package Length	-	-	-	-
Flow Level	Total Package Length	Total Number of Packages	Flow Duration	-	-
Session Level	Number of Packets Sent by Sender	Total Packet Length Sent by Sender	Average Packet Length Sent by Sender	Minimum Packet Length Sent by Sender	Maximum Packet Length Sent by Sender
	Packet Length Standard Deviation Sent by Sender	Packet Count Standard Deviation Sent by Sender	Packet Length Slope of the Sender	-	-
	Number of Packets Received by Receiver	Total Packet Length Received by Receiver	Average Packet Length Received by Receiver	Minimum Packet Length Received by Receiver	Maximum Packet Length Received by Receiver
	Packet Length Standard Deviation Received by Receiver	Packet Count Standard Deviation Received by Receiver	Packet Length Slope of the Receiver	-	-
	Total Number of Packets in the Session	Total Packet Length in the Session	Average Packet Length in the Session	Minimum Packet Length in the Session	Maximum Packet Length in the Session
	Packet Length Standard Deviation in the Session	Packet Count Standard Deviation in the Session	Packet Length Slope in the Session	Session Duration	-

(*continued*)

utilized by the user, while the number of ports, as an essential attribute at the host level, can effectively reflect the quantity and variety of accessed services in communication behavior. In the context of malicious communication behavior, attackers may concentrate attacks from one or a few IP addresses. For instance, a botnet might infect a large number of hosts with bot programs, forming a one-to-many controllable network between the controller and infected hosts. Such a network often utilizes a group of hosts to launch denial-of-service attacks against a specific target. In this scenario, the feature related to port count tends to remain relatively low and fixed. Furthermore, to consume the resources of the target under attack, the total number of transmitted bytes tends to increase.

Based on the summarized observations, this study constructs the feature space for device interaction behaviors as shown in Table 1. This research intends to represent device interaction behaviors without relying on any biased features. As a result, features such as packet length, packet count, session interval, information entropy, and packet skewness are selected to characterize the stability and regularity of the interaction process.

The standard deviation, denoted by "σ", is the arithmetic square root of the variance. It reflects the dispersion of a dataset and is defined as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2}$$

In formula (2), N represents the length of the dataset, μ represents the mean of the dataset, and x_i represents any random variable within the dataset.

The Packet Inclination Rate (PIR) is defined as the ratio of the number of packets exchanged to the average packet length during a certain time window in the interaction process between two devices. Its calculation formula is shown in formula (3), where n represents the number of interactions within the time window, FP_i stands for the number of packets in the i - th interaction session ($i = 1,2,3,\ldots,n$), and FB_i stands for the number of bytes in the i - th interaction session ($i = 1,2,3,\ldots,n$):

$$\mathrm{PIR} = \frac{\left(\sum_{i=1}^{n} FP_i\right)^2}{\sum_{i=1}^{n} FB_i} \tag{3}$$

When devices extensively utilize small packets for network communication, the packet skewness exhibits higher values.

5 Detection Model

The IoT anomaly traffic detection model based on entity interaction protocol consists of three essential components: traffic collection, feature extraction, and anomaly detection. This section provides a comprehensive overview of each key step.

the inherent structural characteristics of networks, while time captures the trends and patterns of network dynamic evolution. Therefore, this study extracts network flow features from both temporal and spatial perspectives to characterize interaction behaviors.

Temporal features of traffic mainly refer to the temporal characteristics of traffic, including time series and time distribution features. The most fundamental attribute within the temporal domain is the time series, which describes the flow unit at different levels (host, session, flow, packet layers).

As temporal flow statistics are based on the relationship between the current connection and statistics within a certain time range, spatial features of flow are needed as a complement. Spatial features of flow primarily encompass attributes such as packet size and packet count. According to the division of network traffic levels, spatial features are analyzed and extracted at the packet, unidirectional flow, session level, and host level.

(1) Packet-level Spatial Features

Studying extractable spatial features at the packet level includes considerations such as packet length. Packet length refers to the size of an individual data packet, effectively capturing the size attribute of network behavior at the packet level. In certain attacks, variations in packet length may exhibit consistent and relatively fixed patterns. In comparison to the more random nature of legitimate communication behavior, these variations often possess discernible distinctions.

(2) Flow-level Spatial Features

Investigating extractable spatial features at the flow level encompasses factors such as packet count and flow byte count. Packet count refers to the quantity of data packets within a single flow, while flow byte count quantifies the size of the entire flow in terms of bytes.

In certain attack scenarios, attackers may frequently initiate communication within the scope of a single flow to achieve their objectives. Legitimate communication behavior, on the other hand, typically displays temporal randomness within a single flow, leading to the capability of differentiating between normal and malicious communication behavior using features related to packet count and flow byte count.

(3) Session-level Spatial Features

Examining session-level considerations reveals primary features such as flow count and session byte count. Similar to flow-level features, flow count and session byte count respectively reflect the characteristics of session quantity and size. In some attack scenarios, attackers might exploit specific services or a combination thereof to execute their attacks. For instance, in attacks like SYN-FLOOD, attackers often exhibit features characterized by a large quantity of the same type of port usage, indicative of an attempt to rapidly deplete server resources.

(4) Host-Level Spatial Features

From the perspective of host-level analysis, significant features that can be extracted include the number of ports and the total number of transmitted bytes. The total number of transmitted bytes can indicate the amount of communication exchange resources

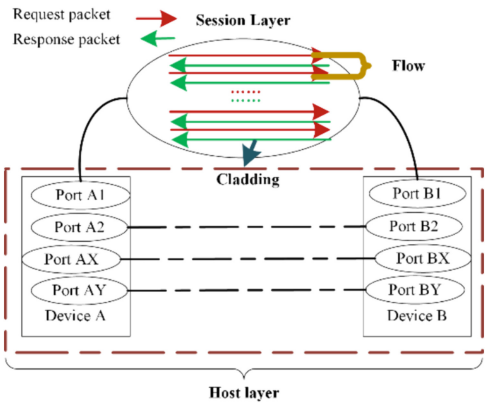


Fig. 2. A hierarchical description of device interaction behavior.

Definition 1. Host Level: In a network, a host is usually identified by one or more IP addresses as its communication address within the network. For analytical and definitional purposes, this study assumes that in the context of the IoT, one IP corresponds to one actual device node. All communication between two IP addresses, i.e., data packets with identical or reverse-source and destination IP addresses, is considered communication data between two devices, referred to as host-level data.

Definition 2. Session Level: Traffic data generated between devices comprises interactions involving multiple applications or services. Different applications or services are represented by distinct port numbers and application protocols in data packets. In this study, a session is formed by aggregating all data packets with the same five-tuple (source IP, source port, destination IP, destination port, and transport layer protocol) or reverse identical five-tuple (source IP, source port, destination IP, and destination port interchanged) within the same session.

Definition 3. Flow Layer: A session consists of one or multiple data streams. Based on the division rules of data streams, within a session, the temporal distribution of data packets in the same direction is considered. Temporal sequence features are essentially consistent when extracted across different hierarchical dimensions, as expressed in Eq. (1).

$$f_{interval} = \{t_0, t_1, t_2, \ldots, t_n, \}, f_{duration} = \{t_0, t_1, t_2, \ldots, t_n, \} \tag{1}$$

where $f_{interval}$ represents the time interval feature set, composed of the initial time representations of the smallest analytical units at this layer, $f_{duration}$ represents the duration feature set (which is absent at the packet level as the smallest analytical unit is the data packet). Additionally, packets with a time difference not exceeding the threshold both before and after, and not exceeding the flow aging time since the first packet in the stream, are aggregated into unidirectional flows.

Definition 4. Packet Level: Data packets constitute the smallest unit of traffic analysis. Every phenomenon is constructed upon the framework of time and space. Space captures

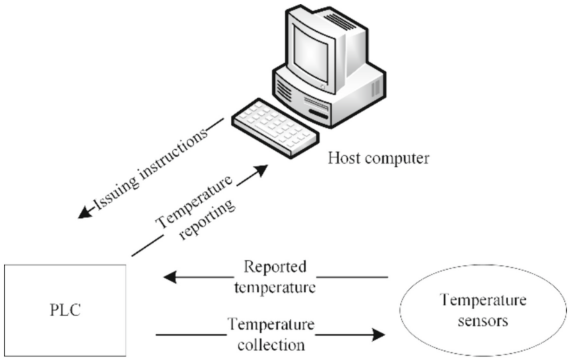


Fig. 1. The example of interactive behaviors of IoT devices.

As illustrated in Fig. 1, taking a temperature sensor as an example, its factory-set function is to gather data on the ambient temperature. Within the context of the IoT, the upper-level computer issues commands to a controlling Programmable Logic Controller (PLC) to schedule periodic temperature measurements and reporting by the temperature sensor. The PLC then forwards the timed measurement instruction to the temperature sensor, which subsequently collects data and reports it. Under normal circumstances, this functional logic is executed repetitively and periodically in a stable manner, assuming no interference or attack. Therefore, by thoroughly studying and modeling all data in this interaction process, any anomalies that arise can be promptly detected.

Existing research has rarely addressed the unique interaction behaviors within network communication. This is primarily because human-initiated interactions are usually driven by human consciousness, displaying attributes of variability, complexity, and subjectivity. However, in practical IoT environments, interaction processes among entities possess certain complexity. Yet, these processes are coordinated to achieve and maintain the dynamic stability of the network environment. Thus, representing entity interaction processes is feasible. In light of this, by comprehensively learning the baseline of interaction behaviors among IoT devices, the detection of any anomalies can be timely, comprehensive, and accurate.

4.2 Device Interaction Behavior Abstraction

According to the analysis in Sect. 4.1, the comprehensiveness and accuracy of the description of interaction behaviors between device nodes directly impact the accuracy of anomaly detection. To comprehensively describe interaction behaviors, this study proposes a method for characterizing device interaction behaviors based on hierarchical flow feature extraction. This method divides the interaction behavior between two device nodes into different hierarchical levels, as shown in Fig. 2, and extracts features from both the temporal and spatial dimensions to achieve a comprehensive representation of the interaction behavior.

The network traffic in the IoT contains interaction data between different nodes, and these interactions involve three levels: session, flow, and packet. Therefore, in this study, we first define the four delineated levels as follows:

achieves the dual objectives of preserving privacy and enhancing anomaly detection precision. Furthermore, Liu et al. [10] develop a convolutional neural network model with long short-term memory to enhance detection accuracy. They also employ gradient compression in FL to reduce communication costs and improve communication quality. Addressing issues in the context of the IoT, where traditional anomaly detection methods suffer from low detection accuracy due to imbalanced massive data and poor model generalization caused by data heterogeneity, Thing et al. [11] utilize a neural network composed of multi-layer sparse autoencoders and stack autoencoders to detect various types of attacks within the network. Doshi et al. [12] leverage IoT-specific network behaviors as features and employ neural networks and ML techniques to achieve high-precision detection of DDoS attacks in IoT communication. Considering the long-term dependencies in streaming data, Cheng et al. [13] introduce a semi-supervised hierarchical stacked temporal convolutional network. This network design fully embraces the characteristics of streaming data in IoT while eliminating uncertain records.

As the IoT continues to evolve, based on a survey of recent research findings [14, 15], it can be observed that ML-based anomaly detection solutions primarily focus on feature selection, model optimization, and improving model generalization capabilities. However, inherent challenges such as strong data dependencies and the inability to adapt to unknown attacks persist within ML approaches. While the integration of deep learning partially mitigates these issues, a fundamental resolution remains elusive.

Analyzing the characteristics of network attacks themselves, these attacks aim to compromise the confidentiality, integrity, and availability of system information. They typically induce deviations from normal network operations, manifesting as abnormal behaviors. As a result, identifying anomalies can be achieved through the discovery of patterns in data that deviate from expected behaviors. System invariance refers to a condition within the "physical" or "chemical" characteristics of a system's operational process, which must hold whenever the system is in a given state. Analyzing physical invariances for anomaly detection has been applied in numerous cyber-physical systems involving networked information and physical components. The underlying processes of the IoT are governed by their operational principles, with inter-device process states being predictable and foreseeable. Hence, the assimilation of the inherent and objective attributes of the IoT, coupled with their modeling, serves as a robust approach for achieving anomaly detection. This approach is versatile and well-suited for detecting novel and emerging types of attacks.

4 Device Interaction Behavior Representation

4.1 Device Interaction Behavior Abstraction

The fundamental building blocks of the IoT comprise a diverse array of sensor devices. The functionality of these devices is preconfigured during their manufacturing, resulting in a set of fixed and limited executable actions. By comprehensively learning the finite operational behaviors among these devices, precise and efficient anomaly detection can be achieved.

Just as network attackers employ network reconnaissance tools for scanning and gathering information about hosts, subnets, ports, and protocols within a network, analogous tools targeting IoT devices now exist. These tools are capable of probing and scanning IoT devices, revealing pertinent information about them. Presently, a substantial portion of IoT devices available in the market utilizes wireless communication protocols, such as ZigBee, ZWave, Bluetooth-LE, and Wi-Fi 802.11. Unfortunately, these protocols are susceptible to wireless surveillance and probing attacks, akin to their network counterparts.

Malware infections, denial of service attacks, unauthorized access, and the injection of falsified data packets are all common attack methods within the realm of the IoT. To safeguard IoT systems from network attacks, a myriad of security measures have emerged, including encrypted communication data, data integrity validation, and access control techniques. These methods serve to shield systems from a variety of attack types. However, even with these security measures in place, attackers can still successfully target systems, employing tactics like malicious packet injection and DDoS attacks. Therefore, the implementation of network anomaly detection becomes crucial to further enhance the security of the IoT.

3 Current State of Research on Anomaly Detection in the Internet of Things

Due to the extensive access of IoT devices, network traffic is experiencing exponential growth. The shared data between devices and on the network becomes more susceptible to cyberattacks. To mitigate the emerging network attacks on IoT devices, greater efforts are required to research anomaly detection [2, 17]. Machine learning (ML)-based detection methods have gained popularity in recent years as a prevalent approach for anomaly detection in the IoT [3, 4]. Employing machine learning techniques, these methods autonomously learn patterns and characteristics from IoT data, enabling independent anomaly detection.

Qiu et al. [5] focus on the deception problem in mobile social networks and propose an adaptive social spammer detection model that improves the security of social users. Dutta et al. [6] propose an integrated approach that leverages deep models and stacking techniques, utilizing a heterogeneous flow-based dataset containing IoT data to achieve reliable anomaly classification. In response to the lower performance and predictive accuracy of intrusion detection systems based on anomaly-driven ML in IoT intrusion detection, Abdelmoumin et al. [7] introduce a method for optimizing models using hyperparameter tuning and ensemble learning to enhance IoT security. Nie et al. [16] focus on the safety and malicious attack issues related to connected vehicle networks. They have designed a data-driven intrusion detection system and a deep learning architecture based on convolutional neural networks, aiming to detect intrusions targeting roadside units. As a novel distributed learning paradigm, Federated Learning (FL) facilitates decentralized training of predictive models through collaborative means [8]. In order to ensure the efficiency, accuracy, and effectiveness of anomaly detection in the context of industrial IoT, Wang et al. [9] introduce an architecture for anomaly detection and an empowered approach utilizing federated deep reinforcement learning. This approach

Due to the aforementioned factors, sensor devices are susceptible to manipulation and destruction by malicious actors, making them vulnerable to security threats such as data breaches and zombie networks. Physical attacks involve tampering with sensor devices to achieve malicious tracking and data acquisition objectives. Attackers dismantle the physical casings of devices like hosts or embedded systems, subjecting them to physical dissection and analysis. This enables the extraction of sensitive components like processors and memory, thus obtaining critical sensitive parameters including key information, passwords, and configuration details.

2.2 Authentication Attacks

Attackers can exploit the default credentials of physical devices in the IoT to initiate attacks and gain unauthorized access to data. Devices with weak authentication processes or those sharing identical authentication information are susceptible to authentication-based attacks.

2.3 Communications Protocol Attacks

The application of multiple communication protocols has indeed introduced threats to the security landscape of the IoT. On one hand, the dynamic nature of network structures presents security vulnerabilities. Due to the ever-evolving nature of IoT network architectures and the emergence of various new network protocols, vulnerabilities within these new protocols and the ongoing upgrade and update of network devices may potentially introduce novel security threats. On the other hand, the convergence of diverse networks generates security risks. The IoT is a confluence of multiple networks that support the IP protocol, each employing distinct security strategies. This convergence process can give rise to fresh security risks and vulnerabilities. In many instances, communication protocols might have introduced vulnerabilities during their initial design or subsequent implementation and configuration phases, thereby rendering the IoT susceptible to security protocol attacks at the transport layer.

2.4 Wireless Detection

In the realm of the IoT, wireless communication is a prevalent means of data transmission. However, wireless signals that are exposed can be easily disrupted and intercepted by malicious actors. This susceptibility results in severe consequences like the paralysis of wireless communication networks, the theft of sensitive user information, and the fabrication of data. Wireless probing predominantly occurs within the sensing layer of the IoT, giving rise to security threats that encompass two primary facets. Firstly, the pilfering of information from sensing nodes is a considerable concern. These nodes typically serve as basic information repositories with limited computational and processing capabilities. Unauthorized users can effortlessly access relevant data stored within these nodes. Secondly, the impersonation of sensing nodes poses a critical challenge. Attackers pilfer label information from these nodes, subsequently replicating or altering these labels. By impersonating the identity of the nodes, attackers gain access to valuable information, undermining the credibility and efficacy of the compromised nodes.

Network traffic encompasses all data communication of IoT devices, and the prevalent approach in IoT environments for safeguarding involves the sidestream collection of network traffic for data analysis and anomaly detection. Leveraging the inherent characteristics of IoT, network communication between IoT devices abides by established objective constraints. Under normal operation, devices tend to adhere to predefined behaviors in a repetitive manner. In light of this, this study introduces an IoT anomaly traffic detection method focused on device interaction behaviors, enabling precise and timely anomaly detection through a comprehensive depiction of interaction behaviors between devices. Specifically, this research contributes in two key aspects:

1. A device interaction behavior representation method based on hierarchical flow features is proposed. This study comprehensively characterizes and portrays the interaction behaviors between IoT device nodes at four levels: packet, flow, session, and host. This realization of thorough interaction behavioral representation establishes a benchmark feature space for IoT anomaly detection.
2. A novel, universal, accurate, and efficient anomaly detection method is innovatively presented from the perspective of device interactions. Escaping the constraints of traditional anomaly detection methods tailored to hardware or specific applications, and addressing the gradual drop in detection rate, this approach capitalizes on IoT's intrinsic attributes. It introduces an anomaly detection scheme centered on the constraints adhered to by device interactions. The propose scheme is experimentally evaluated using the BoT-IoT dataset [1], with an average detection rate of 98.4% and a false positive rate of 1.3%, directly validating the feasibility and superiority of the approach.

2 Security Threats Faced by the Internet of Things

In the course of interactions, the IoT inevitably gives rise to information security issues, encompassing physical security, operational security, and data security. This section provides an analysis of the primary security threats confronting IoT.

2.1 Physical Attacks

An investigation conducted by the MPI Group in 2017 revealed that only 47% of IoT manufacturers consider security concerns during the conceptual or design stages. Furthermore, 21% of manufacturers only begin to contemplate security during the production phase, while 18% address security issues only at the quality management stage. Shockingly, the remaining manufacturers never take security into account. The proliferation of zombie networks like Mirai can be attributed to the fact that many IoT devices lack even the most rudimentary security measures. IoT is extensively employed to replace human intervention in complex, hazardous, or mechanical tasks, with sensor devices predominantly operating without human supervision. Moreover, the majority of these devices possess simplified functionality, and the diverse applications of sensor devices employ distinct standards and protocols, precluding the adoption of unified security measures against external attacks.

1 Introduction

Since the International Telecommunication Union formally put forward the concept of the Internet of Things (IoT) in 2005, sensor networks, cloud computing, microchips, and other technologies have been developing and maturing, and the IoT industry has been rapidly developing and expanding. Over the past five years, the number of IoT devices has experienced explosive growth. According to data released by authoritative statistical organizations, the global count of IoT devices connected to networks reached 2.035 billion in 2017 and is projected to exceed 7.544 billion by 2025. IoT is poised to profoundly impact various aspects of human production and daily life. Simultaneously, the rapid advancement of network technologies is poised to drive the comprehensive realization of the IoT era. Taking 6G mobile communication as an example, the 6G mobile communication network not only connects people but also links computing resources, vehicles, devices, sensors, and robotic agents to fulfill the requirements of a fully interconnected, intelligent digital world. In the 6G mobile communication system, the widespread deployment of IoT leads to a rapid increase in network access points. According to Ericsson's predictions, by the year 2025, over 24.9 billion devices will be connected to networks.

Amidst the thriving development of the IoT, existing security mechanisms have struggled to meet the escalating security demands, resulting in a proliferation of security concerns across diverse application scenarios. The susceptibility of numerous devices to malicious code threats or unauthorized control has given rise to an array of security issues, and in some instances, triggered large-scale security incidents. A notable case occurred in 2016 when the infamous Mirai worm exploited IoT devices, causing a massive Distributed Denial of Service (DDoS) attack that led to widespread internet disruption on the U.S. East Coast. Prominent websites experienced service outages. More recently, smart speakers have been exploited by attackers for eavesdropping on user privacy, underscoring how IoT security threats evolve in tandem with technological advancements. Timely threat detection and proactive defense are pivotal strategies in countering such threats. Anomaly detection plays a crucial role in mitigating malicious activities within defense systems and networks.

Over the past years, research into IoT anomaly detection has surged to combat network attacks, resulting in the proposal of numerous detection mechanisms. Nonetheless, the distinctive attributes of IoT systems present substantial challenges to implementing comprehensive security measures. For instance, a lack of unified standards in IoT platform design, development, communication interaction, and access control, coupled with inadequately protected internal and external operational environments, hamper effective security. Existing solutions exhibit limitations such as narrow applicability and insufficient automation. Additionally, many manufacturers are inclined to believe that augmenting security measures won't enhance a device's market value but rather escalate production costs. Consequently, post-sales support, patch provisioning, and updates are often overlooked, leading to a proliferation of vulnerable devices with high-risk vulnerabilities like default credentials and plaintext transmission of keys. Consequently, a holistic and generalized anomaly detection mechanism, tailored to the distinct IoT network environment, is paramount for bolstering IoT security.

An Abnormal Detection Method Based on the Device Interaction Behavior in the Internet of Things

Wenjing Jin[1], Xiaofei Cui[1], Chengsheng Zhou[1(✉)], Hanxue Li[2], and Jianbo Zheng[3,4(✉)]

[1] China Academy of Information and Communications Technology (CAICT), Beijing 100089, People's Republic of China
zhouchengsheng@caict.ac.cn

[2] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
jianbo.zheng@smbu.edu.cn

[4] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

Abstract. With the development of smart homes, digital medicine, the Internet of vehicles, and other technologies, the application of the Internet of Things (IoT) is becoming more and more popular, and its security issues have attracted more and more attention from researchers. Anomaly detection schemes based on traffic can find anomalies at different levels by external means, which is a key part of the security protection of the IoT. However, existing researchers are faced with the problems of insufficient generality and strong method limitations. In view of this, based on the stability and constraint reflected by the physical constraints followed by the operation of the IoT system and the domain specification on the device interaction behavior, this study proposes a hierarchical traffic characteristic based on the integration of spatiotemporal characteristics of different levels such as packet, stream, session, host, etc. Secondly, based on the complete interaction behavior feature space, an integrated anomaly detection model is established by learning the interaction behaviors of different device pairs to realize accurate and efficient security event discovery. Finally, the propose method is evaluated on a BoT-IoT dataset. Ten-fold cross-check and the detection accuracy under different attack traffic and normal traffic ratio show the feasibility and superiority of the propose method.

Keywords: Internet of Things · Anomaly Detection · Device Interaction Behaviors · Machine Learning

This work was supported by Ministry of Industry and Information Technology Industrial Internet Innovation and Development Project-Internet of Things Ba- sic Security Access Monitoring Platform Project (TC210H023), and also supported by the Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 302–319, 2024.
https://doi.org/10.1007/978-3-031-65126-7_27

8. Ji, L., Guo, S.: Energy-efficient cooperative resource allocation in wireless powered mobile edge computing. IEEE Internet Things J. **6**(3), 4744–4754 (2018)
9. Chen, H., Xiao, L., Yang, D., Zhang, T., Cuthbert, L.: User cooperation in wireless powered communication networks with a pricing mechanism. IEEE Access **5**, 16895–16903 (2017)
10. Li, B., Si, F., Zhao, W., Zhang, H.: Wireless powered mobile edge computing with NOMA and user cooperation. IEEE Trans. Veh. Technol. **70**(2), 1957–1961 (2021)
11. Mao, S., Jinsong, W., Liu, L., Lan, D., Taherkordi, A.: Energy-efficient cooperative communication and computation for wireless powered mobile-edge computing. IEEE Syst. J. **16**(1), 287–298 (2020)
12. Wang, L., Shao, H., Li, J., Wen, X., Zhaoming, L.: Optimal multi-user computation offloading strategy for wireless powered sensor networks. IEEE Access **8**, 35150–35160 (2020)
13. Wang, R., Chen, J., He, B., Lv, L., Zhou, Y., Yang, L.: Energy consumption minimization for wireless powered NOMA-MEC with user cooperation. In: 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–5. IEEE (2021)
14. Wang, X., Ning, Z., Guo, L., Guo, S., Gao, X., Wang, G.: Online learning for distributed computation offloading in wireless powered mobile edge computing networks. IEEE Trans. Parallel Distrib. Syst. **33**(8), 1841–1855 (2021)
15. Wang, X., Wang, S., Wang, Y., Ning, Z., Guo, L.: Distributed task scheduling for wireless powered mobile edge computing: a federated-learning-enabled framework. IEEE Network **35**(6), 27–33 (2021)
16. Wang, X., et al.: Wireless powered mobile edge computing networks: a survey. ACM Comput. Surv. **55**(13), 1–37 (2023)
17. Ning, Z., Yang, Y., Wang, X., Song, Q., Guo, L., Jamalipour, A.: Multi-agent deep reinforcement learning based UAV trajectory optimization for differentiated services. IEEE Trans. Mob. Comput. (2023)

to the other three algorithms. In addition, as the number of terminals in the system increases, the performance of the DBO algorithm becomes increasingly weak. Although the performance of the AWPSO algorithm has caught up with the increase in the number of terminals, overall it is still not as good as the FHO algorithm, and the performance of the FHO algorithm is only second to the Q-IADE algorithm.

7 Conclusion

In this paper, we design a WPT-MEC system model with multiple base stations and terminals. In this model, terminals collect energy for task computation, and we use the FDMA technology to achieve simultaneous task offloading for multiple terminals [16,17]. We propose a relay collaborative offloading scheme to overcome the double near-far effect. Our goal is to minimize the weighted sum of normalized system delay and normalized energy consumption rate by jointly optimizing variables such as wireless energy transmission time, offloading time, computing offloading and local computing task allocation, and offloading target selection. In order to improve the solving efficiency, we have improved the immune algorithm and proposed the Q-IADE algorithm. The simulation results show that the performance of our proposed Q-IADE algorithm is superior to other algorithms, and the convergence effect is also more stable. In the future, we will use existing wireless charging devices in actual scenarios to verify the feasibility of our proposed solution and the effectiveness of the Q-IADE algorithm.

References

1. Mao, Y., Zhang, J., Letaief, K.B.: Dynamic computation offloading for mobile-edge computing with energy harvesting devices. IEEE J. Sel. Areas Commun. **34**, 3590–3605 (2016)
2. Ning, Z., et al.: Mobile edge computing and machine learning in the internet of unmanned aerial vehicles: a survey. ACM Comput. Surv. **56**(1), 1–31 (2023)
3. Wang, X., Ning, Z., Guo, L., Guo, S., Gao, X., Wang, G.: Mean-field learning for edge computing in mobile blockchain networks. IEEE Trans. Mob. Comput. **22**(10), 5978–5994 (2022)
4. Ning, Z., et al.: Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE Trans. Mob. Comput. **22**(5), 2628–2644 (2021)
5. Zhou, F., Hu, R.Q.: Computation efficiency maximization in wireless-powered mobile edge computing networks. IEEE Trans. Wireless Commun. **19**(5), 3170–3184 (2020)
6. Feng, J., Pei, Q., Yu, F.R., Chu, X., Shang, B.: Computation offloading and resource allocation for wireless powered mobile edge computing with latency constraint. IEEE Wireless Commun. Lett. **8**(5), 1320–1323 (2019)
7. Xiaoyan, H., Wong, K.-K., Yang, K.: Wireless powered cooperation-assisted mobile edge computing. IEEE Trans. Wireless Commun. **17**(4), 2375–2388 (2018)

Figure 3 shows the comparison of the effectiveness of each improvement of IA when the number of base stations is $N = 1$ and the number of terminals is $M = 10$. From the graph, it can be seen that the first improved IA can search for higher quality feasible solutions. After our second improvement on IA, the objective function values further converge and become better.

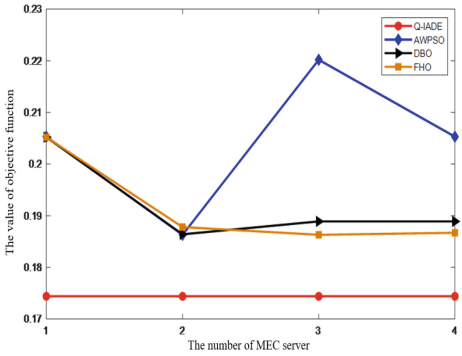


Fig. 4. M=10, the objective function value changes with the number of servers.

Figure 4 shows the comparison of the objective function value and the optimization effect of the number of base stations for four algorithms when the number of terminals is M=10. From the figure, it can be seen that the performance of the Q-IADE algorithm is very stable, which is largely due to the fixed configuration of the terminals in our experiment. It can also be seen that the AWPSO algorithm is the most unstable and prone to falling into local optima.

Figure 5 shows the comparison of the optimization effects of four algorithms on the objective function value and the number of terminals when the number of base stations is $N = 4$. From the graph, it can be seen that our proposed Q-IADE algorithm has the most stable output and better performance compared

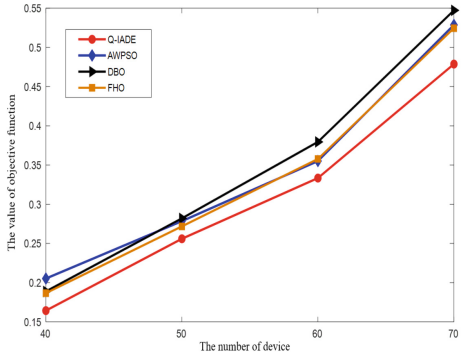


Fig. 5. $M = 10$, the objective function value changes with the number of servers.

Algorithm 1. The procedure of Q-IADE.

Input: t, D
Output: the value v of the objective function
  while the number of training $<N_1+1$
    randomly generate state S
    for each action A in $A(S)$
      Calculate rewards R and generate Q tables
    end for
    select the offload-object U_i of individual i
  end while
  return offloading policy U
  while iteration $<N_2+1$
    randomly generate population P_0 under the condition of U
    calculate the fitness and the similarity between the solution and the solution, and order P_0
    for each individual j from 0 to $0.25P_0$
      variation,cross,select
    end for
    for each individual j from $0.25P_0$ to $0.5P_0$
      immune manipulation
    end for
    for each individual j from $0.5P_0$ to P_0
      flash P_0
    end for
    calculate the fitness and the similarity between the solution and the solution, and order P_0
  end while
  return the value v of the objective function

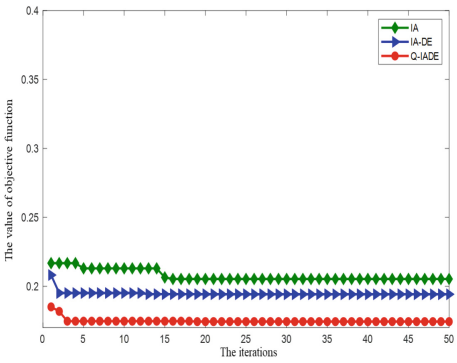


Fig. 3. Comparison of Algorithm Improvement Effects.

Beetle Optimizer (DBO) algorithm and Fire Hawk Optimizer (FHO) algorithm, and compare and simulate to prove the superiority of the proposed algorithm. All emulators are written in the MATLAB programming language.

5 The Q-IADE-Based Resource Scheduling Algorithm

Immune algorithm(IA) is a heuristic algorithm designed by scholars inspired by the biological immune system. This paper improves and enhances it. The IA combines the concept and theory of immunity with genetic algorithm, and adds antibody concentration evaluation operators and incentive degree calculation operators to maintain the diversity of individual populations, avoiding the "premature" problem in the general optimization process.

5.1 Searching Ability Improvement

It is difficult for IA to achieve a global optimal position, and as the population iterates, the convergence speed of the algorithm and the accuracy of feasible solutions will also decrease.

The differential evolution(DE) algorithm has the characteristics of strong robustness and fast convergence speed. Therefore, this article takes the mutation, crossover, and selection operators in DE as part of the IA iteration process, allowing them to participate in the antibody cloning operator of IA, thereby enhancing the local search ability of the original IA.

5.2 Relay Selection Optimization

The model in this paper has a relay offloading scheme, which corresponds to the selection of the optimal offloading path. The optimal path selection involves dynamic planning, and although heuristic algorithms can be used to solve it, its actual effect is poor.

In fact, reinforcement learning is developed from dynamic programming, and the most important thing is that dynamic programming is best at dealing with dynamic optimization problems. The Q-learning algorithm is a model-independent reinforcement learning algorithm, which is theoretically supported by Markov's decision-making process. Therefore, this paper intends to use Q-learning to deal with the problem of how to select relay objects, and help us choose the optimal offloading strategy in a given environment.

Compared with heuristic algorithms, reinforcement learning can give full play to the role of information in historical samples. Finally, the pseudocode of the Q-IADE algorithm obtained by improving the IA is shown in Algorithm 1.

In Algorithm 1, t represents the allocation of the time, D represents the location coordinates of the terminals and base stations.

6 Numerical Simulation

In this section, simulation experiments are designed to verify the effectiveness of the proposed algorithm in our model.

In this paper, three newly proposed evolutionary algorithms in the past two years are selected, namely Adaptive Weighting PSO (AWPSO) algorithm, Dung

The energy consumption when local computing can be calculated by:

$$E_{loc,i} = N_{loc,i} * q_i * e_i, \forall i \in \mathbf{M}, \tag{8}$$

where $e_i = k_i f_i^2$ represents the energy consumption generated by the CPU of terminal i, k_i represents the effective capacitance coefficient of terminal i.

Similarly, the execution of local calculations by terminal i is limited by energy consumption: $E_{loc,i} \leq E_{res,i} + E_{har,i}, \forall i \in \mathbf{M}$.

Based on the above computing offloading and local calculation processes, there are constraints: $E_{off,i} + E_{loc,i} \leq E_{res,i} + E_{har,i}, \forall i \in \mathbf{M}$.

4 Problem Formulation

The problem studied in this paper is to minimize the weighted sum of the normalized system delay and the normalized system energy consumption rate.

Therefore, the problem can be expressed as a formula:

$$\min_{\forall i \in \mathbf{M}} \beta * \frac{t_h + \max(t_i, t_{loc,i})}{T} + (1-\beta) * \frac{\sum(E_{off,i} + E_{loc,i})}{\sum(E_{res,i} + E_{har,i})} \tag{9}$$

s.t.

$$\begin{aligned}
&C1: 0 < \eta_i \leq 1, \forall i \in \mathbf{M} \\
&C2: E_{har,i} \leq C_{max,i} - E_{res,i}, \forall i \in \mathbf{M} \\
&C3: \sum_{i=1}^{M} B_i \leq B_{max}, \forall i \in \mathbf{M} \\
&C4: 0 \leq N_{off,i} \leq N_i, \forall i \in \mathbf{M} \\
&C5: t_h + t_i \leq T, \forall i \in \mathbf{M} \\
&C6: \sum_{i=1}^{M} N_{off,i} * q_i \leq Q, \forall i \in \mathbf{M} \\
&C7: t_h + t_{loc,i} \leq T, \forall i \in \mathbf{M} \\
&C8: f_i \leq f_{max,i}, \forall i \in \mathbf{M} \\
&C9: E_{off,i} + E_{loc,i} \leq E_{res,i} + E_{har,i}, \forall i \in \mathbf{M}
\end{aligned} \tag{10}$$

In (9), the β represents the weight of the normalized system delay, and its value is 1 with the weight of the normalized system energy consumption rate.

Due to the close coupling relationship between the variables involved in the problem studied in this paper, the problem studied in this paper has become a mixed integer nonlinear programming problem. Obviously, the problem being studied involves multivariate combinatorial optimization, which is limited to a certain range of values, making it a NP-hard problem. To solve this problem, the following algorithm has been designed.

According to Shannon's formula, the offloading rate of terminal i task is

$$R_i = B_i \log_2 \left(1 + \frac{h_{i,j} a P_i}{\sigma_0^2}\right), \forall i \in \mathbf{M}, \tag{2}$$

where B_i represents the bandwidth occupied by terminal i, as we adopt the FDMA scheme, which affects the total bandwidth B_{max} has constraints: $\sum_{i=1}^{M} B_i \leq B_{max}, \forall i \in \mathbf{M}$, $h_{i,j}$ represents the channel gain that terminal i chooses to offload to the server or other terminal j; P_i indicates the transmission power of terminal i chooses to offload; a is a constant used to constrain the unloading power:

$$a = \begin{cases} 1, \forall J \in \mathbf{N} = \{1, 2, ..., N\}, \\ 0.6, \forall J \in \mathbf{M}, \end{cases} \tag{3}$$

σ_0^2 represents the additive white Gaussian noise power near the offloaded data receiver.

Assuming the total task volume of terminal i is $N_i \geq 0$ bits, the task of $N_{off,i}$ bits needs to be offloaded to edge servers or other terminals, so there are $0 \leq N_{off,i} \leq N_i, \forall i \in \mathbf{M}$,

$$N_{off,i} = R_i * t_i, \forall i \in \mathbf{M} \tag{4}$$

The energy consumption of terminal i for offloading task data is

$$E_{off,i} = aP_i * t_i + P_c * t_i, \forall i \in \mathbf{M} \tag{5}$$

wherein P_c is the constant circuit power consumption of the terminal. $E_{off,i}$ is constrained by the actual remaining energy in the terminal i battery at this time: $E_{off,i} \leq E_{res,i} + E_{har,i}, \forall i \in \mathbf{M}$.

This paper uses q_i represents the CPU revolutions required for terminal i to calculate 1 bit of data. To ensure that the delay in result return can be ignored, assuming there are limitations: $\sum_{i=1}^{M} N_{off,i} * q_i \leq Q, \forall i \in \mathbf{M}$, where Q represents the computing power that the CPU of the edge server.

3.2.2 Local Computation Model

After terminal i offloaded $N_{off,i}$ bits, perform local calculations on the remaining bits:

$$N_{loc,i} = N_i - N_{off,i}, \forall i \in \mathbf{M}, \tag{6}$$

where $N_{loc,i}$ represents the amount of task data for local computation.

Thus, we can calculate the time that terminal i is used for local calculation:

$$t_{loc,i} = \frac{N_{off,i} * q_i}{f_i}, \forall i \in \mathbf{M}, \tag{7}$$

where f_i represents the CPU frequency of terminal i, which cannot exceed the maximum frequency limitations on $f_{max,i}$.

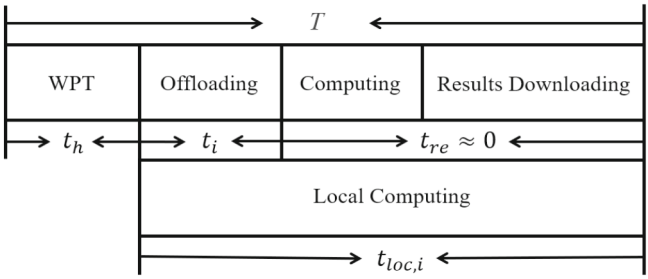


Fig. 2. Time slots allocation within the time block T.

each terminal. This paper adopts a block fading channel model, assuming that the channel state remains unchanged within a time block T, but can change between different time blocks T. In addition, the energy collected by the terminal and the downlink channel gain between the base station and the terminal are both h_i. And the channels have reciprocity. The block fading channel model in the system is $h = 10^{-3}d^{-\alpha}\varphi$, where φ represents short-term fading, where d is the distance, α is the path loss index.

Next, we will continue to introduce the two phases involved in the system model in this paper.

3.1 Energy Harvesting Phase

At the time slot t_h, the energy transmitting station wirelessly charges the terminal. The energy collected by terminal i is

$$E_{har,i} = \eta_i P_0 h_{edo,i} t_h, \forall i \in \mathbf{M}, \tag{1}$$

where η_i represents the energy conversion efficiency of terminal i, and satisfies $0 < \eta_i \leq 1, \forall i \in \mathbf{M}$, P_0 represents the transmission power of the energy transmission station; $h_{edo,i}$ represents the downlink channel gain from the energy transmitting station to terminal i.

Due to the battery capacity $C_{max,i}$ of terminal i is limited, therefore $E_{har,i}$ needs to meet: $E_{har,i} \leq C_{max,i} - E_{res,i}, \forall i \in \mathbf{M}$, where $E_{res,i}$ represents the remaining energy in the terminal i battery before charging.

3.2 Task Data Processing Phase

In this stage, the terminal performs task processing, including two parts: task data offloading and local computing.

3.2.1 Offloading Model

This paper assumes that the task data is bit by bit independent and considers partial offloading. To overcome the dual near far effect, we propose a relay offloading transmission scheme.

results strongly depend on the initial values. However, heuristic algorithms do not require the mathematical properties of solving the target problem.

3 System Model

The system model of this paper is shown in Fig. 1. The system model is mainly composed of a single energy transmitting station, N base stations with integrated mobile edge computing servers and M low-power terminals. The energy station and base stations in the model have stable power supply.

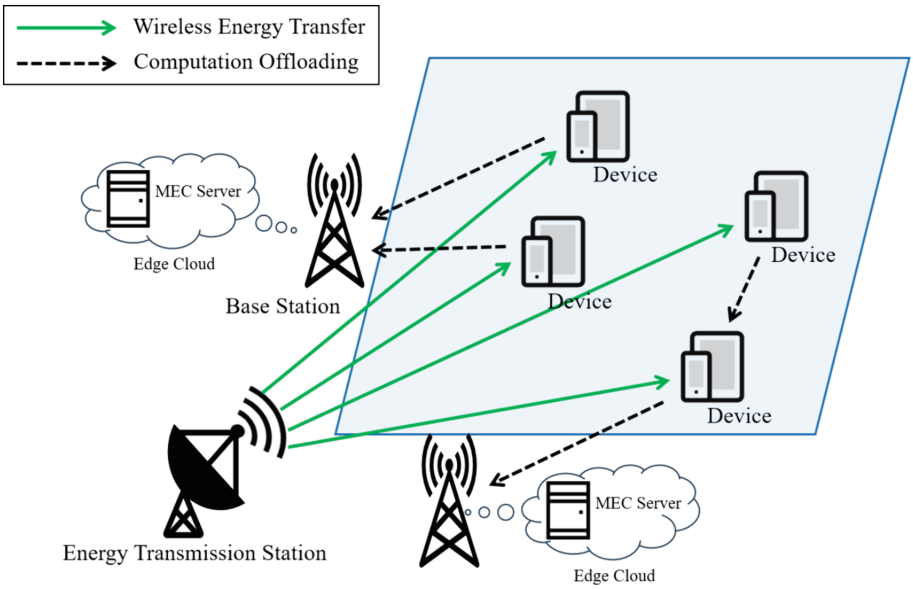


Fig. 1. The overview of system model.

This paper considers the relay collaboration among terminals in the model. This paper applies the Frequency Division Multiple Access (FDMA) technology to the system, that is, the base station can simultaneously receive offloading tasks from M terminals. Figure 2 shows the allocation of time slots. In the first stage, the energy transmitting station wirelessly charges the terminal battery, and this time slot is marked as t_h; In the second stage, the terminal performs task processing. This paper sets the local computation and offloading of the terminal in the model to be possible simultaneously. We mark the time slot for local calculation of terminal i as $t_{loc,i}, i \in \mathbf{M} = 1, 2, ..., M$, and the offloading time slot as t_i. We record the sum of the time between edge server calculations and the return of calculation results as $t_{re} \approx 0$.

This paper assumes that the base station has prior knowledge of channel state information (CSI) between each terminal and the status information of

The main structure of the paper is as follows: Sect. 2 is related work in this area. Section 3 is the model for multi-device relay cooperation WPT-MEC system model. Section 4 is the description for the promotion of Q-IADE. Section 5 gives the simulation experiments. Section 6 is the conclusion of this paper and the future work.

2 Related Works

The literature [6] proposes a multi-user MEC system, there is also a "double near-far effect", which is clearly unfair to distant devices. The literature [7] proposes a WPT-MEC system with only two mobile devices. And experiments have shown that the offloading of remote mobile devices by relay will have lower total emission energy and better performance than APs without cooperative systems. The literature [8] also takes into account the "double near-far effect", but the authors do not take into account the computing power of the mobile device itself, that is, the data tasks can only be completely offloaded. The literature [9] proposes a pricing mechanism based on dual-user collaboration. However, the authors only studied distant devices in the article, simplifying the task situation of nearby devices.

The system models described in the above literature show that the impact of the "double near-far effect" on remote device nodes in the WPT-MEC system cannot be ignored, and there is a lack of research on the existence of multiple device terminals in the WPT-MEC system of inter-user collaboration.

The authors design an iterative algorithm by using the Lagrange dual method [6]. The literature [7] is to obtain the optimal energy transmission power by the bisecting search algorithm. The literature [5] uses the Dinkelsbach method to transform the studied problem into a convex optimization problem. The authors use the classical Lagrange method, Newtonian iterative method and subgradient algorithm to obtain the optimal solution of the convex optimization problem [8]. The literature [10] cleverly uses mathematical methods to directly solve nonconvex problems. The literature [11] uses variable substitution and semi-definite relaxation to transform the original problem into a convex optimization problem, and then obtains the optimal solution by the Lagrange method. The literature [12] designs an unloading algorithm based on the process of merging and splitting. The paper [13] uses the equivalent substitution method to convert the problem into a convex optimization problem, and then use the Lagrange method to find solution. The literature [14] designs a neural model to learn the offloading and time-division decisions of each time slot. The literature [15] designs a framework to support federated learning in the WPT-MEC system.

From the above literature, it can be seen that most of the algorithms used to solve resource scheduling problems in the WPT-MEC model are traditional optimization algorithms, and the application of scheduling algorithms based on reinforcement learning is still limited. However, traditional optimization algorithms based on mathematical theory, which are based on calculus, make it difficult to get started, and their application scale is small, and the solution

ability, which can better solve the problems we raised. Finally, the simulation results show that the proposed Q-IADE algorithm has strong global search ability and stable convergence effect, and the algorithm performance is superior to the other three comparison algorithms, especially when relay offloading can be performed.

Keywords: WPT-MEC · FDMA · Relay collaboration · IA

1 Introduction

In the context of the era of the Internet of Everything, the data types generated by the devices at the edge of the network are varied, and the data is also generated all the time, but the amount of these data is actually much smaller than the amount of tasks that need to be transmitted to the cloud computing in the era of centralized big data processing, and the real-time requirements of the network edge devices for data processing are very high, so the tasks that originally need to be transmitted to the cloud computing are migrated to the edge cloud near the device terminal for processing, which can improve the data transmission performance to ensure the real-time processing.

Because the batteries of edge devices are limited, how to provide sustainable energy supply for edge devices is a major challenge in the era of the Internet of Things. Of course, we can use wired charging for it, but in many cases, wired charging is unlikely or inconvenient [1–4]. In order to overcome this bottleneck of limited battery of network edge devices, the wireless power transfer (WPT) technology is combined with mobile edge computing (MEC) to form a WPT-MEC system.

However, if the energy station and the edge cloud server are configured together, there is a "double near-far effect" in the multi-node WPT-MEC system [5], under the influence of this effect, the terminals close to the energy station will have better channel conditions, which means that the terminals farther away from the energy station collect less energy but consume more energy from/to the edge cloud.

In summary, the WPT-MEC system has received widespread attention from the academic community both domestically and internationally. However, there is currently very little research on the application of intelligent evolutionary algorithms for resource allocation methods in WPT-MEC systems where multiple user devices can collaborate with each other.

The major contributions of this paper are as follow:

1) We propose a WPT-MEC model that takes into account the double near-far effect.
2) We propose a scheme for multi-terminal relay cooperation to overcome the double near-far effect.
3) We select Immune Algorithm(IA) as a resource allocation algorithm for this model and we propose a improved Q-IADE algorithm to improve the performance of IA. In this algorithm, we improve the local search capability and dynamic programming performance of IA.

Joint Delay and Energy Optimization for WPT-MEC System Based on Immune Algorithm

Lu Sun[1], Dianju Li[1], Hao Lu[2,3], Liangtian Wan[4(✉)], Jianbo Zheng[3,5(✉)], and Xianpeng Wang[6]

[1] Department of Communication Engineering, Institute of Information Science Technology, Dalian Maritime University, Dalian 116026, China
{sunlu,11202102251i}@dlmu.edu.cn
[2] City University of Macau, Macau, China
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
[4] School of Software, Dalian University of Technology, Dalian 116620, China
wanliangtian@dlut.edu.cn
[5] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
jianbo.zheng@smbu.edu.cn
[6] School of Information and Communication Engineering, Hainan University, Haikou 570228, China
wxpeng2016@hainanu.edu.cn

Abstract. In some production workshops, because the space is too small to be suitable for the layout of the equipment charging device, the joint application of mobile edge computing (MEC) and wireless power transfer (WPT) can solve such problems. However, this scenario has a double near-far effect that is unfair to terminal devices far away from edge servers and energy transfer stations, so this paper considers relay collaboration among devices when computing offloading. This paper uses the frequency division multiple access (FDMA) technology to enable multiple terminals to perform tasks offloading simultaneously. In this paper, the total communication delay and total energy consumption of the system are optimized through effective computing resources scheduling and reasonable tasks allocation, which is a weighting and minimizing problem of normalized system delay and energy consumption rate, and also a NP-hard problem. This paper improves the immune algorithm (IA) in the scenario of multi-server and multi-terminal to obtain the Q-IADE algorithm. The improved Q-IADE algorithm not only has the characteristics of wide application, but also further improves the global search

This work is supported by National Natural Science Foundation of China (62101088, 61801076), National Natural Science Foundation of Liaoning Province (2022-MS-157, 2023-MS-108, Fundamental Research Funds for the Central Universities (3132023250), Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 290–301, 2024.
https://doi.org/10.1007/978-3-031-65126-7_26

Acknowledgement. This research was supported by the National Natural Science Foundation of China (No. 62171449).

References

1. Cantero, M., Inca, S., Ramos, A., Fuentes, M., Martín-Sacristán, D., Monserrat, J.F.: System-level performance evaluation of 5G use cases for industrial scenarios. IEEE Access **11**, 37778–37789 (2023)
2. Wang, D., et al.: 5G integrated radio transmission scheme for low earth orbit satellite access network. In: 2022 IEEE 22nd International Conference on Communication Technology (ICCT), pp. 1417–1420 (2022)
3. Leyva-Mayorga, I., et al.: LEO small-satellite constellations for 5G and beyond-5G communications. IEEE Access **8**, 184955–184964 (2020)
4. Su, Y., Liu, Y., Zhou, Y., Yuan, J., Cao, H., Shi, J.: Broadband LEO satellite communications: architectures and key technologies. IEEE Wirel. Commun. **26**(2), 55–61 (2019)
5. Chiha, A., Van der Wee, M., Briggs, K., Colle, D.: Techno-economic evaluation of a brokerage role in the context of integrated satellite-5G networks. In: 2020 6th IEEE Conference on Network Softwarization (NetSoft), pp. 1–5 (2020)
6. Liu, Y., et al.: Joint optimization based satellite handover strategy for low earth orbit satellite networks. IET Commun. **15**, 1576–1585 (2021)
7. Shi, L., Yang, F., Wu, W., Sun, A., Sun, Y., Sun, T.: Load balancing and remaining visible time based handover algorithm for LEO satellite network. In: 2022 IEEE 8th International Conference on Computer and Communications (ICCC), pp. 391–395 (2022)
8. Dai, C.-Q., Xu, J., Wu, J., Chen, Q.: Multi-objective intelligent handover in satellite-terrestrial integrated networks. In: 2022 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 367–372 (2022)
9. Miao, J., Wang, P., Yin, H., Chen, N., Wang, X.: A multi-attribute decision handover scheme for LEO mobile satellite networks. In: 2019 IEEE 5th International Conference on Computer and Communications (ICCC), pp. 938–942 (2019)
10. He, S., Wang, T., Wang, S.: Load-aware satellite handover strategy based on multi-agent reinforcement learning. In: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, pp. 1–6 (2020)
11. Wu, Y., Hu, G., Jin, F., Zu, J.: A satellite handover strategy based on the potential game in LEO satellite networks. IEEE Access **7**, 133641–133652 (2019)
12. Lei, Y.H., Cao, L.F., Da Han, M.: A handover strategy based on user dynamic preference for LEO satellite. In: 2021 7th International Conference on Computer and Communications (ICCC), pp. 1925–1929 (2021)
13. Zhang, S., Liu, A., Han, C., Ding, X., Liang, X.: A network-flows-based satellite handover strategy for LEO satellite networks. IEEE Wirel. Commun. Lett. **10**(12), 2669–2673 (2021)
14. Badini, N., Marchese, M., Patrone, F.: ns-3-based 5G satellite-terrestrial integrated network simulator. In: 2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON), pp. 154–159 (2022)
15. Ali, Z., Lagén, S., Giupponi, L., Rouil, R.: 3GPP NR V2X mode 2: overview, models and system-level evaluation. IEEE Access **9**, 89554–89579 (2021)
16. Easthope, P.F.: Examination of SGP4 along-track errors for initially circular orbits. IMA J. Appl. Math. **80**(2), 554–568 (2015)

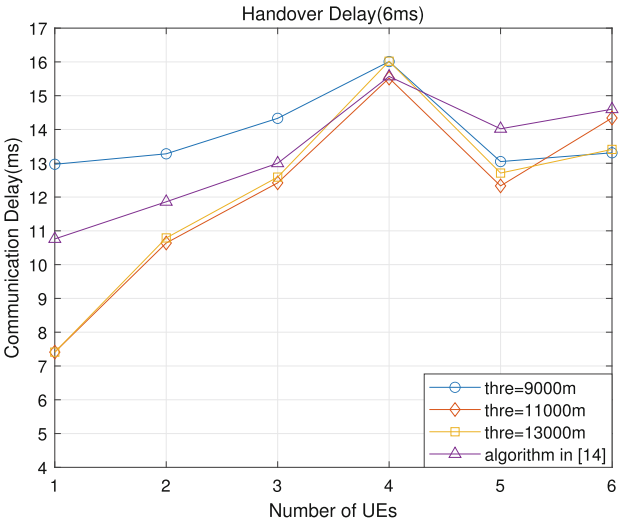


Fig. 7. Communication delay (handover delay is 6 ms).

reduced by setting a threshold and decreases with the increase of threshold, which also validates the conclusions obtained from Fig. 5, Fig. 6 and Fig. 7 well.

5 Conclusion and Future Work

In order to make the simulation more suitable for the actual scenario, this paper constructs 5G LEO satellite network based on ns-3, introduces handover delay and proposes a handover algorithm based on distance difference threshold. In addition, we consider visibility between UEs and satellite gNBs, as well as spectrum resource allocation. Simulation results indicate that when the handover delay is small, the lower distance difference threshold should be selected or the minimum distance algorithm can still be used. With the increase of handover delay, the impact caused by handover gradually increases, and the impact caused by the change of communication link distance is further weakened. In this case, the distance difference threshold should be increased to reduce the number of handover. The handover threshold should be determined based on the actual scenario and prior knowledge, which can reduce the number of handover events while ensuring that the communication distance is not too long, thereby minimizing the delay of the network.

Our future work plan is as follows. Considering the handover delay, the algorithm based only on the distance difference threshold is still not comprehensive enough. In the future, machine learning and intelligent decision-making technology will be used to optimize the handover strategy and network performance. Additionally, we will increase the number of UEs and satellite gNBs, and deploy the handover algorithm based on distance difference threshold to a larger scale 5G LEO satellite network.

threshold should be set to reduce the number of handovers, thereby reducing the communication delay.

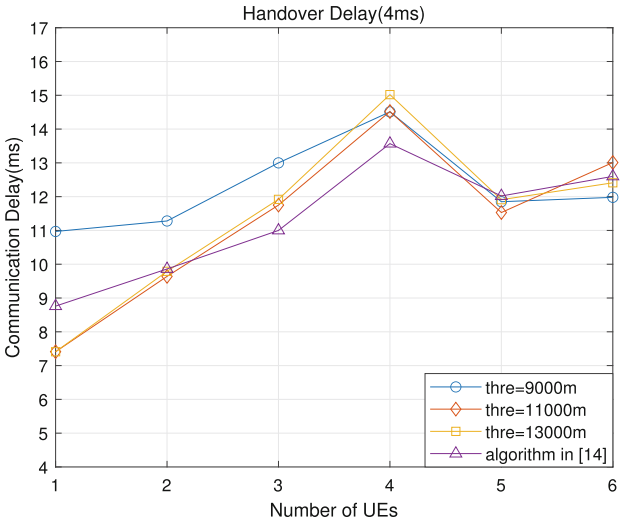


Fig. 6. Communication delay (handover delay is 4 ms).

Figure 7 presents the change of communication delay when handover delay is 6ms. For each number of UEs from 1 to 6, the delay obtained by using the mini-mum distance handover algorithm is no longer the minimum. For example, when the number of UEs is 5, the minimum delay can be obtained by setting a threshold as 11000m, which is about 2ms less than that using the minimum distance handover algorithm. In this case, the delay caused by the handovers has a more significant impact on the performance of network, while the impact of the increasing communication distance is further weakened. In extreme cases, we can even let the number of handovers approach zero to reduce the large effect on the satellite network.

Table 2 shows the number of handovers for different numbers of UEs under the set thresholds. It can be observed that the number of handovers can be

Table 2. Number of handovers for different thresholds.

Threshold	Number of Handovers					
	1UE	2UEs	3UEs	4UEs	5UEs	6UEs
9000m	1	2	2	3	3	4
11000m	0	1	1	2	2	4
13000m	0	1	1	2	2	3
Algorithm in [14]	1	2	3	4	5	6

in this paper, we set three groups of handover delay and distance difference thresholds. Table 1 shows the simulation parameter configuration.

4.3 Results and Analysis

In order to present the superior performance of the proposed algorithm, we compare our work with the minimum distance handover algorithm in [14], which can represent the state-of-the-art.

Figure 5 shows the change of communication delay when the handover delay is 2 ms. In this case, the delay corresponding to the minimum distance handover algorithm is always the minimum. This indicates when the handover delay is small, setting a threshold can reduce the number of handovers and the extra time cost caused by handovers. However, the effect of the increasing distance between UEs and the satellite gNBs (without using the minimum distance handover algorithm) is more significant. Therefore, the handovers will not have a great impact on the network, so it is more appropriate to attach UEs to nearest satellite gNBs or set a small threshold.

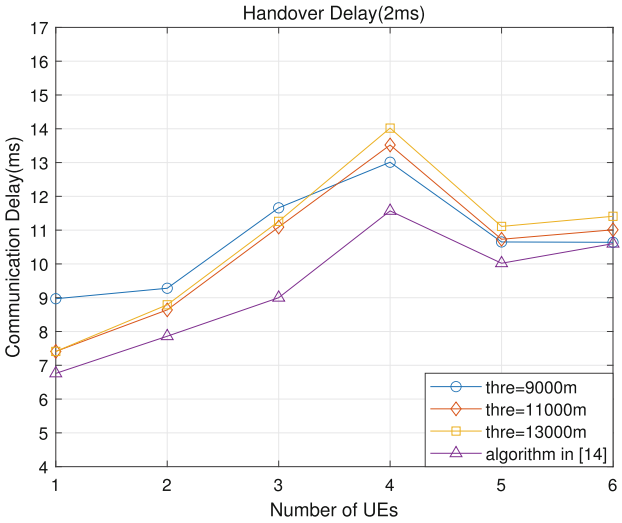


Fig. 5. Communication delay (handover delay is 2ms).

Figure 6 illustrates the change of communication delay when handover delay is 4ms. It can be observed that when the number of UEs is 1, 2, 4, and 5, the delay corresponding to the minimum distance algorithm is not the minimum. By selecting an appropriate threshold, the delay of the network can be minimized. As the handover delay increases, the occurrence of handovers will bring a more significant delay effect to the network. In contrast, the impact of the increase of communication distance will decrease. Therefore, an appropriate handover

4 Simulation and Performance Evaluation

4.1 Introduction of Simulation Platform

The work carried out in this paper is based on the ns-3 environment, which requires the use of 5G-LENA module and Satellite module. Here is a brief introduction of simulation environment and core modules.

- *ns-3*: ns-3 is a discrete network simulator that can abstract a continuous process in the physical world into a series of discrete events in the virtual world. This technology enables ns-3 to simulate various network protocols in the physical world very realistically.
- *5G-LENA module*: This module is a pluggable module of ns-3 that supports configurable Time Division Duplexing (TDD) and Frequency Division Duplexing (FDD) modes. It can also accurately model the numerology-dependent slot and OFDM symbol granularity [15]. With the 5G-LENA module, the construction and simulation of 5G communication networks can be completed effectively.
- *Satellite module*: It is developed based on the mathematical model SGP4 [16]. By inputting the Two-Line Element (TLE) data, it can output the speed and position of satellites in ECEF coordinate system and realize the node movement according to TLE data, which can be used to simulate the satellite gNBs.

4.2 Simulation Parameter Configuration

In this network, each communication link is set as Down Link (DL), meaning that the information transmission direction is from remote Host to UEs. Considering the general value of handover delay and the orbital characteristics of satellites

Table 1. Simulation Parameter Configuration.

Simulation Parameter	Value
Number of satellite gNBs	2
Number of UEs	from 1 to 6
Satellites altitude	800 km
Orbital planes eccentricity	0
Orbital planes inclination	80°
Right ascension of ascending intersection	100°
Perigee argument	90°
Minimum elevation angle	25°
Total number of transmitted packets	1800
Packet size	1280 Byte
Simulation duration	3 s
Handover delay	2 ms; 4 ms; 6 ms
Distance difference threshold	9000 m; 11000 m; 13000 m

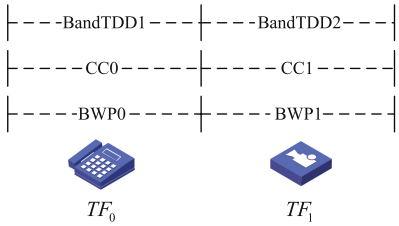


Fig. 4. Spectrum resource allocation diagram.

introduce handover delay and proposes a handover algorithm based on distance difference threshold, which replaces minimum distance handover algorithm.

The coordinates of UEs and the satellite gNBs are set in Earth-Centered Earth-Fixed (ECEF) coordinate system. The number of UEs is N, and the coordinate of UEi is (x_{ui}, y_{ui}, z_{ui}), $i = 1, 2, \cdots, N$. The number of satellites gNBs is two and the coordinate of gNBj is (x_{sj}, y_{sj}, z_{sj}), $j = 1, 2$. The positions of UEs are fixed, and the positions of satellite gNBs change with orbit. The distance from UEs to satellite gNBs can be calculated as

$$d_{ij} = \sqrt{(x_{ui} - x_{sj})^2 + (y_{ui} - y_{sj})^2 + (z_{ui} - z_{sj})^2}. \tag{3}$$

Set the simulation time as T and set the inquiry period as T_0, where $T = MT_0$. This means that handover decisions are made every T_0 for a total of M times. The moment of every decision is $t_k = kT_0$, $k = 0, 1, \cdots, M - 1$.

The specific process of the algorithm is presented in Algorithm 1. At the initial time (at time t_0), each UE is attached to the nearest satellite gNB, which is also the initial condition of the algorithm. At the moment $t_k (k \neq 0)$ when the handover decision needs to be made, follow steps in Algorithm 1.

Algorithm 1: Handover algorithm at time t_k

Input: Coordinates of UEs (x_{ui}, y_{ui}, z_{ui}), Coordinates of satellite gNBs (x_{sj}, y_{sj}, z_{sj}), Distance difference threshold *thre*

Output: Connection relationship between UEs and satellite gNBs at time t_k

```
for i in 1, 2, ··· , N do
    Calculate the distance between UEi and gNB1, which is defined as d_i1;
    Calculate the distance between UEi and gNB2, which is defined as d_i2;
    if UEi is attached to gNB1 after the last decision (at time t_{k-1}) then
        if d_i1 − d_i2 < thre then attach UEi to gNB1else attach UEi to gNB2
    else
        if d_i2 − d_i1 < thre then attach UEi to gNB2else attach UEi to gNB1
    end
end
```

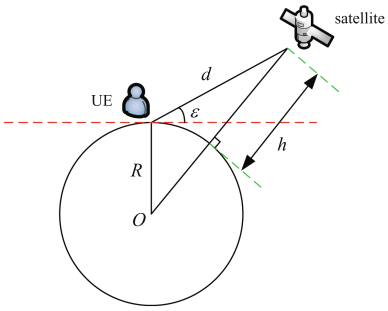


Fig. 3. Elevation diagram between UE and satellite.

where, ε is the elevation angle, R=6378km is the earth radius, h is the orbital altitude of the satellite, and d is the distance between UE and the satellite. The distance can be obtained as

$$d = -R \cdot \sin(\varepsilon) + \sqrt{R^2 \cdot \sin^2(\varepsilon) + h^2 + 2 \cdot R \cdot h}. \tag{2}$$

In the 5G LEO satellite network, the minimum elevation angle and orbital altitude are constrained, and the distance range can be calculated by Eq. (2), which can ensure the visibility between UE and satellite during simulation time.

2.3 Spectrum Resource Allocation

Band Width Part (BWP) and Component Carrier (CC) configurations are also implemented. Services of each UE may have different requirements on communication performance, and the 5G LEO satellite network can support a variety of applications with different requirements. In this paper, UE services are divided into two groups according to communication performance requirements, namely TF_0 and TF_1, and available spectrum resources are allocated to them. The available frequency band is divided into two segments for the transmission of the above two services. Both segments of spectrum are set as Time Division Duplexing (TDD) mode, and each segment only has one CC and BWP. In addition, each satellite gNB can support access to two services. Figure 4 shows the TDD based spectrum resource allocation diagram.

3 Handover Algorithm Based on Distance Difference Threshold

In non-terrestrial network with LEO satellite as gNBs, the distance between UE and satellite is changing rapidly, so UEs should handover the connection relationship with satellite gNBs effectively. The traditional handover algorithm requires each UE to be attached to the nearest satellite gNB. However, the occurrence of handover events will inevitably bring delay effects. Therefore, we

2 5G LEO Satellite Network Model

2.1 Model Description

The 5G LEO satellite network model is shown in Fig. 2.

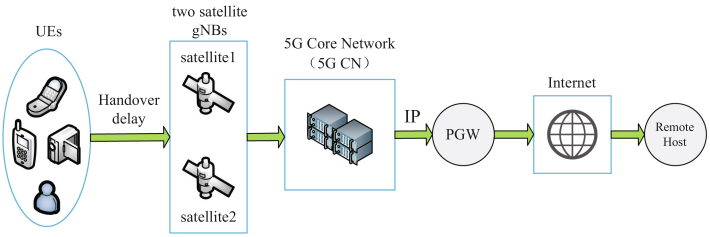


Fig. 2. 5G LEO satellite network model.

The core components are listed as follows.

- *UEs*: User devices on the ground that can access the Internet through 5G New Radio (NR) to send and receive data. In ns-3, UEs are abstracted as ground nodes.
- *gNBs*: Two satellite base stations, both of which can provide direct access to UEs. Usually, a satellite gNB receives access from many UEs. Therefore, gNBs are responsible for managing UEs, including resource allocation, scheduling and access policy management.
- *5G Core Network (CN)*: Manages all functions related to the establishment and maintenance of a 5G communications network, which may consist of one or more physical or virtual nodes connected to each other.
- *PGW*: Packet Data Network Gateway (PGW) connects the architecture to the Internet. If a UE wants to access the Internet, it must pass through the PGW entity. PGW allocates IP addresses for UEs and provides IP routing and forwarding functions. In addition, PGW has other functions, such as different billing and different policies based on users and services.

2.2 Guarantee of Visibility Between UEs and Satellites

When a UE located on the ground communicates with a satellite gNB, the visibility should be taken into account, that is, the elevation angle should not be less than a certain threshold. When the elevation angle is too small, there will be a large communication delay. Figure 3 visually shows the elevation angle between UE and satellite.

According to the law of cosines

$$(R+h)^2 = d^2 + R^2 - 2 \cdot d \cdot R \cdot \cos(90^\circ + \varepsilon), \tag{1}$$

relevant technologies have also been applied in the research of satellite network handover algorithms [9,10]. Considering the signal strength, the remaining service time and the number of idle channels of the candidate satellite, Miao et al. [9] described a handover algorithm for LEO mobile satellite networks based on multi-attribute decision. He et al. [10] proposed a novel satellite handover strategy based on multi-agent reinforcement learning that aims to minimize average satellite handovers. Simulation results show that the above two algorithms have outstanding performance in reducing the blocking rate of UEs. The current LEO satellite handover algorithms focus on considering satellite network parameters, and some scholars have developed a new approach based on preferences of UEs [11,12]. Wu et al. [11] proposed a handover algorithm to maximize the benefits of mobile terminals of UEs based on their preferences, which can greatly improve the call quality. Similarly, according to the known dynamic preference information, Lei et al. [12] adopted dividing the time period of different services to screen out candidate handover satellites, and used the decision matrix to select the satellite for UEs to handover to meet their requirements. In addition, for the satellite handover of massive User Terminals (UTs) in mega constellation, Zhang et al. [13] proposed an improved handover algorithm based on the existing net-work-flows (HSNF) algorithm to enhance the algorithm performance by preventing infinite loop.

Previous literature have analyzed the handover algorithm of satellite networks from multiple perspectives, but the simulation tools used cannot simulate the network in the physical world well and the evaluation indicators of network quality focus only on throughput and blocking rate of UEs. Therefore, in [14], the author built a satellite-ground integrated network using network simulator (ns-3) which performs better than other simulators. Moreover, UEs are attached to the nearest satellite and communication delay is used as the performance indicator. Simulation results show that the delay of the satellite network is larger than that of the ground network due to the longer communication link. However, this analysis is based on an ideal scenario where the occurrence of handover events will not bring additional effects such as delay, which is limited in practice.

To overcome the shortcomings of [14], this paper adopts ns-3 to build a 5G LEO satellite network, introduces handover delay for each handover event, and proposes a handover algorithm based on distance difference threshold. By setting multiple groups of handover delay and distance difference thresholds and comparing them with the minimum distance handover algorithm in [14], we explore the influence of handover delay and threshold selection on network performance.

The rest of this paper is organized as follows. Section 2 describes the model of the 5G LEO satellite network, including the role of each functional module, the description of visibility between UEs and satellite gNBs, spectrum resource allocation. Section 3 introduces the handover algorithm based on distance difference threshold in detail. Section 4 uses ns-3 to conduct network simulation, introduces the simulation platform as well as parameter configuration, and analyzes the simulation results. Finally, Sect. 5 summarizes the work of the full paper and puts forward the future work plan.

5G has been widely adopted across various fields, including industrial control, autonomous driving, smart home, and many more.

Compared to 5G ground communication network, satellite communication offers significant advantages, such as wide coverage and stable communication. In remote areas such as deserts and oceans, ground network may not provide coverage or may be too costly [2], while satellite network can offer low-cost and efficient network access services. Additionally, satellite communication is not affected by weather or terrain, and the probability of signal interruption and other failures is low, which make it suitable for emergency communication. The Low Earth Orbit (LEO) satellite network, in particular, offers lower communication delay, lower link loss, and higher transmission rate compared to High Earth Orbit satellite network, and can achieve super-large system capacity through giant constellations [3,4]. Combining LEO satellites as the Next Generation Base Station (gNB) with 5G ground networks [5] can make up for the shortcomings of 5G ground networks in transportation, maritime communication, telemedicine for remote areas, and other fields, with broad development prospects. Figure 1 shows the application scenario of 5G LEO satellite network. However, due to the fast movement of LEO satellites, when only one satellite gNB is servicing User Equipment (UE), it is easy to lose connection between UE and the satellite. To address this issue, it is necessary to construct LEO satellite clusters and handover the connection relationship between UEs and satellite gNBs.

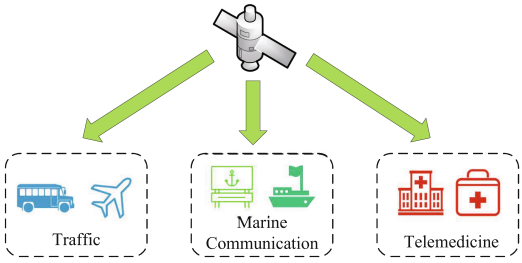


Fig. 1. Application scenario of 5G LEO satellite network.

Recent years have witnessed a growing number of works on handover in satellite networks. Related researches have been carried out from multiple perspectives. Some scholars study handover algorithms from the perspective of load balancing [6–8]. Based on the minimum handover frequency algorithm, Liu et al. [6] outlined a load balanced satellite handover strategy, which can optimize the power allocation of the satellite and improve the system capacity. Shi et al. [7] designed a handover algorithm named as Load Balancing and Remaining Visible Time based Handover (LBRVTH) in LEO satellite network. Simulation results show that the algorithm can effectively ensure better quality of service (QoS). Dai et al. [8] proposed a multi-objective intelligent handover (MIHO) algorithm to increase balance load. This algorithm has good performance in both throughput and load balancing. With the vigorous development of artificial intelligence,

Research on Handover Technology for 5G LEO Satellite Network Based on ns-3

Zheng Wang[1], Li Zhou[2(✉)], and Yankun Wang[2]

[1] Beijing Institute of Astronautical Systems Engineering, China Academy of Launch Vehicle Technology, Beijing 100076, China
2809940902@qq.com

[2] College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China
zhouli2035@nudt.edu.cn

Abstract. The 5G Low Earth Orbit (LEO) satellite network offers several advantages, including wide coverage, stable communication, and strong flexibility, making it an ideal solution for high-quality communications. Due to the fast movement of satellites, research on handover technology for LEO satellite networks is crucial. This paper uses the network simulator (ns-3) to build a 5G LEO satellite network and introduces delay for each handover process. We propose a handover algorithm based on distance difference threshold, taking into account spectrum allocation. In order to evaluate quality of service (QoS) of the network, we explore the influence of handover delay and threshold selection by setting multiple groups of handover delay and distance difference thresholds. Simulation results indicate that as handover delay increases, the impact of handover events on the network also increases gradually. Moreover, an appropriate threshold based on the actual situation can reduce the number of handovers and minimize communication delay. In particular, when the handover delay is 6ms, the communication delay can be reduced by about 2ms with an appropriate distance difference threshold.

Keywords: ns-3 · 5G LEO satellite network · handover delay · distance difference threshold

1 Introduction

The 5th Generation Mobile Communication Technology (5G) boasts superior network speed and lower latency, with advanced features such as ultra-high frequency band and large-scale multiple-input multiple-output (MIMO) technology. This enables a greater number of devices to connect to the network simultaneously and also supports an increased number of smart device connections. The International Telecommunications Union (ITU) has classified 5G services into three primary categories [1], namely Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communication (URLLC), and Massive Machine Type Communication (mMTC). As a hot research topic worldwide,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 279–289, 2024.
https://doi.org/10.1007/978-3-031-65126-7_25

Artificial Intelligence and Machine Learning I

13. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**(1), 1–7 (2020)
14. Salari, N., et al.: Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. Glob. Health **16**(1), 1–11 (2020)
15. Sarma, K.V., et al.: Federated learning improves site performance in multicenter deep learning without data sharing. J. Am. Med. Inform. Assoc. **28**(6), 1259–1264 (2021)
16. Sun, X., et al.: Sleep behavior and depression: findings from the China Kadoorie biobank of 0.5 million Chinese adults. J. Affect. Disord. **229**, 120–124 (2018)
17. Wang, R., et al.: Predicting symptom trajectories of schizophrenia using mobile sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 1–24 (2017)
18. Wu, H., Wang, P.: Fast-convergent federated learning with adaptive weighting. IEEE Trans. Cogn. Commun. Netw. **7**(4), 1078–1088 (2021)
19. Xu, X., Peng, H., Sun, L., Bhuiyan, M.Z.A., Liu, L., He, L.: Fedmood: federated learning on mobile health data for mood detection. arXiv preprint arXiv:2102.09342 (2021)
20. Yang, F., et al.: Internet-of-things-enabled data fusion method for sleep healthcare applications. IEEE Internet Things J. **8**(21), 15892–15905 (2021)
21. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-IID data. arXiv preprint arXiv:1806.00582 (2018)

Therefore, the gradient divergence can be further bounded by:

$$\begin{aligned}
&||G_{mE+t}^{(k)} - G_{mE+t}^{(c)}|| \\
\leq &\alpha^t * ||G_{mE}^{(g)} - G_{mE}^{(c)}|| \\
&+(1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * ||\nabla F_{mE+j-1}^{(k)} - \nabla F_{mE+j-1}^{(c)}|| + \beta * t \\
\leq &\alpha^t * ||G_{mE}^{(g)} - G_{mE}^{(c)}|| + (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * \sigma + \beta * t \\
= &\alpha^t * ||G_{mE}^{(g)} - G_{mE}^{(c)}|| + (1-\alpha^t) * \sigma + \beta * t.
\end{aligned}$$

Hence, Proposition 2 is proved.

References

1. Aledavood, T., Torous, J., Triana Hoyos, A.M., Naslund, J.A., Onnela, J.P., Keshavan, M.: Smartphone-based tracking of sleep in depression, anxiety, and psychotic disorders. Curr. Psychiatry Rep. **21**(7), 1–9 (2019)
2. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: an overview. IEEE Signal Process. Mag. **35**(1), 53–65 (2018)
3. Garcia-Ceja, E., et al.: Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In: Proceedings of the 9th ACM on Multimedia Systems Conference, MMSys 2018. ACM, New York (2018). https://doi.org/10.1145/3204949.3208125
4. Krizhevsky, A., Hinton, G.: Convolutional deep belief networks on CIFAR-10. Unpublished manuscript, vol. 40, no. 7, pp. 1–9 (2010)
5. Kumar, R., et al.: Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. IEEE Sens. J. **21**(14), 16301–16314 (2021)
6. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010). http://yann.lecun.com/exdb/mnist/
7. Li, X., Orabona, F.: On the convergence of stochastic gradient descent with adaptive stepsizes. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 983–992. PMLR (2019)
8. Liu, B., Yan, B., Zhou, Y., Yang, Y., Zhang, Y.: Experiments of federated learning for COVID-19 chest X-ray images. arXiv preprint arXiv:2007.05592 (2020)
9. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
10. Qayyum, A., Ahmad, K., Ahsan, M.A., Al-Fuqaha, A., Qadir, J.: Collaborative federated learning for healthcare: multi-modal COVID-19 diagnosis at the edge. arXiv preprint arXiv:2101.07511 (2021)
11. Reddi, S., et al.: Adaptive federated optimization. arXiv preprint arXiv:2003.00295 (2020)
12. Ren, X., Huang, W., Pan, H., Huang, T., Wang, X., Ma, Y.: Mental health during the COVID-19 outbreak in China: a meta-analysis. Psychiatr. Q. **91**(4), 1033–1045 (2020)

$$
\begin{aligned}
&G^{(k)}_{mE+t} \\
=&\alpha * G^{(k)}_{mE+t-1} + (1-\alpha) * \nabla F^{(k)}_{mE+t-1} \\
=&\alpha * \left[\alpha * G^{(k)}_{mE+t-2} + (1-\alpha) * \nabla F^{(k)}_{mE+t-2}\right] \\
&+ (1-\alpha) * \nabla F^{(k)}_{mE+t-1} \\
=&\alpha^2 * G^{(k)}_{mE+t-2} + \alpha * (1-\alpha) * \nabla F^{(k)}_{mE+t-2} \\
&+ (1-\alpha) * \nabla F^{(k)}_{mE+t-1} \\
=&\alpha^3 * G^{(k)}_{mE+t-3} + \alpha^2 * (1-\alpha) * \nabla F^{(k)}_{mE+t-3} \\
&+ \alpha * (1-\alpha) * \nabla F^{(k)}_{mE+t-2} + (1-\alpha) * \nabla F^{(k)}_{mE+t-1} \\
=&\alpha^t * G^{(k)}_{mE} + (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * \nabla F^{(k)}_{mE+j-1} \\
=&\alpha^t * G^{(g)}_{mE} + (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * \nabla F^{(k)}_{mE+j-1}.
\end{aligned}
$$

Since $\alpha^t + (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} = 1$ and $G^{(c)}_t = \nabla F^{(c)}_t$ for centralized method, $G^{(k)}_{mE+t}$ as the local estimated function of gradient after t times of centralized updates is equal to ($t \geq 2$):

$$
G^{(c)}_{mE+t} = \alpha^t * G^{(c)}_{mE+t} + (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * \nabla F^{(c)}_{mE+t}.
$$

Then, to calculate the difference of $G^{(c)}_{mE+t}$ and $G^{(k)}_{mE+t}$, we can have the gradient divergence:

$$
\begin{aligned}
&||G^{(k)}_{mE+t} - G^{(c)}_{mE+t}|| \\
\leq&\alpha^t * ||G^{(g)}_{mE} - G^{(c)}_{mE+t}|| \\
&+ (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * ||\nabla F^{(k)}_{mE+j-1} - \nabla F^{(c)}_{mE+t}|| \\
\leq&\alpha^t * \left[||G^{(g)}_{mE} - G^{(c)}_{mE}|| + ||G^{(c)}_{mE+t} - G^{(c)}_{mE}||\right] \\
&+ (1-\alpha) * \sum_{j=1}^{t} \alpha^{t-j} * \Big[||\nabla F^{(k)}_{mE+j-1} - \nabla F^{(c)}_{mE+j-1}|| \\
&+ ||\nabla F^{(c)}_{mE+j-1} - \nabla F^{(c)}_{mE+t}||\Big].
\end{aligned}
$$

As we assume $\beta-smooth$ for the centralized gradient $\nabla F^{(c)}_t$, $\forall j \in \{1, 2, .., t+1\}$, we have:

$$
||\nabla F^{(c)}_{mE+j-1} - \nabla F^{(c)}_{mE}|| \leq \beta * t.
$$

Since $\sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} = 1$, the equation can be transformed into:

$$w_{mE}^{(g)} = w_{(m-1)E}^{(g)} - \eta * \sum_{k=1}^{K} \sum_{t=0}^{E-1} \frac{|D^{(k)}|}{|D|} * G_{(m-1)E+t}^{(k)}.$$

Meanwhile, for parameters under centralized learning, similarly we can formulate it as:

$$\begin{aligned} w_{mE}^{(c)} &= w_{(m-1)E}^{(c)} - \eta * \sum_{t=0}^{E-1} G_{(m-1)E+t}^{(c)} \\ &= w_{(m-1)E}^{(c)} - \eta * \sum_{k=1}^{K} \sum_{t=0}^{E-1} \frac{|D^{(k)}|}{|D|} * G_{(m-1)E+t}^{(c)}. \end{aligned}$$

Therefore, the parameter divergence between global parameters and centralized parameters can be formulated as:

$$\begin{aligned} & ||w_{mE}^{(g)} - w_{mE}^{(c)}|| \\ =& ||w_{(m-1)E}^{(g)} - \eta * \sum_{k=1}^{K} \sum_{t=0}^{E-1} \frac{|D^{(k)}|}{|D|} * G_{(m-1)E+t}^{(k)} \\ & - w_{(m-1)E}^{(c)} + \eta * \sum_{k=1}^{K} \sum_{t=0}^{E-1} \frac{|D^{(k)}|}{|D|} * G_{(m-1)E+t}^{(c)}|| \\ \leq & ||w_{(m-1)E}^{(g)} - w_{(m-1)E}^{(c)}|| \\ & + \eta * \sum_{t=0}^{T-1} \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * ||G_{(m-1)E+t}^{(k)} - G_{(m-1)E+t}^{(c)}||. \end{aligned}$$

Hence Proposition 1 is proved.

A.2 Proof of Proposition 2

Based on the assumption that $||G_t^{(k)} - G_t^{(c)}|| \leq \sigma$, the bounding for FedAVG can be easily proved:

$$\sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} ||G_{mE+t}^{(k)} - G_{mE+t}^{(c)}|| \leq \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * \sigma.$$

As for the second inequality, we first prove that according to Eqs. 3 and 4 (adaptive gradient algorithm), $G_{mE+t}^{(k)}$ as the local estimated function of gradient after t times of local updates is equal to ($t \geq 2$):

for aggregation can be reduced potentially, thus saving network resources and energy consumption.

6 Conclusion

In this article, an adaptive gradient method for FL optimization is proposed to achieve faster training convergence on non-IID datasets. We also analytically show that the parameter divergence under non-IID datasets is reduced by our method. By adaptively updating the local gradients based on the global gradient, we demonstrate that our algorithm can reduce the gradient divergence during local training. As a result, the convergence performance for non-IID data training case is improved significantly. Upon that, an application to depression detection is developed based on a sleep monitoring system. The experiments validate that our proposed algorithm outperforms the commonly adopted FedAVG as well as other adaptive FL optimization algorithms. Furthermore, the effectiveness and feasibility of our proposed method on the depression detection system is demonstrated.

A Appendix

A.1 Proof of Proposition 1

According to Eq. (1), after E times of local updates, the local parameters can be formulated as:

$$\begin{aligned} w_{mE}^{(k)} =& w_{mE-1}^{(k)} - \eta * G_{mE-1}^{(k)} \\ =& w_{mE-2}^{(k)} - \eta * G_{mE-1}^{(k)} - \eta * G_{mE-2}^{(k)} \\ =& w_{(m-1)E}^{(k)} - \eta * G_{mE-1}^{(k)} - \eta * G_{mE-2}^{(k)} \\ & - ... - \eta * G_{mE-1}^{(k)} \\ =& w_{(m-1)E}^{(k)} - \eta * \sum_{t=0}^{E-1} G_{(m-1)E+t}^{(k)} \\ =& w_{(m-1)E}^{(g)} - \eta * \sum_{t=0}^{E-1} G_{(m-1)E+t}^{(k)}. \end{aligned}$$

We notice that due to parameter synchronization after global aggregation, we have $w_{(m-1)E}^{(k)} = w_{(m-1)E}^{(g)}$. According to Eq. (2), after global parameter aggregation, the global parameters can be formulated as:

$$\begin{aligned} w_{mE}^{(g)} &= \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * w_{mE}^{(k)} \\ &= \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * \left[w_{(m-1)E}^{(g)} - \eta * \sum_{t=0}^{E-1} G_{(m-1)E+t}^{(k)} \right]. \end{aligned}$$

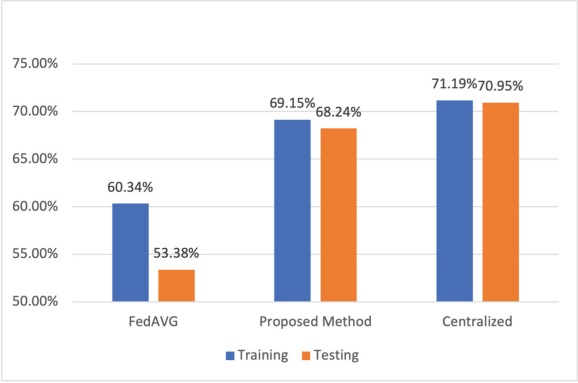


Fig. 7. Training and testing accuracy of depression classification under non-IID training, with FedAVG, our proposed method and centralized learning respectively

5.3 Depression Classification

To validate the feasibility of the depression classification with our proposed method, a preliminary experiment is conducted according to the dataset collected by [3]. In the experiment context, 23 depressive patients and 32 non-depressive subjects were involved, with wearable actigraph watches collecting subjects' motor activity. The activity data during sleep is used to classify depression. To simulate the non-IID FL training conditions, the dataset is divided into 5 groups, with 2 groups containing entirely depressive samples. In reality, the majority of depressive samples are collected in hospitals, while non-depressive samples can be found located in schools, homes, etc.

After feature engineering, 7 significant features are selected for training. We develop a 3-layer multi-layer perceptron (MLP) model for depression classification with the following architecture: 7×8 Fully connected $\rightarrow 8 \times 32$ Fully connected $\rightarrow 32 \times 1$ Fully connected. We compare the non-IID training performance of our proposed adaptive gradient method with FedAVG and the centralized learning method. The hyperparameter α of our proposed method is adjusted to 0.4 due to the simplicity of the training dataset.

The accuracy of training and testing after 100 epochs is illustrated in Fig. 7. As indicated by the result, with our proposed method, the accuracy of training and testing is approaching to the centralized method, while with FedAVG the training can hardly converge. The poor convergence speed of FedAVG can significantly increase the global rounds of FL aggregation, thus intensifying the network communication burden. The result demonstrates the feasibility of applying adaptive gradient FL optimization to medical applications, which typically contain non-IID training datasets. The advantage of fast convergence speed and robustness of training can accelerate the FL non-IID training process. Meanwhile, with the adaptive gradient algorithm, the global communication rounds

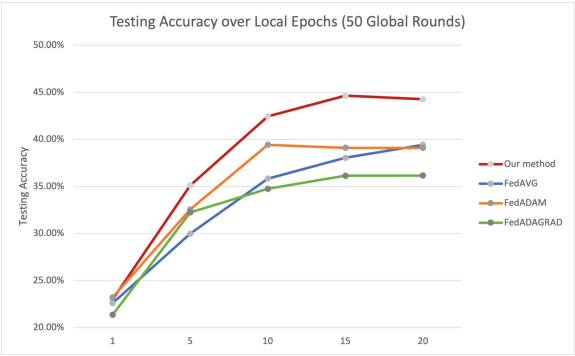


Fig. 6. Testing accuracy trained on CIFAR-10 with 50 global rounds over local training epoch, with FedAVG, FedADAM, FedADAGRAD and our proposed method respectively

on CIFAR-10 is much harder than non-IID training on MNIST [21]. Therefore, we observe testing accuracy converges only to around 50%.

Regardless, we can observe that with local epochs more than 5, the convergence rate of our proposed method is much faster compared to other optimization methods. While the performance is similar when the local epoch is equal to 1. The reason is that according to Proposition 1, with more local epochs, local gradient divergence will be accumulated to bias the global training, making it difficult to converge due to the high variance. Since the designed adaptive gradient can reduce the gradient divergence caused by non-IID datasets, the advantage stands out as the number of local epochs increases. Therefore, our proposed method can potentially enable much more local training epochs because the local updates are stabilized using the adaptive gradient. With more local train epochs, the communication cost can be reduced consequently.

In Fig. 6, we validate that our proposed adaptive gradient method can support more iterations of local training. The testing accuracy after 50 global aggregations is plotted, given local epochs ranging from 1 to 20. We can observe that with our proposed method, the testing accuracy increases much faster than the other algorithms as local epochs get larger. FedADAM and FedADAGRAD have little accuracy improvement when the number of local training epochs is more than 10, while FedAVG's accuracy increases in a relatively slow pace. Meanwhile, when the number of local epochs is more than 15, the testing accuracy of our proposed method stops increasing. The reason is that according to Proposition 2, the number of local training epochs cannot be too large, otherwise the advantage of our proposed algorithm (variance reduction in gradient divergence) would be compromised.

5.1 Experiment Setting

1. **Dataset**: For the simulated image classification training, we evaluate different algorithms by training on the typical image classification datasets: MNIST and CIFAR-10 [4,6]. MNIST (60000 training samples and 10000 testing samples) is the handwritten images dataset with 10 classes. CIFAR-10 (50000 training samples and 10000 testing samples) is the colored images dataset with 10 classes. For IID training, each dataset is split into 100 equal-size subsets randomly for clients (MNIST: 100*600; CIFAR-10: 100*500). For non-IID training, we consider the extreme case where each client contains samples from only one class. The dataset is firstly sorted according to labels, then each client selects a subset of samples with the same label.
2. **Model**: CNN models are applied as deep learning models for the FL training. For MNIST, the CNN model has 6 layers with the following architecture: $5 \times 5 \times 10$ Convolutional $\rightarrow 5 \times 5 \times 20$ Convolutional $\rightarrow$ p = 0.5 Dropout $\rightarrow 320 \times 50$ Fully connected $\rightarrow 50 \times 10$ Fully connected. For CIFAR-10, the CNN model has 6 layers with the following architecture: $5 \times 5 \times 6$ Convolutional $\rightarrow 2 \times 2$ MaxPool $\rightarrow 5 \times 5 \times 16$ Convolutional $\rightarrow 400 \times 120$ Fully connected $\rightarrow 120 \times 84$ Fully connected $\rightarrow 84 \times 10$ Fully connected.
3. **Federated Learning**: The number of clients: $K = 200$; Local dataset size: $|D_i| = 300$ for MNIST and $|D_i| = 250$ for CIFAR-10; Local batch size: $B = 10$. Global round m and local update epochs E are two variants to control the experiments. We set $\eta = 0.01$, $\alpha = 0.9$ for hyperparameters.

5.2 Image Classification

The simulation result over MINST is illustrated in Fig. 4. We plot the testing accuracy vs. global rounds of FL with IID training and non-IID training respectively. From the diagram, we can observe from the non-IID training result that our proposed method has a much faster convergence speed than the other three methods, with local epochs ranging from 1 to 20. In particular, our proposed method converges very fast for the initial rounds since the gradient divergence is more significant at the early stage. It demonstrates that our proposed adaptive gradient method outperforms other optimization methods with adaptive learning rate (FedADAM and FedADAGRAD). With a large number of clients and extremely unbalanced local datasets, both FedADAM and FedADAGRAD algorithms can hardly deal with the parameter divergence with huge variance, while our proposed method can reduce the variance to enable robust training. As for IID training, the proposed adaptive gradient method has a similar convergence performance to the other three algorithms, which indicates that our algorithm is very suitable for solving non-IID data training.

The simulation result over CIFAR-10 is illustrated in Fig. 5. According to MNIST's result, the performance of different algorithms over IID training is similar. Thus we only demonstrate the non-IID data scenario. Due to the complexity of CIFAR-10 and the simplicity of our CNN model, the non-IID training

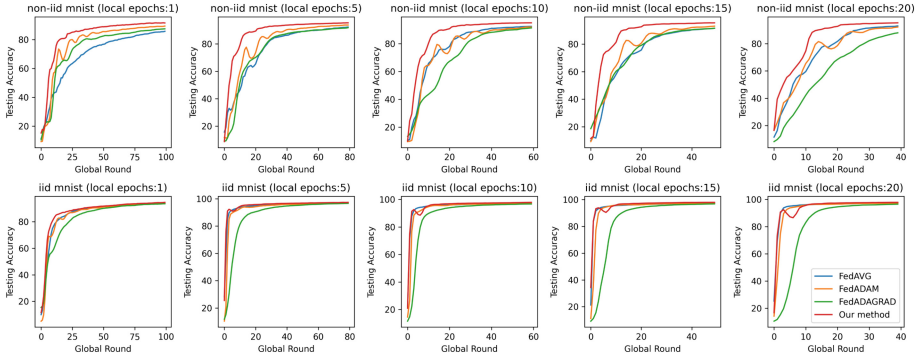


Fig. 4. Testing accuracy over global rounds of the proposed algorithm, FedAVG, FedADAM, FedADAGRAD over MNIST, with IID datasets and non-IID datasets respectively

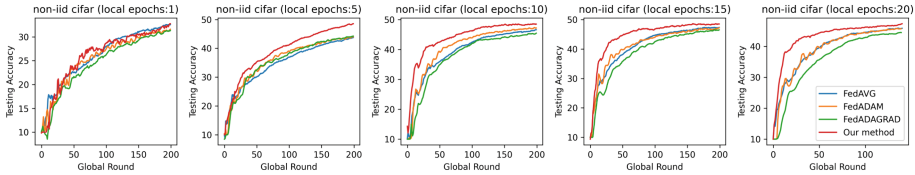


Fig. 5. Testing accuracy over global rounds of the proposed algorithm, FedAVG, FedADAM, FedADAGRAD over CIFAR-10 with non-IID datasets

As the number of non-depressive samples in the dataset is much larger than the number of depressive samples, the classifiers generally prefer to classify the sample as non-depressive. To ameliorate this problem, we design a voting mechanism for the ensemble method, where classifiers with a higher preference for depressive samples would be weighed more. The performance can also be improved by adding penalty terms to the weak classifiers.

5 Experiment and Validation

To validate the performance of our proposed adaptive gradient optimization, we have implemented the algorithm with PyTorch environment and conducted a series of experiments to demonstrate the convergence performance. In Sect. 5.1, we firstly introduce the experiment setting. Then, in Sect. 5.2 we perform and compare the simulated FL training for non-IID image classification problems, using our proposed method, FedAVG, FedADAM, and FedADAGRAD respectively. In Sect. 5.3, we validate the performance of our algorithm on the depression detection application using real-life data collected from wearable smartwatches.

features including root mean square (rms), variance (var), and mean (avg) are calculated and a low-pass filter is applied to classify users' body movement. Finally, we extract sleep features including movement rate (MR), average movement amplitude (AMA), and average movement interval (AMI), which can measure users' sleep quality.

- **Activity Data**: The activity data is a series of data points, which are constituted by a timestamp and an activity value. Due to the low intensity of activity during sleep, we develop a voting mechanism to determine whether a data point is sleep data or not. We define the interval before and after 15 min of a data point as the interval of interest. If there are more than 13 min of inactivity (activity value equal to 0) within the interval of interest, the data point would be considered as sleep data. Based on the time series classified as sleep data, we can extract the following sleep features: start time (ST), end time (ET) and total sleep duration (TSD), etc. Meanwhile, by analyzing the discontinuity of the sleep data, we can extract features such as the total time of waking-up (TW), the number of waking-up (NW), and the frequency of waking-up (FW). Finally, we extracted statistical features such as the highest activity value (HA) and the lowest activity value (LA).
- **Other**: The age and gender of the users are added as additional features to be considered in the analysis.

4.2 Feature Engineering

When collecting the data from different devices, it is common that data is abnormal or missing due to improper operation. Therefore, extra attention needs to be paid to handle the missing data. We use a generative adversarial network (GAN) [2] to interpolate the missing data. Using a generator to generate the interpolated data and a discriminator to discriminate between the interpolated values and the true values, we make the generated values as close to the actual values as possible through a continuous competition. To avoid the generation of anomalous data values (i.e., the generated sleep end time is earlier than the sleep start time), we reprocessed the generated values by a transformation function so that all generated values would fall within the maximum and minimum values of the feature. After data interpolation, we also normalize the data to make it comparable across dimensions, thus improving the model training's effectiveness.

4.3 Depression Classification

A machine learning model was built to classify whether a user is depressive based on the extracted feature set. The model was built following an ensemble learning approach, integrating a variety of weak classifiers, including SVM, KNN, logistic regression, case based reasoning (CBR), and decision tree classification algorithms. The ensemble learning design is to improve the accuracy and generalization of the model.

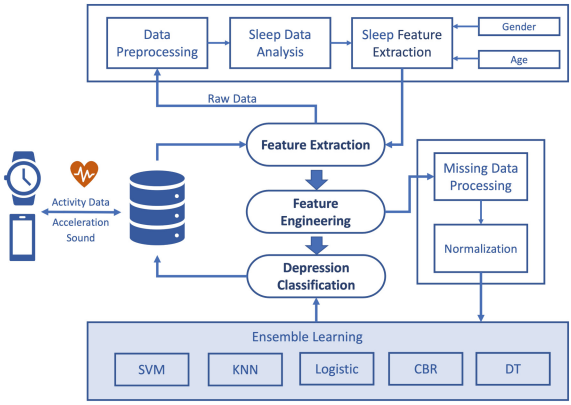


Fig. 3. Depression Classification with Sleep Data

4 Depression Detection with Sleep Monitoring

In this section, we would discuss the detailed process of depression detection with the collected sleep data. As an application of our proposed FL adaptive optimization, the depression classification models can be trained efficiently with non-IID datasets. As indicated in Fig. 3, two sleep monitoring devices are applied in our system to collect data. Smartphone serves as an unobtrusive monitor to record the overnight sound signals and bed acceleration, without physical contact with users. While wearable smartwatch, embedded with an actigraph detector, can detect activity data of wrist movement. The raw data will be analyzed according to the following steps: feature extraction, feature engineering and depression classification.

4.1 Feature Extraction

To extract sleep features related to depression classification, different types of data would follow a different data analysis process:

- **Sound**: The sound signal is sampled at a frequency of 16KHz. During data preprocessing, the sound signal is firstly segmented into 5-second frames, and each frame would be converted into a spectrogram using the short-time fourier transform (STFT). After data preprocessing, a deep learning classifier, SleepDetCNN [20], is applied to classify the spetrograms into "snoring", "coughing" or "background noise". Finally, we would extract the acoustic sleep features including total snoring time (TST) and total coughing time (TCT).
- **Acceleration**: Acceleration data is used to detect the slight 3-dimensional vibration of the bed, and to record users' body movement. According to our previous work [20], acceleration data would be segmented into frames and the noise frames would be eliminated. After data preprocessing, statistical

$||w^{(g)}_{(m-1)E} - w^{(c)}_{(m-1)E}||$ *plus the gradient divergence caused by the recent* E *local updates. Therefore, by reducing the gradient divergence within each local update, the global weight divergence can be reduced cumulatively and significantly.*

Next, we will show that under the aforementioned model, the proposed method can reduce the local gradient divergence $G^{(k)}_{(m-1)E+t} - G^{(c)}_{(m-1)E+t}$ by α^E, compared to the FedAVG algorithm. In FedAVG, the estimated gradient function is equal to the gradient of loss function:

$$G^{(k)}_t = \nabla F^{(k)}_t. \tag{7}$$

While for the adaptive gradient optimization, the estimated gradient function is adaptive according to (3) and (4). Then, we can deduce the following proposition based on these two estimated gradient functions.

Assumption 1: The difference between the estimated gradient functions for local client $k \in [K]$ and centralized method is bounded by σ, i.e., $||G^{(k)}_t - G^{(c)}_t|| \leq \sigma$ with $\sigma \geq 0$.

Assumption 2: The centralized gradient $\nabla F^{(k)}_t$ is $\beta - smooth$.

Then, we have the following proposition.

Proposition 2: *Under Assumptions 1 and 2, the average gradient divergence of all local agents can be computed as:*

– *FedAVG:*

$$\sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} \left[G^{(k)}_{(m-1)E+t} - G^{(c)}_{(m-1)E+t} \right] \leq \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * \sigma. \tag{8}$$

– *Our Method:*

$$\begin{aligned} &\sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} \left[G^{(k)}_{(m-1)E+t} - G^{(c)}_{(m-1)E+t} \right] \leq \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * \\ &\left[(1-\alpha^t) * \sigma + \alpha^t * (G^{(g)}_{(m-1)E} - G^{(c)}_{(m-1)E}) + \beta * t \right]. \end{aligned} \tag{9}$$

The proof of Proposition 2 is demonstrated in Appendix A.2. Based on Proposition 2, we can have the following remarks:

Remark 3: σ *is the main reason that causes the gradient divergence. With non-IID dataset, the value of* σ *can be potentially huge. Therefore, to improve convergence, it is necessary to alleviate the effect caused by* σ*. As the number of clients increases, it is reasonable to assume that* $G^{(g)}_{(m-1)E} = G^{(c)}_{(m-1)E}$*, neglecting the gradient divergence caused by previous updates. Therefore, the proposition indicates that our proposed method can effectively reduce the gradient divergence from* σ *to* $(1-\alpha^t) * \sigma$*, for the consecutive* t *local updates.*

Remark 4: *To reduce the local gradient divergence, we prefer a larger* $\alpha \in [0, 1]$*. However,* α *cannot be too large because it would add up the risk that the optimization would start up for a wrong direction at the beginning. Meanwhile, the gradient divergence reduction will not be prominent as local iteration* t *increases.*

3.2 Convergence Analysis

In this section, we provide mathematical proof that our proposed method can have better convergence compared to traditional FedAVG algorithm, under the context of non-IID training. We evaluate the parameter divergence, $||w_{mT}^{(g)} - w_{mT}^{(c)}||$, between federated learning mechanism (under adaptive gradient optimization approach) and traditional centralized learning approach using stochastic gradient descent (SGD) optimization method. The parameter divergence quantifies the deviation of local updates from centralized training, which directly affect the training performance of FL. Ideally, smaller parameter divergence indicates that the federated learning process is close to the centralized learning process under the non-IID scenario.

In the traditional centralized optimization method, an IID dataset D is applied. Let $w_t^{(c)}$ denotes the parameters found by centralized SGD under the t^{th} iteration. The iterative SGD optimization process can be formulated as:

$$w_{t+1}^{(c)} = w_t^{(c)} - \eta * G_t^{(c)}, \tag{5}$$

where $G_t^{(c)} = \nabla F_t^{(c)}$ is an unbiased estimate of the loss function given the aggregated data set D. We assume the mechanism of federated learning as follows: Central server will aggregate the local parameters after E local updates. After m global aggregations, we denote $w_{mE}^{(k)}$ as the parameters found by local SGD at the mE^{th} iteration, and $G_{mE}^{(k)}$ as the adaptive gradient function proposed in the previous section. Similarly, we denote $w_{mE}^{(g)}$ as global parameters at the mE^{th} iteration.

To formally bound the parameter divergence between federated learning training and centralized training, we provide the following proposition:

Proposition 1: *Given K clients with non-IID datasets $D^{(k)}, k \in [K]$ and performing E local updates each global iteration. After m global iterations, the parameter divergence satisfies:*

$$\begin{aligned} &||w_{mE}^{(g)} - w_{mE}^{(c)}|| \leq ||w_{(m-1)E}^{(g)} - w_{(m-1)E}^{(c)}|| + \\ &\eta * || \sum_{t=0}^{T-1} \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * \left[G_{(m-1)E+t}^{(k)} - G_{(m-1)E+t}^{(c)} \right] ||. \end{aligned} \tag{6}$$

The proof of Proposition 1 is demonstrated in Appendix A.1. Based on Proposition 1, we can have the following remarks:

Remark 1: *The parameter divergence under non-IID context is majorly caused by local gradient divergence at each local iteration, $G_{(m-1)E+t}^{(k)} - G_{(m-1)E+t}^{(c)}$. Using the centralized training gradient as the reference direction for weights update, the wrong direction trained by non-IID samples can slowly drift the global parameters away from the desired output.*

Remark 2: *The parameter divergence after the m^{th} global aggregation can be treated as the accumulative result of the previous global parameter divergence*

Algorithm 2. Adaptive Gradient Optimization

Initiate $w_0^{(g)}, w_{-1}^{(g)}, G_0^{(g)}$, decay parameter $\alpha \in [0, 1]$.
for $m = 0, 1, ..., T-1$ **do**
  $G_t^{(g)} = \frac{w_m^{(g)} - w_{m-1}^{(g)}}{\eta E}$
  for each client $k \in [K]$ in parallel **do**
    $t = m * E$
    $w_t^{(k)} = w_t^{(g)}$
    $G_{t-1}^{(k)} = G_t^{(g)}$
    for $e = 0, 1, ..., E-1$ **do**
      $G_t^{(k)} = \alpha * \nabla F_t^{(k)} + (1-\alpha) * G_{t-1}^{(k)}$
      $w_{t+1}^{(k)} = w_t^{(k)} - \eta * G_t^{(k)}$
      t = t +1
    end for
  end for
  $w_{m+1}^{(g)} = \sum_{k=1}^{K} \frac{|D_k|}{|D|} * w_t^{(k)}$
end for

receives both global parameter $w_{mE}^{(g)}$ and global gradient $G_{mE}^{(g)}$, which is calculated according to the global parameters of current and the previous global iterations $G_{mE}^{(g)} = -\frac{w_{mE}^{(g)} - w_{(m-1)E}^{(g)}}{\eta E}$. The global parameters $G_{mE}^{(g)}$ measures the average gradient in each local iteration, using information from two consecutive central server aggregations. Then, the local gradient of client k at the first local iteration mE is updated as follows:

$$G_{mE}^{(k)} = \alpha * G_{mE}^{(g)} + (1-\alpha) * \nabla F_{mE}^{(k)}, \tag{3}$$

where $\nabla F_{mE}^{(k)}$ is the gradient of models loss function under the local dataset $D^{(k)}$. After the first epoch of local update, the local gradient of client k at the remaining epochs is updated as:

$$G_{mE+t}^{(k)} = \alpha * G_{mE+(t-1)}^{(k)} + (1-\alpha) * \nabla F_{mE+t}^{(k)}, \tag{4}$$

where $t = 1, 2, ..., E-1$.

The idea of the adaptive gradient design is that local training can refer to the global gradient as the "the correct direction". The advantage of such a design can be summarized as follows:

1. The optimization convergence is potentially improved under the context of non-IID training because the variance of local parameter updates is reduced and the direction of local update is corrected by the global gradient.
2. Momentum information of previous training iterations is considered, which can accelerate the training convergence.
3. The global gradient as additional information transmitted through the network reveals no private information and the privacy-protecting nature of FL is preserved.
4. With local training in clients stabilized, more epochs of local updates can be performed to reduce the network communication burden.

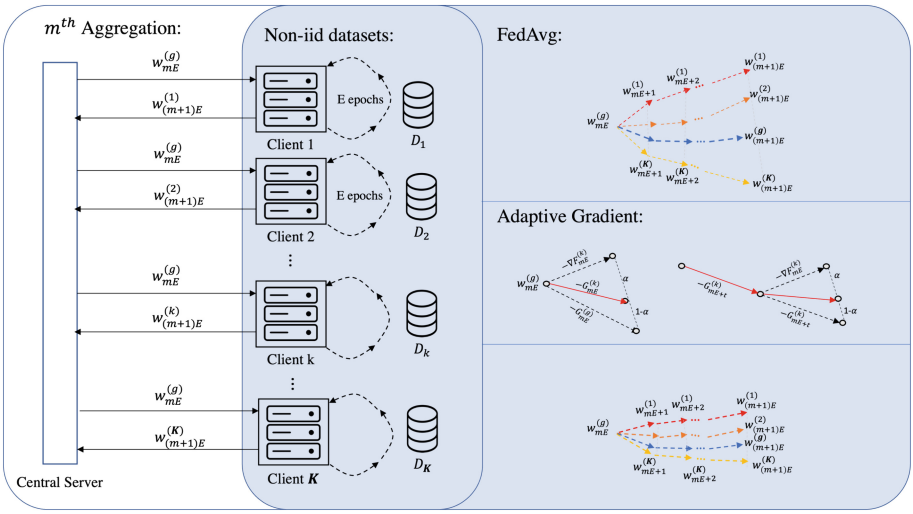


Fig. 2. Federated Learning with Adaptive Gradient

as the estimated gradient function for local client k. Then the local parameters update can be formulated as:

$$w_{mE+1}^{(k)} = w_{mE}^{(k)} - \eta * G_{mE}^{(k)}. \tag{1}$$

We denote the global parameters as $w_{mE}^{(g)}$, which is the weighted average of local parameters $w_{mE}^{(k)}$ received from all participating clients:

$$w_{mE}^{(g)} = \sum_{k=1}^{K} \frac{|D^{(k)}|}{|D|} * w_{mE}^{(k)}. \tag{2}$$

As for FedAVG, the estimated gradient function $G_{mE}^{(k)}$ is assumed to be $\nabla F_{mE}^{(k)}$, the gradient of the loss function $F_{mE}^{(k)}$, as no adaptive method is applied. As indicated in Fig. 2, due to the great variance of different local datasets, the local parameter updates of traditional FedAVG can be diverging. Because the gradient of loss function $\nabla F_{mE}^{(k)}$ is varying with dataset $D^{(k)}$. Therefore, the parameter update of client 1 can be significantly distinct from that of client K. Thus, the equivalent global parameter updates (the blue arrows), as the average of local parameter updates, potentially has huge variance and is considerably unstable. The convergence efficiency is compromised as a result.

Therefore, to alleviate the instability caused by non-IID local datasets, it is intuitive to include certain global information to guide the local updates and reduce the variance. We propose the adaptive gradient optimization method that utilize the global gradient as a reference to guide the local parameter updates. As shown in Algorithm 2, after m global iterations, each local client

Therefore, due to the non-IID nature of local training datasets, the local parameters are updated heterogeneously given different local clients. After epochs of local training, the local parameters transmitted back to the server can have a huge variance. The diversity of local parameters will be exaggerated with the increment of local training epochs as well as the level of unbalance of datasets.

Therefore, we propose the adaptive gradient optimization method for FL training with non-IID data. By utilizing the global gradient shared among clients, we can give direction for the local gradient updates and potentially reduce the parameter divergence. Different from traditional adaptive learning algorithms, which tend to modify learning rate throughout the training, our method adapts the local gradients accordingly. The detailed implementation will be further discussed in the next section.

2.3 Depression Detection Through Sleep Monitoring

As presented in Fig. 1, with our proposed FL optimization method to deal with non-IID data training, an application to the depression detection can be implemented. Traditionally, the training performance is limited due to the non-IID distribution of datasets (i.e., hospitals may contain more depressive sleep data), while our adaptive gradient method potentially improves the convergence performance during the training. As for the depression detection system, two portable devices are utilized to record users' sleep behavior, smartphones and wearable smartwatches. Both devices are common in our daily life. Smartphones can collect the overnight sound signal with microphones and detect body movement with accelerometers, while smartwatches can monitor wrist activities. The collected raw data would be segmented into frames and each frame would be classified into predefined sleep events, including body movement, snoring, coughing, etc. After event classification, features such as sleep duration and sleep efficiency can be extracted. Meanwhile, feature engineering on local datasets, including normalization and null data handling, should be performed before model training. Finally, features will be fed into a classification model for depression classification. The developed system can potentially be deployed for both medical analysis or for home-based healthcare monitoring. The detailed implementation will be further discussed in Sect. 4.

3 Federated Learning with Adaptive Gradient

3.1 Proposed Algorithm for Adaptive Gradient

Since the large heterogeneity of non-IID datasets, local parameters training of different clients may diverge from each other significantly with the traditional FedAVG. The parameter divergence can be modelled as Fig. 2: There are K local agents participating in the training with local non-IID datasets $D^{(k)}$, where $k \in [K]$. The central server will aggregate the local weights after E local updates. After m global iterations, we denote $w_{mE}^{(k)}$ as its local parameters and denote $G_{mE}^{(k)}$

Algorithm 1. FedAVG, FedADAGRAD, and FedADAM

Initiate $w_0^{(g)}, m_{-1}, v_{-1} \geq \tau^2$, decay parameters $\beta_1, \beta_2 \in [0, 1]$.
for $r = 0, 1, ..., T-1$ **do**
  $t = r * E$
  for each client $k \in [K]$ in parallel **do**
    $w_t^{(k)} = w_r^{(g)}$
    for $e = 0, 1, ..., E-1$ **do**
      Compute an unbiased estimate gradient function $G_t^{(k)}$ of $\nabla F^{(k)}(w_t^{(k)})$
      $w_{t+1}^{(k)} = w_t^{(k)} - \eta * G_t^{(k)}$
      t = t +1
    end for
  end for
  $w_{r+1}^{(g)} = \sum_{k=1}^{K} \frac{|D_k|}{|D|} * w_t^{(k)}$ (FedAVG)
  $\Delta_r = w_{r+1}^{(k)} - w_r^{(k)}$
  $m_r = \beta_1 * m_{r-1} + (1 - \beta_1) * \Delta_r$
  $v_r = \beta_2 * v_{r-1} + (1 - \beta_2) * \Delta_r^2$ (FedADAM)
  $v_r = v_{r-1} + \Delta_r^2$ (FedADAGRAD)
  $w_{r+1}^{(g)} = w_r^{(g)} + \eta \frac{m_r}{\sqrt{v_r} - \tau}$
end for

the central server, where a weighted average of all local parameters is computed and sent back to local clients for the next round's training.

Recently, the adaptive learning rate is a widely adopted federated optimization technology to enable faster and more robust convergence. FedADAGRAD and FedADAM [7] are among the state-of-art algorithms. The pseudo-codes are illustrated in Algorithm 1. According to [11], both algorithms share the same idea: remember the historical momentum information about gradients and enlarge the learning rate when the current gradient is steady. The difference between these two adaptive optimization methods is the different coefficients, β_1 and β_2, which are used to compute m_t and v_t. Experiments show that adaptive learning algorithm can have better convergence performance compared to FedAVG, which motivates us to adopt an adaptive optimization method when dealing with non-IID datasets.

2.2 Parameter Divergence with Non-IID Datasets

Even though adaptive-learning-rate algorithms can boost the FedAVG to achieve more robust convergence, FL optimization algorithms generally perform bad with respect to non-IID datasets. The reason behind the poor convergence performance is discussed in [21], where the author proposed the parameter (or weight) divergence model and quantified the parameter divergence by Earth Mover's Distance (EMD).

As illustrated in Fig. 1, datasets collected by different clients can vary significantly from other clients. For depression classification, datasets from hospitals contain more positive samples compared to datasets collected from schools.

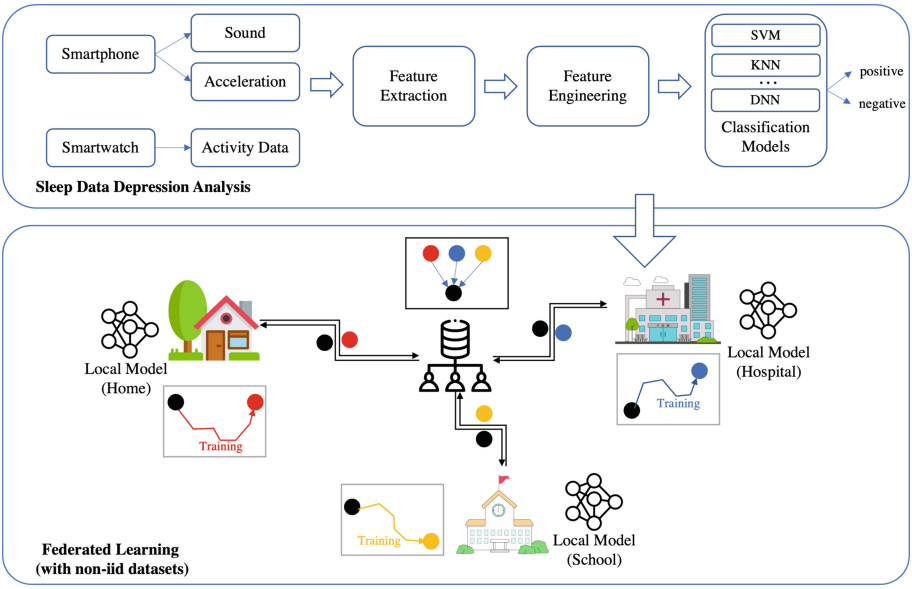


Fig. 1. System Architecture

In the following section, we first discuss several commonly adopted FL algorithms and adaptive optimization methods. Following that, we introduce the issues caused by non-IID data training and the motivation of our proposed method. Finally, we discuss how our proposed method can be applied into depression detection based on sleep monitoring systems.

2.1 Federated Learning and Adaptive Optimization

Federated learning is a decentralized machine learning mechanism, where a globally shared model is trained collaboratively by a set of distributed local clients. As illustrated in Fig. 1, local clients can be located in hospitals, schools, or even at home. Different from the centralized learning method, the local dataset stored in each client is not transmitted to the central server. Rather, only model parameters are transmitted. Therefore, FL is considered as a privacy-preserving technology for many data-sensitive applications.

To collaboratively train a global model, FedAVG is a commonly adopted optimization algorithm, which is based on stochastic gradient descent (SGD) algorithm [9]. The FedAVG pseudo-code is illustrated in Algorithm 1. The central server firstly sends the current global parameters $w_r^{(g)}$ at global iteration r to the set of local clients $[K]$, $[K] = \{1, 2, \ldots, K\}$. After receiving the global parameters, the local client k ($k \in [K]$) would train its local model and update its local parameters with its local dataset $D^{(k)}$. After local training is performed for E iterations, the local parameters $w_t^{(k)}$ of all clients are transmitted back to

how the optimization algorithm can improve the model's training with non-IID datasets. Further, the integration of the proposed method into a depression detection application based on sleep monitoring is explored. The main contributions are summarized as follows:

1. A novel federated optimization method with adaptive gradient is proposed for the non-IID data training case, which achieves faster convergence performance compared to commonly adopted FL algorithms, e.g., FedAVG, FedADAGRAD and FedADAM. Based on the global gradient, we propose to adaptively modify local gradients according to the weighted average of a global gradient and the local gradient, which helps lead the local parameter updates to the desired direction.
2. A systematic analysis for our proposed algorithm is provided. We analytically prove that parameter divergence of the non-IID data training case is reduced by the adaptive gradient method. By reducing the gradient divergence during local training, our proposed algorithm enables more robust training and less number of rounds to converge with non-IID clients, which saves communication costs significantly. Furthermore, the experiments validate our results by demonstrating faster convergence and higher accuracy of our proposed algorithm.
3. A novel depression detection system is implemented with the integration of FL, where depression classification is achieved by analyzing sleep data collected by portable smartphones and wearable smartwatches. Experiments demonstrate that the proposed adaptive gradient method can enable robust and effective model training.

In the rest of this article is organized as follows. The background and motivation are discussed in Sect. 2. In Sect. 3, we propose a FL optimization algorithm with adaptive gradient, and provide a theoretical convergence analysis. Following that, detailed depression detection technology with smartphone and wearable smartwatch is discussed in Sect. 4. In Sect. 5, the performance of the proposed system is evaluated by conducting a simulated experiment. Finally, Sect. 6 concludes this paper.

2 Background and Motivation

The non-IID data training problem is a common issue for IoMT applications. IoMT applications contain datasets recorded from a variety of data sources, including hospitals, schools and homes, etc. The non-IID distribution of datasets causes heterogeneous local parameter updates, which perplexes the global parameter aggregation and requires a large number of rounds to converge. Therefore, it is crucial to develop a federated optimization method for the non-IID data training to enable fast convergence and reduce communication costs. Furthermore, a system architecture for the integration of FL and IoMT is required to enable mental healthcare monitoring system.

pandemic has raised concerns not only over the public's health physically but also psychologically. According to [12], the depression ratio in China has increased during the Covid-19 period, with significant heterogeneity detected. [14] suggests that people under quarantine have a higher risk of suffering from different levels of mental problems. Thus, a timely and effective depression prevention system is demanded.

Recent studies have discovered significant association between depression and sleep behavior, such as sleep duration and sleep disturbance [16]. The findings enable the idea of a depression detecting system supported by Internet of Medical Things (IoMT) and sleep healthcare monitoring. By utilizing portable sleep monitoring devices, such as smartphones or wearable smartwatches, it is possible to apply machine learning (ML) technologies to extract significant sleep patterns for depression classification [1,17]. Meanwhile, federated learning (FL), as a newly proposed decentralized ML training paradigm, is proposed to protect users' data privacy. By integrating FL into IoMT design, there are potentially more public users willing to participate in the model training to achieve higher accuracy and sensitivity [15].

Many implementations of FL enabled healthcare application are proposed to solve medical diagnosis problems, e.g., [10], [5] and [8] try to enable Covid-19 diagnosis with FL framework and [19] applies FL into human emotion detection. However, the majority of the researches assume Independently-and-Identically-Distributed (IID) data across the participants, which usually is not true for many real-life applications [13]. In most reality scenarios, the datasets across the participants are non-Independently-and-Identically-Distributed (non-IID). For example, the datasets from hospital generally contain more positive samples than datasets collected from community users. The non-IID datasets can significantly downgrade the convergence during model training. To achieve desired model accuracy, more training epochs are required compared to IID training, which also results in higher communication costs. Therefore, to alleviate network overload, it is important to develop a novel federated learning method with faster convergence rate under non-IID training.

Currently, there are a number of researches trying to improve the non-IID convergence performance of FL from a different perspective. [21] proposed the weight divergence model to quantify the effect of non-IID datasets, and suggested to improve convergence by sharing a subset of data, which violates the privacy-preserving principle of FL. [11] summarized a general framework for adaptive federated optimizations, including FedADAM, and FedYOGI [7], to improve convergence by changing learning rate throughout the training process. [18] also assign learning rate adaptively according to the clients' contribution. However, there are few articles discussing the optimization method with adaptive gradient, where gradients in back-propagation is adjustable to enhance convergence during training phase. While changing learning rate can adjust the "speed" of optimization, adaptive gradient adjusts the "direction" of global and local FL optimization, which potentially accelerates the convergence.

In this article, we develop and explore a novel adaptive gradient optimization method to alleviate the effect of non-IID data training. We also investigate

Fast Convergence Federated Learning with Adaptive Gradient: An Application to Mental Healthcare Monitoring System

Junqiao Fan[1,3], Xuehe Wang[1(✉)], and Yuzhu Hu[2]

[1] The School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai 519082, China
wangxuehe@mail.sysu.edu.cn
[2] The School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510006, China
huyzh27@mail2.sysu.edu.cn
[3] The School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
fanj0019@e.ntu.edu.sg

Abstract. Nowadays, there is increasing demand for mental health monitoring systems to enable disease diagnoses, such as anxiety and depression. However, the privacy concerns for sensitive data impede its wide adoption. To protect data privacy, federated learning (FL) is proposed to enable decentralized collaborative model learning without sharing sensitive data. Though, FL training process can be slowed with the non-Independent-and-Identically-Distributed (non-IID) datasets across participating clients, causing extra communication costs. In this paper, we propose the FL adaptive gradient optimization method to accelerate the convergence under the context of non-IID training. As the reference direction for parameter update, the gradient has a great impact on the convergence performance throughout the training. By adaptively modifying the local gradients according to the global gradient, we reduce the local parameter divergence to enable robust training and fast convergence. Meanwhile, as an application to our FL optimization algorithm, a novel sleep monitoring system is proposed to detect potential depression. Experiments demonstrate that with our proposed method, faster convergence and higher accuracy can be realized compared to commonly adopted Federated Averaging (FedAVG) and other adaptive optimization methods, which effectively save communication costs.

Keywords: Adaptive Gradient · Federated Learning · Non-IID Datasets · Depression Detection

1 Introduction

Since the report of the first Covid-19 case, the pandemic has spread for more than two years, overwhelming the healthcare system all over the world. The sustaining

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 255–275, 2024.
https://doi.org/10.1007/978-3-031-65126-7_24

sources and devices are employed for performance validation, and the results well demonstrate the effectiveness and advantages of the algorithm.

References

1. Chen, C.Y.: Automated ECG classification based on 1D deep learning network. Methods **202**, 127–135 (2022)
2. Ebrahimzadeh, A., Shakiba, B., Khazaee, A.: Detection of electrocardiogram signals using an efficient method. Appl. Soft Comput. **22**, 108–117 (2014)
3. Fan, X., Yao, Q., Cai, Y., Miao, F., Sun, F., Li, Y.: Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. IEEE J. Biomed. Health Inf. **22**(6), 1744–1753 (2018)
4. Guo, W., Zhang, Y., Yang, J., Yuan, X.: Re-attention for visual question answering. IEEE Trans. Image Process. **30**, 6730–6743 (2021)
5. Huang, Y., Li, H., Yu, X.: A multiview feature fusion model for heartbeat classification. Physiol. Meas. **42**(6), 065003 (2021)
6. Jekova, I., Bortolan, G., Christov, I.: Assessment and comparison of different methods for heartbeat classification. Med. Eng. Phys. **30**(2), 248–257 (2008)
7. Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J., Shao, L.: Understanding more about human and machine attention in deep neural networks. IEEE Trans. Multimedia **23**, 2086–2099 (2020)
8. Liu, P., Sun, X., Han, Y., He, Z., Zhang, W., Wu, C.: Arrhythmia classification of LSTM autoencoder based on time series anomaly detection. Biomed. Signal Process. Control **71**, 103228 (2022)
9. Mei, Z., Gu, X., Chen, H., Chen, W.: Automatic atrial fibrillation detection based on heart rate variability and spectral features. IEEE Access **6**, 53566–53575 (2018)
10. Oh, S.L., Ng, E.Y., San Tan, R., Acharya, U.R.: Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Comput. Biol. Med. **102**, 278–287 (2018)
11. Oh, S.L., Ng, E.Y., San Tan, R., Acharya, U.R.: Automated beat-wise arrhythmia diagnosis using modified U-Net on extended electrocardiographic recordings with heterogeneous arrhythmia types. Comput. Biol. Med. **105**, 92–101 (2019)
12. Pourbabaee, B., Roshtkhari, M.J., Khorasani, K.: Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Trans. Syst. Man Cybern. Syst. **48**(12), 2095–2104 (2018)
13. Mukhopadhyay, S.K., Mitra, S., Mitra, M.: An ECG signal compression technique using ASCII character encoding - sciencedirect. Measurement **45**(6), 1651–1660 (2012)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
16. Wang, J., et al.: Automated ECG classification using a non-local convolutional block attention module. Comput. Methods Programs Biomed. **203**, 106006 (2021)
17. Yeh, Y.C., Wang, W.J., Chiou, C.W.: Cardiac arrhythmia diagnosis method using linear discriminant analysis on ECG signals. Measurement **42**(5), 778–789 (2009)

Table 7. Results of ablation experiments on the 2017 PCCD

Methods	*Acc*	*Sen*	*Pre*	F1
VGG11	0.914	0.647	0.800	0.689
SEVGG11	0.972	0.901	0.898	0.898
VGG11-LSTM	0.952	0.841	0.950	0.885
VGG13-LSTM	0.945	0.868	0.893	0.880
VGG16-LSTM	0.941	0.798	0.950	0.853
SEVGG11-LSTM	0.952	0.877	0.913	0.894
Proposed (SEVGG11-LSTM-O)	**0.962**	**0.931**	**0.914**	**0.922**

4.3 Discussions

In order to determine the optimized network, convolutional networks of various depths are created. In specific, networks with 11, 13 and 16 convolutional layers are compared on the MIT-BIH arrhythmia database. As listed in Table 6, with the increasing counts of convolutional layers, the performance metrics decrease. Compared with the basic models, VGG11-LSTM is finalized for further improvement. When enhanced by the attention mechanism, the overall F1 score increases, indicating that the SEVGG11-LSTM has better classification performance. After oversampling is implemented, the classification performance is further improved, and the strong competitiveness on the MIT-BIH arrhythmia database is advantageous to that of the state-of-the-art algorithms.

Similar ablation experiments are also performed on the 2017 PCCD, and the performance records are comparatively listed in Table 7. Similarly, the F1 score decreases with the increasing counts of the convolutional layers, and is increased after introducing the attention mechanism. When oversampling is further used, the F1 score reaches the optimum value. Tables 6 and 7 well validate the great potentials of attention mechanism and oversampling for ECG classification.

Compared with the validation on a single database, two databases are employed in our paper in a "dual-centers" fashion. More participants are involved to avoid the limitation of a single database, and the conclusion is therefore more reliable. In addition, researchers can draw on the wisdom of the masses in the two databases to improve clinical trials.

5 Conclusion

This paper demonstrates a novel deep learning algorithm for ECG classification. It makes use of the combined convolutional and recurrent neural network for classifying ECG as well as the attention mechanism to assign weights. The input ECG signals are sequentially segmented and normalized. After that, the preprocessed signals are fed into the combined VGG and LSTM network for feature extraction and classification. The core network contains an attention mechanism that increases the weight of significant features. Two databases from different

Table 4. Comparison on MIT-BIH arrhythmia database

Methods	*Acc*	*Sen*	*Pre*	F1
LSTM [8]	0.986	0.980	0.976	0.978
CNN+LSTM [10]	0.981	0.975	–	–
CNN+LSTM+HOS [5]	0.989	0.965	0.969	0.967
U-Net [11]	0.973	–	–	–
RBF-BA [2]	0.952	0.956	0.906	0.930
LDA [17]	0.962	0.925	0.947	0.936
KNN [6]	–	0.809	0.769	0.788
Proposed method	**0.996**	**0.984**	**0.988**	**0.986**

Table 5. Performance comparison on 2017 PCCD

Methods	*Acc*	*Sen*	*Pre*	F1
CNN [12]	0.899	0.671	0.590	0.628
VGGNet [14]	0.976	0.909	0.903	0.906
MS-CNN [3]	0.977	0.943	0.886	0.914
BT [9]	0.966	0.832	–	–
Proposed	**0.962**	**0.931**	**0.914**	**0.922**

Table 6. Results of ablation experiments on the MIT-BIH arrhythmia database

Methods	*Acc*	*Sen*	*Pre*	F1
VGG11	0.972	0.901	0.898	0.898
VGG13	0.979	0.873	0.943	0.895
VGG16	0.963	0.810	0.858	0.824
VGG11-LSTM	0.980	0.911	0.911	0.911
VGG13-LSTM	0.980	0.894	0.919	0.906
VGG16-LSTM	0.977	0.867	0.911	0.888
SEVGG11-LSTM	0.980	0.933	0.909	0.919
Proposed (SEVGG11-LSTM-O)	**0.996**	**0.984**	**0.988**	**0.986**

Table 2. Performance on MIT-BIH arrhythmia database

Types	*Acc*	*Sen*	*Pre*	F1
A	0.994	0.955	0.955	0.955
L	0.999	0.991	1.000	0.995
N	0.992	0.997	0.989	0.993
R	1.000	1.000	1.000	1.000
V	0.994	0.978	0.997	0.987
Overall	0.996	0.984	0.988	0.986

Table 3. Performance on 2017 PCCD

Types	*Acc*	*Sen*	*Pre*	F1
N	0.962	0.974	0.982	0.978
A	0.962	0.888	0.845	0.866
Overall	0.962	0.931	0.914	0.922

4.2 Comparison Results

In order to further validate the performance of our proposal, some state-of-the-art algorithms are employed for comparison. For fair comparison, only the algorithms developed for the same classification problems and tested on the same database are introduced. Regarding the MIT-BIH arrhythmia database, the selected models for comparison are as follows. Liu *et al.* [8] proposed a model based on LSTM to obtain time-series features of ECG. In [10], the CNN layer is used to extract feature maps, and the LSTM layer captures the temporal dynamics. The model [5] is composed of a eight-layers CNN, a eight-layers LSTM and a fully connected layer. Detailed performance records of the compared models are listed in Table 4. The F1 score, *Acc*, *Sen* and *Pre* of the proposed model are higher than those of the compared models. The advantages of our proposal are therefore validated.

In addition, the two-class classification performance of wearable ECG is also compared on the 2017 PCCD. The results are listed in Table 5. Compared with the peer algorithms that have reported their F1 performance, the proposed model has the highest F1 score of 0.922. The comparison results further verify the good performance of the proposed model.

Table 1. Oversampling on MIT-BIH arrhythmia database

Type	Before oversampling	After oversampling
N	6735	6735
V	3005	6010
L	1202	7212
R	1179	7074
A	771	6939
Total	12892	33970

3.3 Cross Validation

Under the condition of limited size of dataset, 10-fold cross validation is able to achieve multiple random partitioning of the training set and test set. The original dataset is divided into 10 equal sized parts, of which nine parts are considered as training dataset and the other one is used for testing. 10-fold cross validation avoids overlap between the training and testing datasets. In each iteration, the balanced training data set is used to get the optimized parameters of the model, and then the corresponding test set is employed for performance evaluation. After 10 iterations, the results from each iteration are combined to yield the average performance of the model.

4 Results and Discussion

4.1 Overall Performance

Table 2 lists the results when using the proposed algorithm to classify various heart rhythms in the MIT-BIH arrhythmia database. The overall *Acc*, *Sen*, *Pre* and F1 values are 0.996, 0.984, 0.988, and 0.986, respectively. By considering both *Sen* and *Pre*, the F1 score better demonstrate the overall performance of the proposed algorithm. For A, L, N, R, and V types of rhythms, the achieved F1 scores of our method are 0.955, 0.995, 0.993, 1.000, and 0.987, respectively. For R rhythm, all of the performance metrics are optimal, that is, 1, which indicates the good performance in identifying R rhythms. In addition, all of the *Acc* values are higher than 0.992, demonstrating that the classification performance of our algorithm is very satisfactory.

In order to verify the universality of the proposed algorithm, it is further applied to 2017 PCCD, and the performance records are listed in Table 3. As can be observed, the *Acc*, *Sen*, *Pre* and overall F1 records are 0.962, 0.931, 0.914, and 0.922, respectively. As concluded above, the proposed algorithm shows satisfactory performance on both databases.

The fully connected layers reassembles the local features into a complete graph through the weight matrix. The representative features are integrated into a value, which has the advantage of reducing the influence of feature locations on classification results, and improving the robustness of the whole network. In addition, multiple fully connected layers are connected, so that the ability of nonlinear expression is improved.

3 Experiment Configuration

3.1 Database and Performance Metrics

Two different databases are employed for performance evaluation, the MIT-BIH arrhythmia database and the 2017 PhysioNet/CinC Challeng database (2017 PCCD). The MIT-BIH arrhythmia database includes 48 recordings, each of which is sampled at 360 Hz and contains a series of two-leads ECG data. Based on the proportion of various arrhythmias in the clinic, normal rhythm (N), left bundle branch block (R), right bundle branch block (L), ventricular precontraction (V) and atrial premature contractions (A) are selected for classification. On the other hand, there are 8528 single-lead ECG records in 2017 PCCD. All of them are collected by wearable devices at a sampling frequency of 300 Hz. The types of normal (N) and atrial fibrillation (AF) samples are selected to validate the performance of our model. Because of the data imbalance problem, samples in the training set are balanced by oversampling, as detailed in Sect. 3.2.

In this paper, accuracy (Acc), sensitivity (Sen), precision (Pre), and F1 score are employed for performance evaluation. The Acc is the proportion of correctly classified samples, Sen denotes the recognition ability of positive examples, Pre represents the proportion of correct predictions in the samples with positive predictions, and F1 score is the weighted average of Sen and Pre, respectively.

All of the models are implemented on the framework Tensorflow-gpu 2.6.2, using the Windows-11 operation system. The algorithm is deployed on a workstation with Intel(R) Core(TM) i7-12700H at 2.7 GHz, and an RTX 3060 GPU with a 14 GB memory.

3.2 Oversampling

Oversampling is adopted to solve the data imbalance problem, by increasing the number of certain arrhythmia samples. In this paper, the random oversampling method is used. It works as follows. 1) Taking the class with the largest number of samples as the benchmark, calculate the multiple of the benchmark class and the minority classes. 2) If the multiple is greater than 2, the minority classes are copied by corresponding multiples. 3) If the multiple is less than 2, a certain proportion (multiple minus one) of samples from the minority classes are randomly selected for duplication. Taking MIT-BIH arrhythmia database as an example, and samples counts before and after oversampling is listed in Table 1.

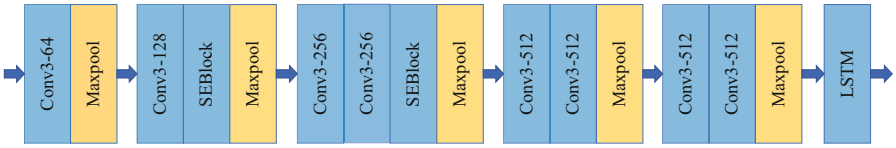


Fig. 2. Core network of the proposed model.

pooling layers. The numbers of the convolutional layers in each part are 1, 1, 2, 2, and 2, and the sizes of convolutional kernels are 64, 128, 256, 512, and 512, respectively. Each part of convolutional layer is followed by a maximum pooling layer to reduce the data length and computational burden of the model. A LSTM layer is connected after the convolutional layers and maximum pooling layers of VGGNet to avoid the gradient disappearance and gradient explosion. In addition, two SE blocks are added to the fusion model of VGGNet and LSTM, which strengthen the features of R peak. After that, the output of LSTM is given to the fully connected layer for ECG classification. During the model training process, we selected the following optimal parameters: optimizer = "Adam", learning rate = 0.001, epoch = 50, and batch size = 32 for improved performance.

The SENet, a kind of attention mechanism, is presented in the form of the SE block. It contains the squeeze and excitation modules. Squeeze operation is carried out after the traditional convolution module, that is, all spatial features in a channel are encoded into a global feature. It is defined as

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), z_c \in R^c, \tag{2}$$

where Z_c refers to the c-th element of the squeezed channels, u_c represents the c-th channel of the input, F_{sq} is the squeeze function, and H and W are the height and width of the input, respectively. After that, two fully connected layers are employed for better generalization, an activation function is used to obtain the channel-wise dependencies. The process is described by

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \tag{3}$$

$$\tilde{x}_c = Fscale(u_c, s_c) = s_c u_c, \tag{4}$$

where F_{ex} denotes the excitation function, W_1 and W_2 represent the widths of the inputs in dimensionality reduction and increasing layers, δ denotes the ReLU activation function, σ is the sigmoid function, and $Fscale$ represents channel-wise multiplication.

2.4 Classification Module

At the end of our model, a 3-layers fully connected layer of the VGGNet is selected, and the softmax function is used to realize multi-classification problem.

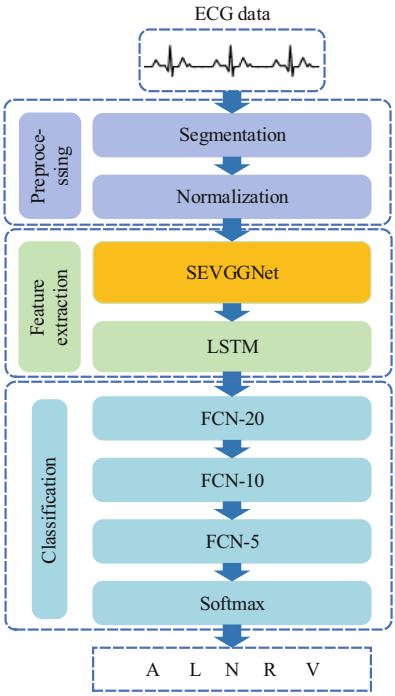


Fig. 1. Block diagram of the proposed model.

2.2 Data Preprocessing

The data preprocessing module primarily concentrates on two key tasks: signal segmentation and normalization. To alleviate computational overhead, ECG signals are segmented into multiple 10-s segments. Since ECG segment amplitudes can significantly vary owing to individual differences and lead positions, normalization becomes crucial. This normalization process involves adjusting the amplitudes, ensuring uniformity, and improving data consistency for further analysis and model training. The normalization is conducted by

$$Normalized(X) = \frac{X - \bar{X}}{S}, \tag{1}$$

where X is value of each record, $\bar{X}$ and S refer to the average and standard deviation of all the records, respectively.

2.3 Feature Extraction

The core network of the proposed algorithm is illustrated in Fig. 2. It includes eight convolutional layers, one LSTM layer, five maximum pooling layers, and two SE blocks. The convolutional layers are split into five parts by maximum

the classification performance. This advantage has been widely exploited in the fields of neural machine translation and computer vision [4,15]. It is therefore plausible to infer that adding attention mechanisms to the ECG classification model is likely to achieve performance improvement. [16] presented a CNN model with a non-local convolutional block attention module capable of distinguishing relationships between local and global segments. In addition, previous works [1,10] have demonstrated the great potentials of the combined structure of convolutional neural network (CNN) and rerrent neural network (RNN) in the ECG classification problems. As a popular CNN, the visual geometry group network (VGGNet) [14] has more nonlinear transformation and enhanced ability to learn features. On the other hand, long short-term memory neural network (LSTM) is an improved RNN, and it is able to avoid the problems of gradient explosion or gradient disappearance.

This paper proposes the SEVGGNet-LSTM model for ECG classification. Our proposal takes advantages of the combined structure of convolutional and recurrent neural network for ECG classification, and also makes full use of the attention mechanism's weight allocation capability. First, ECG data records are split into 10 s segments. After that, the amplitudes of the divided segments are normalized. The preprocessed ECG signals are then fed into the SEVGGNet-LSTM, which is a sequential combination of VGG and LSTM, with an attention mechanism (squeeze-and-excitation block, SE block) to increases the weight of important features. Finally, ECG classification is completed by the fully connected layers. Two databases from different sources and devices are employed for performance validation, and the experimental results well demonstrate the effectiveness and advantages of the proposed algorithm.

Our main contributions are as follows. 1) A fused deep convolutional neural network is constructed for ECG classification. 2) The combined convolutional and recurrent neural network and the weight allocation capability of the attention mechanism are exploited for performance promotion. 3) Two databases with ECG records collected by different devices are employed for performance evaluation.

2 The Proposed Method

2.1 Architecture

The proposed algorithm mainly consists of the preprocessing module, feature extraction module, and classification module, as illustrated in Fig. 1. The preprocessing module includes signal segmentation and normalization. The ECG signals are split into 10 s segments, after that a normalization technique is used to normalize their amplitudes. In the feature extraction module, SEVGGNet and LSTM constitute the deep neural network, and the preprocessed ECG signals are fed into the deep neural network for feature extraction. Final, a three-fully connected layer and a softmax layer constitute the classification module to realize the multi-classification problem.

SEVGGNet-LSTM: A Fused Deep Learning Model for ECG Classification

Tongyue He[1], Yiming Chen[1], Bo Fang[1], and Junxin Chen[2(✉)]

[1] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110167, China
[2] School of Software, Dalian University of Technology, Dalian 116620, China
junxinchen@ieee.org

Abstract. With the dramatic progress of smart sensing and wearable device, continuous and real-time acquisition of electrocardiograph (ECG) tends to be realized in a convenient way. Data mining of ECG signals has therefore been extensively researched, among which ECG classification is a hot topic. This paper presents a fused deep learning algorithm for ECG classification. It takes advantages of the combined convolutional and recurrent neural network for ECG classification, and the weight allocation capability of attention mechanism. The input ECG signals are firstly segmented and normalized, and then fed into the combined VGG and LSTM network for feature extraction and classification. An attention mechanism (SE block) is embedded into the core network for increasing the weight of important features. Two databases from different sources and devices are employed for performance validation, and the results well demonstrate the effectiveness and robustness of the proposed algorithm for classifying ECG signals obtained from wearable ECG devices and professional medical equipment.

Keywords: ECG classification · Deep learaning · Arrhythmia · Attention mechanism

1 Introduction

The electrocardiogram (ECG) analysis is an important non-invasive means for diagnosing and evaluating cardiac diseases [13]. However, the inherent complexity of arrhythmia often brings difficulties to medical workers in ECG classification, and may lead to mis-diagnosis. With the popularity of artificial intelligence (AI), developing computer-aided ECG classification is in a high demand.

Deep learning based solution for ECG classification has drawn world-wide concerns in recent years. It is able to automatically extract features, and hence get rid of the dependence of manual feature extraction in traditional machine learning methods. About features extraction, as reported in [7], attention mechanism is able to increase the weight of important features and further promote

Supported by the National Natural Science Foundation of China (No. 62171114).

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 245–254, 2024.
https://doi.org/10.1007/978-3-031-65126-7_23

17. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
18. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
19. Zhou, D., Wei, X.X.: Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. Adv. Neural. Inf. Process. Syst. **33**, 7234–7247 (2020)

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Alarcao, S.M., Fonseca, M.J.: Emotions recognition using EEG signals: a survey. IEEE Trans. Affect. Comput. **10**(3), 374–393 (2017)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Process. Syst. **27**, 1–9 (2014)
5. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
6. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural Inf. Process. Syst. **31**, 1–11 (2018)
7. Gao, Y., Archer, E.W., Paninski, L., Cunningham, J.P.: Linear dynamical neural population models through nonlinear embeddings. Adv. Neural Inf. Process. Syst. **29** (2016)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Jazayeri, M., Ostojic, S.: Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Curr. Opin. Neurobiol. **70**, 113–120 (2021)
10. Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. **8**, 94 (2014)
11. Pandarinath, C., et al.: Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods **15**(10), 805–815 (2018)
12. Quitadamo, L.R., et al.: Support vector machines to detect physiological patterns for EEG and EMG-based human-computer interaction: a review. J. Neural Eng. **14**(1), 011001 (2017)
13. Rossini, P.M., et al.: Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis: report from the IFCN-sponsored panel of experts. Clin. Neurophysiol. **131**(6), 1287–1310 (2020)
14. Sadtler, P.T., et al.: Neural constraints on learning. Nature **512**(7515), 423–426 (2014)
15. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 1–9 (2023)
16. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)

Table 1. Decoding accuracy from different embeddings to different labels with different decoders

Decoding Label	Emotion		Subject	
	DNN	CNN	DNN	CNN
Discovery-driven embedding-8D	0.327	0.357	0.722	0.958
Discovery-driven embedding-16D	0.355	0.355	0.746	0.962
Discovery-driven embedding-32D	0.346	0.340	0.785	**0.962**
Emotion-guided embedding-8D	0.417	0.413	0.427	0.108
Emotion-guided embedding-16D	0.420	0.410	0.357	0.091
Emotion-guided embedding-32D	0.394	0.409	0.304	0.091
Subject-guided embedding-8D	0.339	0.341	0.979	**0.995**
Subject-guided embedding-16D	0.327	0.340	0.980	**0.995**
Subject-guided embedding-32D	0.343	0.331	0.978	**0.996**

the identity of a person. Moreover, in the self-supervised learning case where no labels are provided, the accuracy of identifying a subject is still 96.2%. Recall that the input feature to generate the embedding is only EEG collected within 0.5 s, which is quite a short time. The feasibility of input features and the close-to-perfect classification accuracy indicate the promising future of our method be to applied to identify persons in various HCI scenarios.

The accuracy of detecting emotions is low, probably because the emotional changes are rarely reflected in temporal voltage features. Including more EEG features in the inputs may improve emotion detection performance.

4 Conclusion and Future Work

In this paper, we propose to use contrastive learning to generate low-dimension EEG latent embeddings that are consistent and identifiable. The contrastive learning encoder can be trained in either the supervised or self-supervised manner and the encoder trained in both manners can have very high decoding accuracy. This indicates the potential of using the EEG latent embeddings for various downstream tasks. Excitingly, we can use the EEG latent embeddings to identify different persons at close-to-perfect accuracy of 99.6% with EEG input data with only a 0.5-s window. Our method can be promisingly used in emerging modern HCI devices and applications, e.g., automatically connecting people to their roles in video games via EEG-capable AR devices.

In the future, we will try to integrate our EEG latent embedding method into industrial HCI solutions for entertainment and health management. To achieve this, we need to improve our method for a wider range of downstream applications, which may require exploring more informative EEG features as inputs. We will also verify it with various datasets and in EEG devices of different specifications.

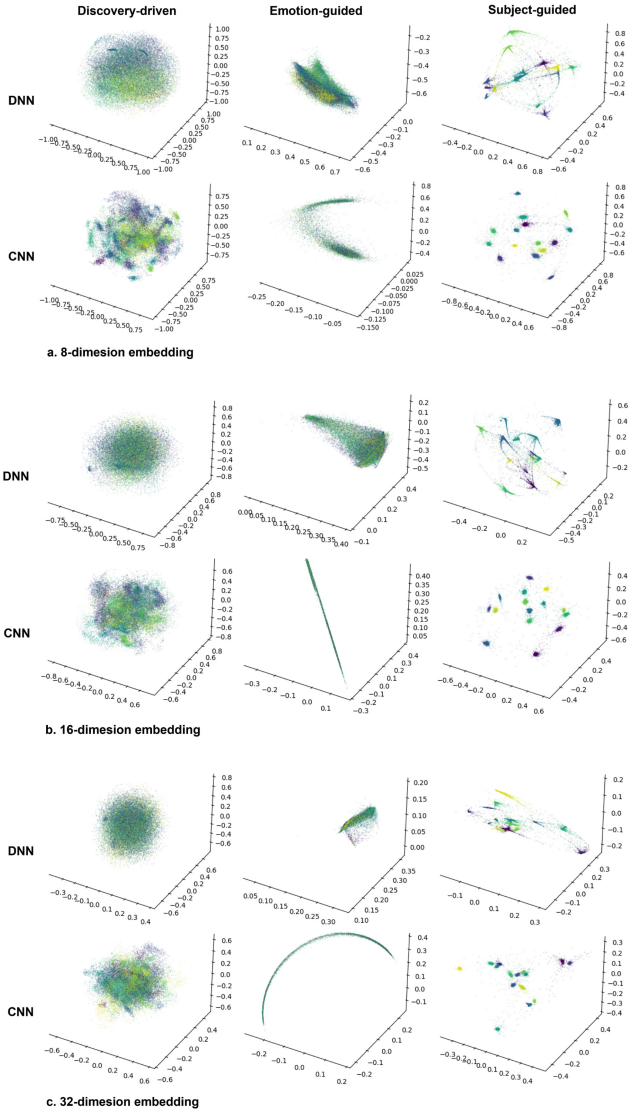


Fig. 3. The first three dimensions of DNN or CNN-learned embeddings of 8D, 16D, and 32D by discovery-driven, emotion-label-guided, and subject-label-guided contrastive learning, respectively.

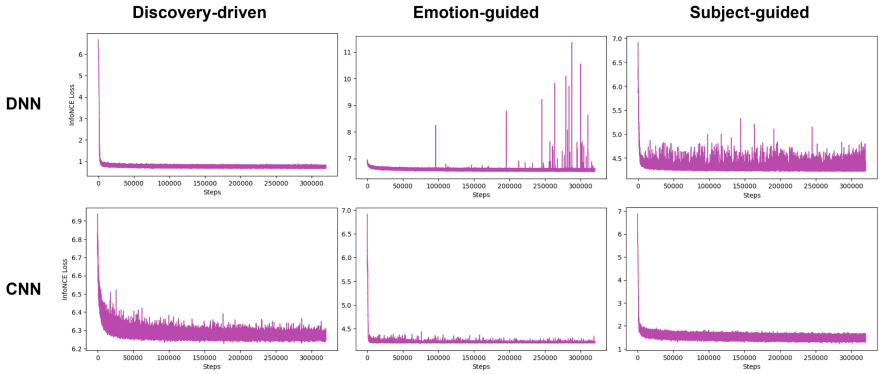


Fig. 2. The convergence performance of the DNN and CNN encoders using contrastive learning in discovery-driven, emotion-label-guided, and subject-label-guided manners, respectively

with different labels, and with different output dimensions (Fig. 3). The related labels of the data points can be distinguished by colors. From the figures, we can see that in the CNN encoder cases, where points of the same color or similar colors tend to cluster into color clouds according to some patterns, generally generate more identifiable embeddings than DNN ones, where color point clouds are more chaotic. The reason is that the CNN model considers the sequential context of the input data.

Specifically, in the CNN subject-guided cases even when the output dimension is as low as 8, we can clearly recognize 15 color clouds, which corresponds exactly to 15 subjects in the dataset. It shows that we can use a low-dimension EEG latent embedding to distinguish the identities of people, which is an exciting result and can greatly benefit new applications of HCI.

For CNN encoders trained in a discovery-driven manner, color clouds are formed in a more complex pattern and are not completely separated from each other. It indicates that the embeddings are generally informative and have great potential to be applied to various tasks. The higher the embedding dimension, the clearer the patterns and the more identifiable the color clouds.

3.3 Decoding Accuracy

The results of top-1 accuracy of decoding embeddings from different encoders to different labels are shown in Table 1. The accuracy of distinguishing different subjects has reached as high as 99.6% using 32-dimension latent embedding generated by the CNN encoder via supervised contrastive learning. Even for these latent embeddings of dimensions 8 and 16, the accuracy is as high as 99.5%, which is almost the same as the 32-dimension ones. It shows that we can compress EEG data to float vectors as small as 8 dimensions to represent

Positive and negative samples are taken from a minibatch of the training input. The identification of positive and negative samples depends on the scientific problem we are solving. In the discovery-driven manner when no label is provided, samples near x along the timeline are positive and those far away from x along the timeline are negative. While in the hypothesis manner, specific labels are provided and the learning process is similar to supervised contrastive learning in concept. Samples with the same label as that of x are positive, while those with different labels from x are negative. We train different encoder models without any label, with emotion labels, and with subject labels, respectively, and compare their convergence, latent visualization, and decoding performance, respectively.

We test the encoder with two different neural network structures, where one is a five-layer 1D convolutional network with skipping connections (CNN) and the other is a four-layer fully connected network (DNN). Perceptrons in both networks are activated by GELU functions. The mini-batch size of the input is 1,024, the learning rate is 0.001. The encoder output dimension can be 8, 16, and 32, and the number of mini-batch training iterations is 10,000 times the encoder output dimension.

Embedding Decoding. The decoder is a classification or regression model that fits the EEG latent embedding to the task-specific labels. We use the K-nearest neighbors (KNN) method (where $k = 5$) as the model and emotions and subjects as the labels, respectively. Both types of labels are discrete, where emotion labels have 3 values and subjects have 15. The embedding decoder is trained separately from the embedding encoder, where the encoder output embeddings are generated from the training input features by the well-learned encoder and the corresponding labels are the training inputs of the decoder.

3 Results and Discussions

3.1 Contrastive Learning Convergence

We explore the convergence performance of the encoder using contrastive learning in discovery-driven and hypothesis manners. Figure 2 shows the NCE loss of the DNN and CNN encoders to encode EEG to 32-dimension latent embeddings provided without labels, with emotion labels, and with subject labels, respectively. The encoder tends to converge in all cases, which indicates that the encoder will generate consistent latent embeddings for EEG features that are considered similar. The NCE losses of encoding 8-dimension and 16 dimension latent embeddings also converge but jitter in a smaller magnitude, which is not depicted here to prevent redundancy.

3.2 Embedding Visualization

Low-dimension representations easily can be visualized to help the interpretation [9]. We visualize the first three dimensions of the testing embeddings generated by encoders of different neural network structures, contrastively learned

2 Method

Model. The whole procedure of the model is depicted in Fig. 1. We use an EEG dataset of emotion detection with neural activity with continuous long-lasting stimuli (Fig. 1.a). The encoder learns from time-domain features via a neural network with contrastive learning either in a discovery-driven manner or guided by task-specific labels and outputs EEG latent embeddings (Fig. 1.b). The decoder classifies the latent embeddings into different task-specific labels (Fig. 1.c).

Dataset. We use the SEED [18] dataset, which is a widely used open dataset designed for exploring the relationship between EEG and emotions. The dataset comprises EEG data from 15 people subjects joining a 3-session testing, with each testing session stimulated by watching 15 movie clips of a total of about 3600 s, which are continuous lasting stimuli. The movie stimuli are related to 3 emotions, i.e., positive, neutral, and negative. The EEG signals are collected by 62 electrode channels, down-sampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data grouped by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

Most work extracts EEG features from the frequency-domain representation [1,10,16], but some recent work [5,6] shows that interpretable latent embeddings can be learned from time-domain representation. We further down-sample the time-domain signals to 2 Hz so that within every 0.5 s, we can extract 5 voltage features for each of the 62 channels, namely the maximum, minimum, mean, median, and standard deviation. Finally, each encoder input is a 310-dimension vector with a 0.5-s time window, the volume of the training set is $\mathbb{R}^{270855\times310}$, and that of the testing set is $\mathbb{R}^{35280\times310}$.

Contrastive Learned Embeddings. The encoder is a non-linear convolutional neural network (CNN) or deep neural network (DNN) that applies contrastive learning optimizing the NCE loss, which follows a similar procedure as in [15].

For the input features x and y, where y is a positive or negative contrastive sample of x, let $p(x)$ be the probability density function of x, $p(y|x)$ and $q(y|x)$ be the probability density function of the positive and negative samples conditioned on x, respectively. Encoding x and y can be represented by a function f with normalized outputs, and $f(x)$ and $f(y)$ are the normalized latent embeddings, respectively. We use the dot product of $f(x)$ and $f(y)$ adjusted with a temperature parameter τ as the similarity function between these two latent embeddings, which is denoted as $\psi(x, y) = f(x)^T f(y)/\tau$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{x\sim p(x), y_+\sim p(y|x)\\ y_1,y_2,\ldots,y_n\sim q(y|x)}} [-\psi(x, y_+) + log \sum_{i=1}^{n} e^{\psi(x,y_i)}].$$

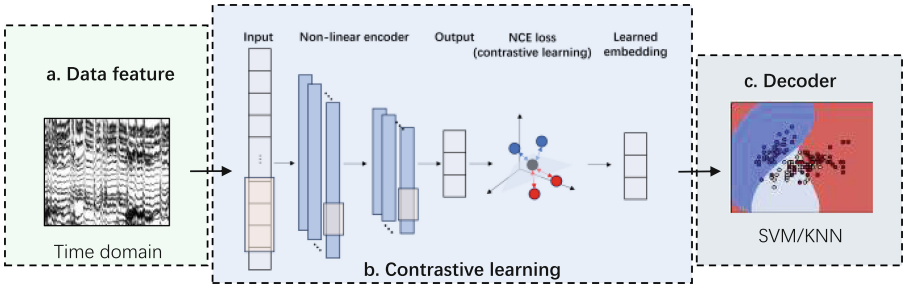


Fig. 1. The procedure of encoding time-domain EEG into embeddings by contrastive learning and decoding the learned embeddings

(HCI) and affective computing [10,16]. Encoding EEG data into consistent and identifiable latent embeddings [5,6] can greatly extend the application of EEG in various downstream tasks. The generated embeddings facilitate EEG applications by incorporating valuable information on EEG characteristics and filtering some irrelevant noise in the reduced-dimension representation. Most existing work on encoding EEG [7,14] uses supervised learning methods that depend on specific tasks, which limits its representation ability and application to other tasks. Recently, unsupervised learning [2,4,17] and self-supervised learning [3,8,11,19] methods have shown their capability of learning discriminative latent embeddings which can be generally used for downstream tasks. For example, [15] shows that contrastive learning can generate consistent and identifiable neural latent embeddings from instant spike-stimulated signals in a discovery-driven or hypothesis manner. However, neural signals are usually affected by continuous long-lasting stimuli, e.g., disease and environmental influence. The ability of contrastive learning to generate embeddings from continuous long-lasting stimulated EEG data is unknown.

In this paper, we investigate the ability of contrastive learning to generate consistent and identifiable EEG latent embeddings We use features of the time-domain representation of EEG as the input of a learnable encoder, which optimized by the noise-contrastive estimation [8] (NCE) in both the discovery-driven (self-supervised) and hypothesis (supervised) manners. We explore properties of consistency and interpretability by testing the convergence performance of the model and the visualization effect of the latent embeddings, respectively. We also verify the identifiablility of the embeddings by exploring the decoding performance of different downstream tasks. We find that encoding the time-domain features by contrastive learning can generate general EEG latent embeddings for some downstream applications. Excitingly, using the EEG latent embeddings that are encoded by 0.5-s time window features, the accuracy of recognizing the identities different persons is 96.2% in the self-supervised case and as high as 99.6% in the supervised case. The close-to-perfect decoding performance proves the potential of applying our method to emerging HCI and other EEG-related scenarios.

Contrastive Learning Consistent and Identifiable Latent Embeddings for EEG

Feng Liang[1,2(✉)], Zhen Zhang[1,2,3], Jiawei Mo[4], and Wenxin Hu[1,2]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{fliang,huwenxin}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
zhangzhen19@lzu.edu.cn
[4] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Extracting informative EEG data into low-dimension latent embeddings is important for storing and analyzing these neuron signals and applying them to various applications, such as modern human-computer interaction (HCI) techniques. We use the contrastive learning algorithm on time-domain features of EEG in both discovery-driven (self-supervised) and hypothesis (supervised) manners to encode the EEG data into latent embeddings that are proven consistent and identifiable. The self-supervised embeddings have the potential to be used for a range of downstream tasks, while the supervised embeddings have very high decoding accuracy for specific tasks. With embeddings encoded from EEG features collected within every 0.5-s window, the accuracy of recognizing the identities of persons by decoding the self-supervised and supervised embeddings is as high as 96.2% and 99.6%, respectively. Our method and results can promote new HCI techniques, e.g., automatically connecting users to their roles in AR games once they wear EEG-capable devices. The source code is available at: https://www.github.com/liangfengsid/timeEegContrastive.

Keywords: EEG · Contrastive learning · Latent embedding · Neural representation · Identity recognition

1 Introduction

Electroencephalography (EEG), the technique that intensively collects time-series electronic signals from the scalp, is widely used in clinical screening [13] and has a great potential application in modern human-computer interaction [12]

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 236–244, 2024.
https://doi.org/10.1007/978-3-031-65126-7_22

11. Li, Y., et al.: A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Trans. Cogn. Dev. Syst. **13**(2), 354–367 (2020)
12. Lin, Y.P., et al.: EEG-based emotion recognition in music listening. IEEE Trans. Biomed. Eng. **57**(7), 1798–1806 (2010)
13. Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. **8**, 94 (2014)
14. Pandarinath, C., et al.: Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods **15**(10), 805–815 (2018)
15. Pang, J.C., et al.: Geometric constraints on human brain function. Nature **618**, 566–574 (2023)
16. Rossini, P.M., et al.: Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis: report from the IFCN-sponsored panel of experts. Clin. Neurophysiol. **131**(6), 1287–1310 (2020)
17. Sadtler, P.T., et al.: Neural constraints on learning. Nature **512**(7515), 423–426 (2014)
18. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 360–368 (2023)
19. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)
20. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
21. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
22. Zhou, D., Wei, X.X.: Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. Adv. Neural. Inf. Process. Syst. **33**, 7234–7247 (2020)

5 Conclusion

In this paper, we explore encoding EEG into identifiable low-dimension latent embeddings from differential entropy powers by self-supervised contrastive learning. The latent embedding can be an informative representation uesd for downstream tasks. Using contrastive learning to extract latent embedding for EEG data is an interesting and promising topic and still needs a lot of studies. In the future, we will explore more traditional EEG features or even the raw signals for encoding EEG embeddings with contrastive learning and other self-supervised alternatives. We aim to find the algorithm to generate invariant and identifiable EEG embeddings for general tasks, and explore a wider application of EEG in the fields of neural studies, clinical screening and diagnosis, and human-computer interaction.

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
2. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural. Inf. Process. Syst. **27**, 1–9 (2014)
3. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 81–84. IEEE (2013)
4. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
5. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural. Inf. Process. Syst. **31**, 1–11 (2018)
6. Gao, Y., Archer, E.W., Paninski, L., Cunningham, J.P.: Linear dynamical neural population models through nonlinear embeddings. Adv. Neural. Inf. Process. Syst. **29** (2016)
7. Hinrikus, H., et al.: Electroencephalographic spectral asymmetry index for detection of depression. Med. Biol. Eng. Comput. **47**, 1291–1299 (2009)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Jazayeri, M., Ostojic, S.: Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Curr. Opin. Neurobiol. **70**, 113–120 (2021)
10. Khosla, P., et al.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)

Table 1. Decoding accuracy from different embeddings to different labels with different decoders

Decoding Label	Emotion		Subject
	SVM	KNN	KNN
Time nearness embedding-2D	0.338	0.382	0.040
Time nearness embedding-8D	0.359	0.398	0.044
Time nearness embedding-16D	0.415	0.428	0.065
Emotion-guided embedding-2D	0.417	0.429	0.026
Emotion-guided embedding-8D	0.500	0.447	0.054
Emotion-guided embedding-16D	**0.507**	0.464	0.072
Subject-guided embedding-2D	0.352	0.341	0.097
Subject-guided embedding-8D	0.326	0.426	**0.234**
Subject-guided embedding-16D	0.369	0.383	0.078

to extract low-dimension latent features where EEGs with similar characteristics should have similar embeddings close in distance, the purpose of the DE extraction and the contrastive learning encoder is somewhat overlapped. Besides, as EEG signals tend to vibrate in a short time and exhibit more distinguishable characteristics in a longer observation, the features extracted from short-time windows only fluctuates and may not be representative of specific properties. More input features besides DE, including statistical features about frequency domain and time domain signals and asymmetric features between electrode channels [7,11], can hopefully improve the embedding performance.

About Encoder Model. The encoder we use in this paper is a four-layer DNN. We also tested with a similar complexity CNN, alternatively. Both the encoding convergence and the visualization of the outcome embeddings are similar and the later decoding accuracy is slightly lower. We also used deeper neural networks (up to 16 1-D convolutional layers). The encoding convergence and decoding accuracy do not improve either. The reason is that the dimension of the DE feature, 310, is not large and a very deep network is not necessary. When we add more features for the encoder input, a more complex neural network may improve the embedding quality, which will be left to our future work.

About Embedding Dimension. For time-nearness-guided and emotion-guided embedding, the higher the dimension, the higher the top-1 classification accuracy for emotion labels. It indicates that high-dimension embeddings have the ability to include more useful information than low-dimension ones. But it is not the same case with subject-guided embeddings. Lower-dimension embeddings may lack representation ability, while higher-dimension embeddings may involve more noise than useful information for subject classification.

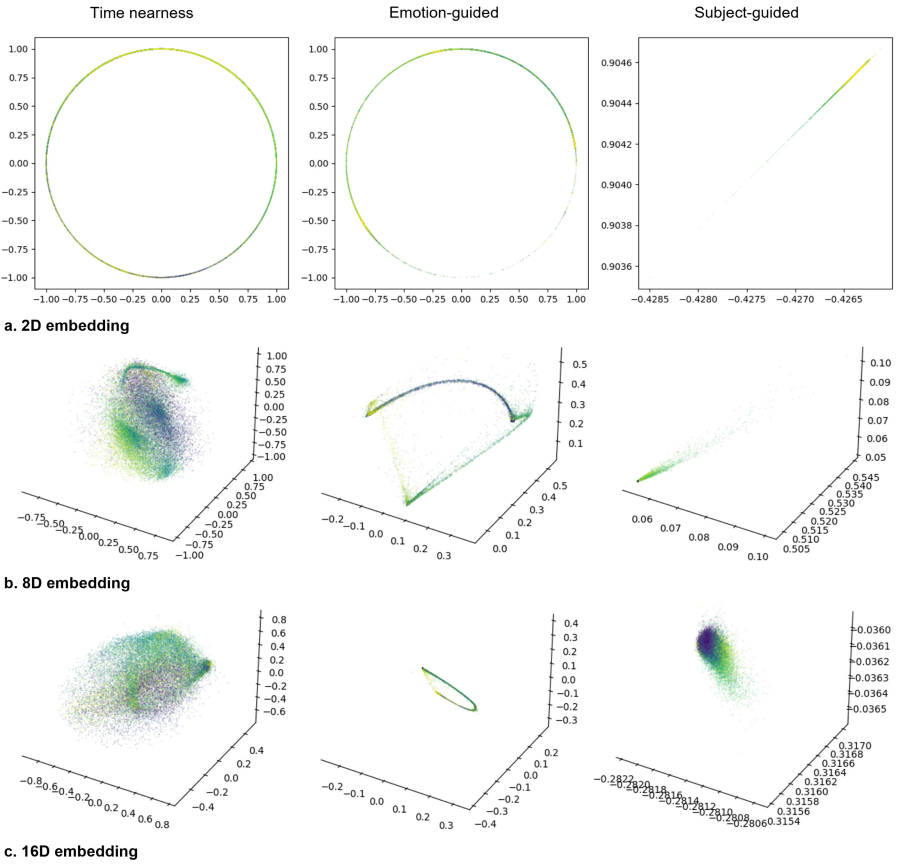


Fig. 3. The first two/three dimensions of learned embeddings of 2D, 8D, and 16D guided by time nearness, emotion labels, and subject labels, respectively.

reason is that the DE of different frequency bands varies more with people's emotional changes, but is more consistent across different people subjects. Decoding subject-guided embeddings to emotion labels also generates low accuracy, because the subject-guided embeddings have removed information about distinguishing emotions.

4 Discussion

About de Feature. The DE is proven a significant single feature which greatly reduces the feature dimension to 5 values for each channel in each short-time window, with some important frequency domain information remained and some noises filtered. But it also leaves out much useful information for encoding an invariant and identifiable embedding. Since the embedding encoder is supposed

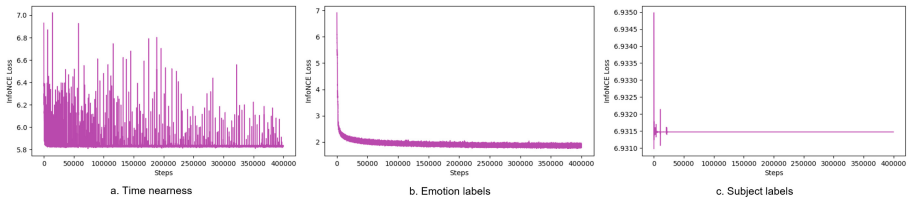


Fig. 2. NCE loss the encoder guided by different criteria

a learning rate of 0.001. The encoder does not converge in the time nearness and subject label cases. The NCE loss in the time nearness case jitters with a lower bound of about 5.8 limited by the time window length, while that in the subject label case is a straight line at the level of about 6.9315. The NCE loss in the emotion label case drops quickly in the first 1,000 iterations and gradually converges to about 1.9 after that.

3.2 Embedding Visualization

Visualizing the embeddings can help to interpret the encoding quality [9]. As shown in Fig. 3, we generate embeddings guided by different information (time or task related labels) with different output dimension sizes, 2, 8, and 16. The first two or three dimensions of the embeddings are drawn where values of the related information are indicated by colors, where embeddings related to the same information have the same color. When the embeddings are identifiable, the embedding points will cluster by color and have a clearer contour intuitively, where points of the same color are closer to each other and farther away from embedding points with different colors.

3.3 Decoding Accuracy

The results of top-1 accuracy of decoding embeddings of different dimensions from different encoder models to different labels are shown in Table 1. The highest classification accuracy for emotion labels is 0.507, which achieved by using the non-linear SVM method with 16-dimension embeddings generated by the emotion-guided encoder, and that for subject labels is 0.234, which achieved by using KNN with 8-dimension embeddings generated by the subject-guided encoder. The accuracy of decoding label-guided embeddings is higher than decoding embeddings that are guided by time nearness when no label is provided. The higher dimension of the embeddings tends to increase classification accuracy, except for decoding the subject-guided embeddings to subject labels. The embeddings generated by time nearness information can also be used for emotion classification, which indicates the potential application of contrastive learning to EEG to a wider range of downstream tasks. The classification accuracy for subject labels is much poorer than that for emotion labels. The possible

encoder is non-linear and is usually a convolutional neural network (CNN) or a deep neural network (DNN) that applies contrastive learning, which follows a similar procedure as in [18].

For the DE features h and g, where g is a positive or negative sample of h, let $p(h)$ be the probability density function of h, $p(g|h)$ and $q(g|h)$ be the probability density function of the positive and negative samples conditioned on h, respectively. After encoding h and g, $c(h)$ and $c(g)$ are their normalized latent embeddings, respectively. The similarity function between $c(h)$ and $c(g)$ is denoted as $\psi(h, g)$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{h\sim p(h), g_+\sim p(g|h)\\ g_1,g_2,\ldots,g_n\sim q(g|h)}} [-\psi(h, g_+) + log \sum_{i=1}^{n} e^{\psi(h,g_i)}].$$

Positive and negative samples are taken from a minibatch of the training input. The identification of positive and negative samples depends on the scientific problem we are solving. It can be based on time nearness between h and g if no label is provided, where samples close to h in time are considered positive and those far from h in time are considered negative. We can also provide labels to guide the training so that samples with the same label as that of h are considered positive and others are negative. The label-guided approach is supervised contrastive learning. With the SEED dataset with emotion labels from different subjects, we learn different encoder models based on time, emotion labels, and subject labels, respectively, and compare their embedding performance.

As to the similarity function, we use the dot product of the normalized latent embeddings adjusted with a temperature parameter τ, i.e., $\psi(h, g) = c(h)^T c(g)/\tau$.

Embedding Decoding. In application, EEG embeddings can be decoded for classification and regression tasks, as shown in Fig. 1.c. We use K-nearest neighbors (KNN) and non-linear support vector machine (SVM) models to classify the EEG embeddings generated by different encoders into emotion and subject labels, respectively. The embedding decoder is trained separately from the embedding encoder, and the training embeddings for the decoder are generated by the well-trained encoder from training DE features.

3 Results

3.1 Contrastive Learning Convergence

We explore the convergence performance of the encoder by contrastive learning with different criteria for identifying the positive and negative samples. Figure 2 shows the NCE loss of training a 4-layer neural network using GELU activation functions to encode EEG to 16-dimension embeddings guided by time nearness (when no label is provided), emotion labels, and subject labels, respectively. The encoder is trained for 10,000 iterations with a minibatch size of 1024 and

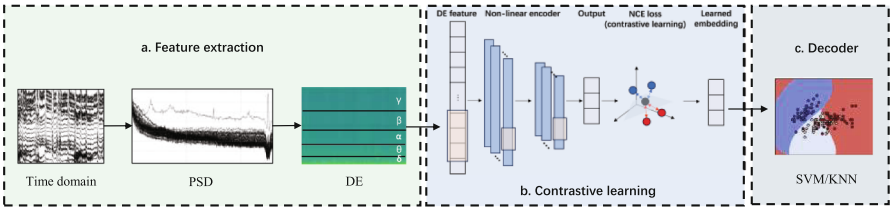


Fig. 1. The procedure of encoding EEG DE into embeddings by contrastive learning and decoding the learned embeddings

2.2 Model

The whole procedure of the model is depicted in Fig. 1. It composes three steps: the DE feature extraction, which generates DE of the frequency domain data from the origin time domain representation; the contrastive learning encoder, which encodes the DE features into latent embeddings by contrastive learning; and the decoder, which decodes the latent embedding to labels of interest.

DE Feature Extraction. DE [3] has the ability to discriminate signals between high and low frequency energy. We first transform the time domain signals to the frequency domain power in non-overlapped short-time Hanning windows and then follow a similar process to [21] to extract the DE of different frequency bands in each electrode channel (Fig. 1.a). The difference is, instead of using the magnitude spectrum as the input, we use the power spectrum density (PSD), which is recognized better than the magnitude spectrum for analyzing random vibration signals as its value is independent of frequency.

If we assume the PSD within a specific frequency band in the electrode channel i, represented by X_i, follows Gaussian distribution, i.e., $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the DE is calculated as

$$\begin{aligned} h(X_i) &= -\int_{X_i} f(x) log(f(x)) dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} log \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} dx \\ &= \frac{1}{2} log 2\pi e\sigma^2, \end{aligned}$$

where $f(x)$ is the probability density of $x \in X_i$. For each electrode channel, we divide the frequency into five bands (delta $\in [1, 4)$ Hz, theta $\in [4, 8)$ Hz, alpha $\in [8, 14)$ Hz, beta $\in [14, 31)$ Hz, and gamma $\in [31, 50)$ Hz), and calculate the DE for each frequency band, respectively. Therefore, for each short-time window, we extract 62×5 DE features, which is represented as $h(X)$.

Contrastive Learned Embeddings. As shown in Fig. 1.b, DE features are fed into a learnable encoder and the output is the EEG latent embedding. The

with specific properties of interest, much work has been done on extracting various EEG features and using different statistical or machine-learning models to retrieve useful information about behavior or health status.

Using proper features or latent embeddings of EEG is critical for EEG analysis. EEG signals usually have large sizes and come with significant noises. Most existing work uses frequency domain signals as features [13,19]. For example, it has been proved that differential entropy (DE) of different bands [3,21] incorporates useful emotional information. The latent approach [4,5] extracts invariant and identifiable latent embeddings of EEG, which can significantly reduce the representation size and extract useful information out of noises. Most work [6,17] uses supervised learning models to get latent embeddings of EEG for tasks with specific outcome labels. But much EEG data such as clinical EEG are recorded without labels, where supervised learning cannot be applied. Methods to generate general EEG latent embeddings that are independent of specific tasks can benefit the application of EEG to a wide range of downstream tasks. Recently, some studies [8,14,22] have worked on self-supervised learning methods and yielded promising results. Cebra [18] indicates that contrastive learning [1,2,10,20] has a great potential to extract invariant and identifiable EEG latent embeddings, which motivates us to explore the feasibility of extracting the general EEG latent embeddings for downstream tasks.

We apply contrastive learning, a powerful self-supervised learning algorithm, to transform DE features of EEG into lower dimensional latent embeddings for downstream tasks. We retrieve the DE of frequency power spectrum density and train a deep neural network by contrastive learning which minimizes the noise-contrastive estimation (NCE) [8] loss between generated latent embeddings of samples in a batch of training data. We use either the (self-supervised) implicit time information or (supervised) specific labels to identify positive and negative samples in the NCE loss. The first case can train the model in scenarios without labels, while the learning in the second case is guided by labels and can generate latent embeddings tuned for the specific downstream task. We explore the visual representation of the latent embeddings of different output dimensions generated by encoders guided by different information and find that the embeddings can be identifiable intuitively. We also decode the embeddings in different tasks to investigate the potential to apply the embeddings to a wide range of EEG applications.

2 Method

2.1 Dataset

We use the SEED [21] dataset, which is designed for exploring the relationship between EEG and emotions. The dataset comprises EEG data from 15 people subjects joining a 3-session testing, with each testing session stimulated by watching 15 movie clips related to 3 emotional labels. The signals are collected by 62 electrodes, downsampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data divided by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

Identifiable EEG Embeddings by Contrastive Learning from Differential Entropy Features

Zhen Zhang[1,2,3], Feng Liang[1,2(✉)], Jiawei Mo[4], and Wenxin Hu[1,2]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{fliang,huwenxin}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
zhangzhen19@lzu.edu.cn
[4] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Encoding EEG data into low-dimension latent embeddings greatly facilitates data analysis and interpretation in neuroscience studies, clinical diagnosis, and human-computer interaction. But generating informative and identifiable latent embeddings that are representative of the origin EEG is not an easy mission. Contrastive learning has the potential to utilize large amounts of unlabelled EEG data and extract informative and identifiable latent embeddings for a wide range of downstream tasks. We explore the feasibility of applying the contrastive learning method to train the EEG latent encoder from the feature of differential entropy of short-time window frequency domain signals. The encoder minimizes the noise-contrastive estimation loss by comparing the embeddings with positive and negative embedding samples, where the distinction of samples is guided by time nearness information or task-specific labels. We test encoders with different output dimensions and the outcome latent embeddings can be identifiable via visualization of a few dimensions. The decoding result also shows that the embeddings preserve information about the original EEG features and can be potentially used for a wide range of downstream tasks. The source code is available at: https://www.github.com/liangfengsid/deContrastiveLearning.

Keywords: EEG · Contrastive learning · Latent embedding

1 Introduction

Electroencephalogram (EEG) are electrical signals on the scalp collected by a set of electrodes and has been widely applied in neuroscience research [15], clinical diagnosis [16], and behavior and affection analysis [11,12]. To relate EEG

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 227–235, 2024.
https://doi.org/10.1007/978-3-031-65126-7_21

6. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
7. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural Inf. Process. Syst. **31** (2018)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Niso, G., et al.: Open and reproducible neuroimaging: from study inception to publication. In: NeuroImage 119623 (2022)
10. Schirrmeister, R.T., et al.: Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. **38**(11), 5391–5420 (2017)
11. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 360–368 (2023)
12. Song, Y., Zheng, Q., Liu, B., Gao, X.: EEG conformer: convolutional transformer for EEG decoding and visualization. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 710–719 (2022)
13. Sun, R., Sohrabpour, A., Worrell, G.A., He, B.: Deep neural networks constrained by neural mass models improve electrophysiological source imaging of spatiotemporal brain dynamics. Proc. Natl. Acad. Sci. **119**(31), e2201128119 (2022)
14. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
15. Wynn, J.K., Roach, B.J., McCleery, A., Marder, S.R., Mathalon, D.H., Green, M.F.: Evaluating visual neuroplasticity with EEG in schizophrenia outpatients. Schizophr. Res. **212**, 40–46 (2019)
16. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)

Visualizing the self-supervised learned EEG embeddings can help judge the potential value of the input feature and the effect of the encoder. In the discovery-driven case, the point cloud is in chaos when the time window size is 0.05 s, but some points of the same color start to gather and form blur contours. When the time window size is 0.5 s, we can already see some overlapping color clouds. This indicates that the corresponding input features contain some valuable information and the encoder properly transforms them into identifiable embeddings. Otherwise, if the color cloud is always in chaos, either we need to try other input features, or we need to investigate the effectiveness of the encoder algorithm.

4 Discussion and Future Work

In this paper, we use time-domain EEG features as an example to show that encoding EEG into latent embeddings and visualizing them can greatly improve the interpretability and understandability of EEG and help EEG readers easily discover some patterns. But other traditional EEG features for various tasks can also apply the latent embedding visualization so that clinical doctors and neuroscientists can make a preliminary decision from the EEG results before they go on further analysis.

In the future, we will explore self-supervised and supervised encoding methods with more traditional EEG features and see how visualizing the low-dimension embeddings can help to reveal identifiable patterns. We will also develop other visualization techniques for EEG that make EEG more understandable to clinical doctors and patients, neuroscientists, and biomedical engineers.

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
3. Daud, S.N.S.S., Sudirman, R.: Pattern of EEG voltage and oscillations under stimulation of Mozart's music and white noise for visual learning process. Biomed. Signal Process. Control **85**, 104986 (2023)
4. Donoghue, T., et al.: Parameterizing neural power spectra into periodic and aperiodic components. Nat. Neurosci. **23**(12), 1655–1665 (2020)
5. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Processing Syst. **27** (2014)

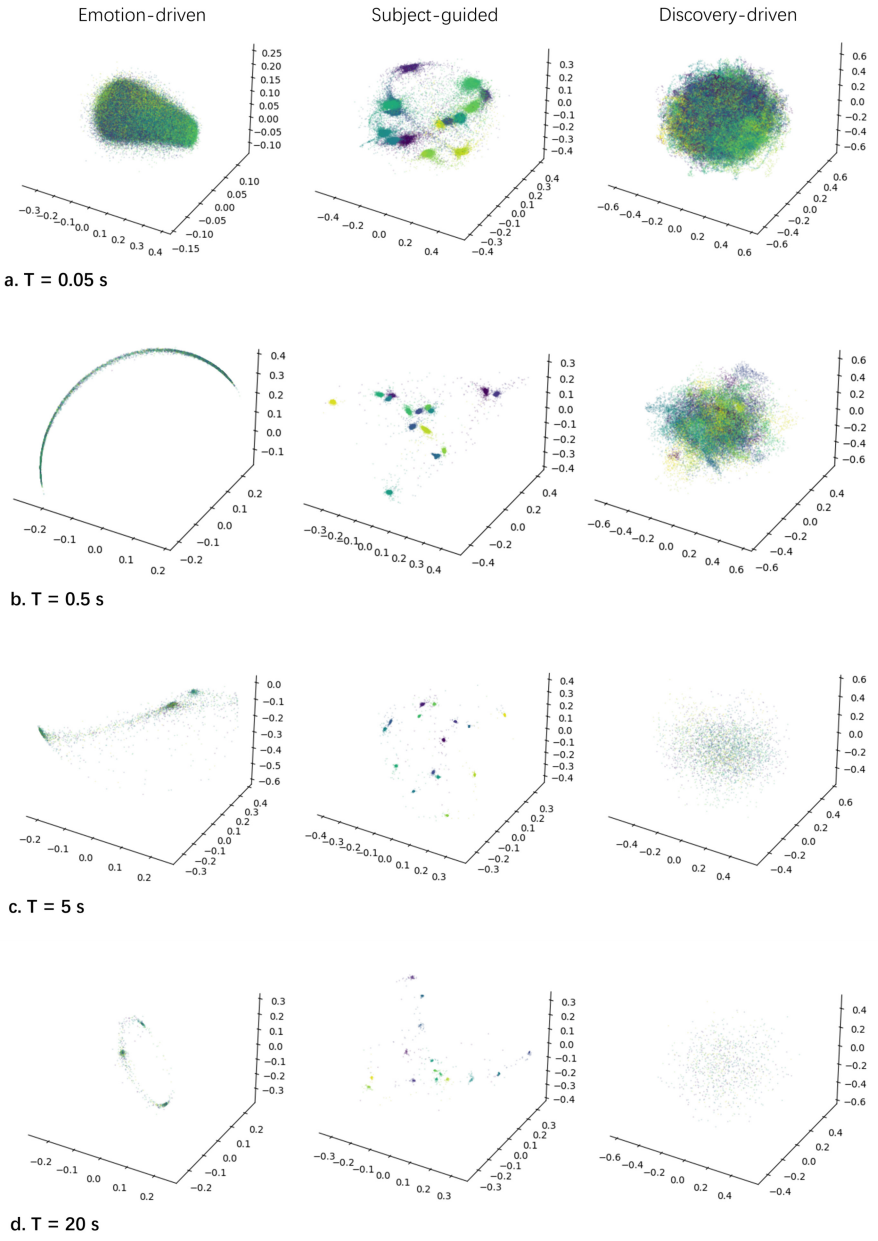


Fig. 1. The first three dimensions of EEG embeddings by discovery-driven, emotion-label-guided, and subject-label-guided contrastive learning, respectively, with features extracted from different time window sizes.

temperature parameter τ as the similarity function between these two latent embeddings, which is denoted as $\psi(x, y) = f(x)^T f(y)/\tau$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{x\sim p(x), y_+\sim p(y|x) \\ y_1, y_2, \ldots, y_n \sim q(y|x)}} [-\psi(x, y_+) + log \sum_{i=1}^{n} e^{\psi(x, y_i)}].$$

Positive and negative samples are taken from a minibatch of the training input. In the discovery-driven manner when no label is provided (self-supervised), samples near x along the timeline are positive and those far away from x along the timeline are negative. In the hypothesis manner, specific labels are provided (supervised), samples with the same label as that of x are positive, while those with different labels from x are negative. We train different encoder models without any label, with emotion labels, and with subject labels, respectively.

The encoder is a five-layer 1D convolutional network with skipping connections with each perceptron activated a GELU function. The mini-batch size of the input is 1,024, the learning rate is 0.001. The encoder output dimension is 32, and the number of mini-batch training iterations is 320,000.

2.3 Visualization

If the low-dimension EEG embeddings are invariant and discriminative, only drawing a few dimensions would be enough for revealing intuitive information about their classes. We plot the first three dimensions of the EEG embeddings of the testing set as points along the axes in the figure, with colors to indicate their related temporality or labels.

3 Results and Discussions

Figure 1 shows the results of plotting the first three dimensions of EEG embeddings by self-supervised and supervised contrastive learning methods with time-domain features extracted from different time window sizes. The clusters of color points reveal abundant information on the classes of the EEG samples.

The EEG embedding visualization can be applied to emotion recognition. In the emotion-label-guided case, we cannot find obvious patterns from the point cloud when the time window size is too small, e.g., at 0.05 or 0.5 s. But as the time window size increases, e.g., to 5 s or 20 s, points tend to cluster into 3 clouds, probably corresponding to 3 types of emotions.

Visualizing EEG embeddings can also be used for identity recognition. In the subject-label-guided case, even when the time window size is 0.05 s, the points are roughly clustered into 15 groups, probably corresponding to the 15 subjects, respectively. As the time window size increases, the contours of the point clouds become more apparent, and people can identify the cluster that a point belongs to with high confidence.

increases information density and improves interpretability when visualized [6,7,10,12]. Therefore, we are motivated to explore the visualization effect and the interpretability of EEG embeddings encoded by different methods.

We visualize the EEG embeddings generated by contrastive-learned encoders [2,8] and investigate their ability to provide distinguishable information. The contrastive-learned encoders can be either self-supervised models or supervised ones, depending on whether it is trained in the discovery-driven manner without labels provided or in the hypothesis manner with task-related labels provided. We encode different EEG features in different training manners and visualize a few dimensions of the embeddings, where the colors of points are related to the labels or their temporal information. We find that by inspecting these figures of the EEG embeddings, people can clearly identify clouds of clustered points, where each cloud consists of points whose original EEG features are considered similar. Our study shows that compressing EEG data to low-dimension embeddings by contrastive learning and visualizing only a few dimensions can help EEG readers easily recognize the inherent patterns and relationships.

2 Method

2.1 Dataset and Feature Extraction

We use the SEED [16] dataset, which comprises EEG data from 15 persons (subjects) joining a 3-session testing, with each testing session stimulated by watching 15 movie clips of a total of about 3600 s. The movie stimuli are related to 3 emotions, i.e., positive, neutral, and negative. The EEG signals are collected by 62 electrode channels, down-sampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data grouped by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

We extract time-domain features from non-overlapping sliding time windows of different sizes, i.e., 0.05, 0.5, 5, and 20 s. For every time window, we extract 5 statistic voltage features for each of the 62 channels, namely the maximum, minimum, mean, median, and standard deviation. The size of each encoder input instance is 310.

2.2 Contrastive Learned Embeddings

The encoder is a non-linear convolutional neural network (CNN) that applies contrastive learning optimizing the NCE loss, which follows a similar procedure as in [11].

For the input features x and y, where y is a positive or negative contrastive sample of x, let $p(x)$ be the probability density function of x, $p(y|x)$ and $q(y|x)$ be the probability density function of the positive and negative samples conditioned on x, respectively. Encoding x and y can be represented by a function f with normalized outputs, and $f(x)$ and $f(y)$ are the normalized latent embeddings, respectively. We use the dot product of $f(x)$ and $f(y)$ adjusted with a

Investigating the EEG Embedding by Visualization

Yongcheng Wen[1,2], Jiawei Mo[3], Wenxin Hu[1,2], and Feng Liang[1,2(✉)]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{1120200244,huwenxin,fliang}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Visualizing EEG data helps clinical doctors and neuroscientists discover potential patterns and abnormalities before further mathematical analysis. Encoding complex EEG data into low-dimension embeddings and visualizing the points in 3-dimension axes with colors can help users quickly recognize some EEG properties. We apply contrastive learning in both self-supervised and supervised manners to extract the time-domain EEG features within different time window sizes. The color points tend to cluster into clouds based on their related classes and graph readers can roughly distinguish people's emotions and identities directly by inspecting the graphs. With self-supervised encoders where the generated embeddings are supposed to be used for general tasks, the visualization method can also uncover the value of the original input features extracted from raw EEG data. The source code is available at: https://www.github.com/liangfengsid/visContrastive.

Keywords: EEG · Contrastive learning · Latent embedding · Visualization · Self-supervised

1 Introduction

Visualizing the electroencephalogram (EEG) helps users easier to discover patterns in clinical diagnosis [15] and neuroscience studies [9]. Traditional visualization elements of EEG include time-domain signals (such as the voltage) [3], frequency-domain signals (such as the power spectrum density) [4], and the source estimate [13]. However, these elements usually do not directly provide intuitively distinguishable patterns and readers may not easily extract valuable results from the visualization without further analysis. The latent approach for visualization is pervasive in computer vision [1, 5, 14]. Encoding high-dimension EEG features to low-dimension embeddings

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 221–226, 2024.
https://doi.org/10.1007/978-3-031-65126-7_20

E-Health Networks II

49. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), vol. 96, no. 34, pp. 226–231 (1996)
50. Nambiar, A., Bernardino, A., Nascimento, J.C., Fred, A.: Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In: International Conference on Automatic Face & Gesture Recognition, pp. 973–980. IEEE (2017)
51. Munaro, M., Ghidoni, S., Dizmen, D.T., Menegatti, E.: A feature-based approach to people re-identification using skeleton keypoints. In: International Conference on Robotics and Automation (ICRA), pp. 5644–5651. IEEE (2014)
52. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: International Conference on Pattern Recognition (ICPR), vol. 4, pp. 441–444. IEEE (2006)
53. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124 (2015)
54. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**(2), 207–244 (2009)
55. Davis, J.V., Kulis, B., Jain P. Sra,, S., Dhillon, I.S.: Information-theoretic metric learning. In: International Conference on Machine Learning (ICML), pp. 209–216 (2007)
56. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_21
57. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. IEEE (2010)
58. Liu, Z., Zhang, Z., Wu, Q., Wang, Y.: Enhancing person re-identification by integrating gait biometric. Neurocomputing **168**, 1144–1156 (2015)
59. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
60. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
61. Chen, C.-H., Ramanan, D.: 3D human pose estimation= 2D pose estimation+ matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7035–7043 (2017)
62. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 30, no. 1, pp. 2058–2065 (2016)

31. Munaro, M., Basso, A., Fossati, A., Van Gool, L., Menegatti, E.: 3D reconstruction of freely moving persons for re-identification with a depth sensor. In: International Conference on Robotics and Automation (ICRA), pp. 4512–4519. IEEE (2014)
32. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1735–1742 (2006)
33. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3733–3742 (2018)
34. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
35. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Adv. Neural Inf. Process. Syst. **33**, 22243–22255 (2020)
36. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: International Conference on Learning Representation (ICLR) (2021)
37. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Process. Syst. **27**, 766–774 (2014)
38. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: International Conference on Artificial Intelligence and Statistics, pp. 297–304 (2010)
39. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 6002–6012 (2019)
40. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
41. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6707–6717 (2020)
42. Ye, M., Zhang, X., Yuen, P.C., Chang, S.-F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6210–6219 (2019)
43. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 9865–9874 (2019)
44. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML), pp. 1597–1607 (2020)
45. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2020)
46. Winter, D.A.: Biomechanics and Motor Control of Human Movement. John Wiley & Sons, Hoboken (2009)
47. Aggarwal, J.K., Cai, Q., Liao, W., Sabata, B.: Nonrigid motion analysis: articulated and elastic motion. Comput. Vis. Image Underst. **70**(2), 142–156 (1998)
48. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representation (ICLR) (2018)

15. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 667–676 (2019)
16. Karianakis, N., Liu, Z., Chen, Y., Soatto, S.: Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–733. Springer, Heidelberg (2018)
17. Ge, Y., Zhu, F., Chen, D., Zhao, R.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Adv. Neural. Inf. Process. Syst. **33**, 11309–11321 (2020)
18. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with RGB-D Sensors. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 433–442. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_43
19. Andersson, V.O., Araujo, R.M.: Person identification using anthropometric and gait data from Kinect sensor. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 425–431 (2015)
20. Rao, H., et al.: Self-supervised gait encoding with locality-aware attention for person re-identification. In: International Joint Conference on Artificial Intelligence (IJCAI), vol. 1, pp. 898–905 (2020)
21. Rao, H., et al.: A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **01**, 1–1 (2021)
22. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 973–980 (2021)
23. Rao, H., Hu, X., Cheng, J., Hu, B.: SM-SGE: a self-supervised multi-scale skeleton graph encoding framework for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1812–1820 (2021)
24. Han, F., Reily, B., Hoff, W., Zhang, H.: Space-time representation of people based on 3D skeletal data: a review. Comput. Vis. Image Underst. **158**, 85–105 (2017)
25. Tanawongsuwan, R., Bobick, A.: Gait recognition from time-normalized joint-angle trajectories in the walking plane. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. II–II (2001)
26. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recogn. **98**, 107069 (2020)
27. Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L.: One-shot person re-identification with a consumer depth camera. In: Gong, S., Cristani, M., Yan, S., Loy, C.C. (eds.) Person Re-Identification. ACVPR, pp. 161–181. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4_8
28. Yoo, J.-H., Nixon, M.S., Harris, C.J.: Extracting gait signatures based on anatomical knowledge. In: Proceedings of BMVA Symposium on Advancing Biometric Technologies, pp. 596–606. Citeseer (2002)
29. Murray, M.P., Drought, A.B., Kory, R.C.: Walking patterns of normal men. J. Bone Joint Surg. **46**(2), 335–360 (1964)
30. Pala, P., Seidenari, L., Berretti, S., Del Bimbo, A.: Enhanced skeleton and face 3D data for person re-identification from depth cameras. Comput. Graph. **79**, 69–80 (2019)

Lastly, we propose a skeleton prototype contrastive learning scheme to cluster unlabeled skeleton graph representations and contrast their inherent similarity with representative skeleton features to learn effective skeleton representations for person re-ID. The proposed SPC-MGR outperforms several state-of-the-art skeleton-based methods, and is also highly effective in more general person re-ID scenarios.

References

1. Nambiar, A., Bernardino, A., Nascimento, J.C.: Gait-based person re-identification: a survey. ACM Comput. Surv. **52**(2), 33 (2019)
2. Zheng, W.-S., Gong, S., Xiang, T.: Towards open-world person re-identification by one-shot group-based verification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 591–606 (2015)
3. Baltieri, D., Vezzani, R., Cucchiara, R.: SARC3D: a new 3D body model for people tracking and re-identification. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011. LNCS, vol. 6978, pp. 107–206. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24085-0_21
4. Vezzani, R., Baltieri, D., Cucchiara, R.: People reidentification in surveillance and forensics: a survey. ACM Comput. Surv. **46**(2), 29 (2013)
5. Tan, H., Liu, X., Yin, B., Li, X.: MHSA-Net: multihead self-attention network for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8210–8224 (2022)
6. Zhu, K., Guo, H., Liu, S., Wang, J., Tang, M.: Learning semantics-consistent stripes with self-refinement for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **34**, 8531–8542 (2022)
7. Zheng, Z., Wang, X., Zheng, N., Yang, Y.: Parameter-efficient person re-identification in the 3d space. IEEE Trans. Neural Netw. Learn. Syst. **35**, 7534–7547 (2022)
8. Miao, J., Wu, Y., Yang, Y.: Identifying visible parts via pose estimation for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(9), 4624–4634 (2021)
9. Wei, Z., Yang, X., Wang, N., Gao, X.: Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(9), 4676–4687 (2021)
10. Zhou, Q., Zhong, B., Liu, X., Ji, R.: Attention-based neural architecture search for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(11), 6627–6639 (2021)
11. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE Trans. Pattern Anal. Mach. Intell. **25**(12), 1505–1518 (2003)
12. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2164–2176 (2011)
13. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. IEEE Trans. Pattern Anal. Mach. Intell. **38**(12), 2501–2514 (2016)
14. Zhao, R., Oyang, W., Wang, X.: Person re-identification by saliency learning. IEEE Trans. Pattern Anal. Mach. Intell. **39**(2), 356–370 (2017)

[58] that fuses RGB appearance and GEI features by up to 57.6% top-1 accuracy, 39.3% top-5 accuracy, and 29.1% top-10 accuracy. Despite only utilizing estimated skeleton data with noise for training, the proposed unsupervised approach can still obtain highly competitive performance compared with supervised appearance-based methods in different conditions, which demonstrates the great potential of our approach to be applied to large-scale RGB-based datasets under more general re-ID settings.

5.4 Application to Generalized Person Re-Identification

Our approach can learn a unified skeleton graph representation for different skeleton data with varying body joints or topologies, which enables the pre-trained model to be directly transferred to different datasets for the generalized person re-ID task. To evaluate the effectiveness of our approach on generalized person re-ID, we exploit the model trained on the source dataset to perform person re-ID on the target dataset, *i.e.*, direct domain generalization (DG), and then further fine-tune the model with the unlabeled data of target datasets, *i.e.*, unsupervised fine-tuning (UF), to compare the generalization performance. As shown in Table 6, we can draw the following observations and conclusions. The model trained on one dataset can be transferred to other unseen target datasets and even achieves better person re-ID performance. Direct generalization is shown to be effective among different datasets, while unsupervised fine-tuning on the target dataset can further improve the person re-ID performance. Such results demonstrate that our approach possesses good generalization ability with robustness to domain shifts [62] and can be promisingly applied to other open person re-ID tasks. Interestingly, we observe that training on different source datasets typically leads to different person re-ID performance on a new dataset. For example, the model trained on the KGBD fails to yield satisfactory performance on IAS-B, BIWI-W and BIWI-S, while the pre-trained model of KS20 with further fine-tuning on those testing sets can achieve superior performance to the original ones, as shown by the bold numbers in Table 6, which implies that an appropriate domain initialization or model pre-training of our model could be potentially exploited to facilitate better generalized person re-ID performance.

6 Conclusion

In this paper, we devise unified multi-level graphs to represent 3D skeletons, and propose an unsupervised skeleton prototype contrastive learning paradigm with multi-level relation modeling (SPC-MGR) to learn effective skeleton representations for person re-ID. We devise a multi-head structural relation layer to capture relations of neighbor body-component nodes in graphs, so as to aggregate key correlative features into effective node representations. To capture more discriminative patterns in skeletal motion, we propose a full-level collaborative relation layer to infer dynamic collaboration among different-level components. Meanwhile, a multi-level graph fusion is exploited to integrate collaborative node features across graphs to enhance structural semantics and global pattern learning.

Table 6. Generalized person re-ID performance of our approach with direct domain generalization (DG) from source datasets ("Source") to target datasets ("Target"). "UF" represents fine-tuning the source model with the unlabeled data of target datasets. BIWI-W/S denotes the Walking/Still testing set of BIWI. Bold numbers indicates that the model using "DG" or "UF" obtains better performance than the original one trained on the same dataset.

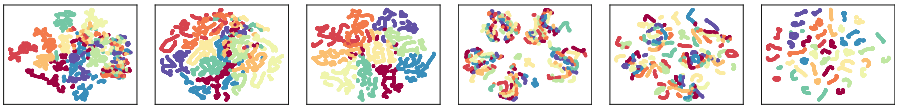
	Source	KS20				KGBD				IAS-Lab				BIWI			
Target	Type	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
KS20	**DG**	—	—	—	—	19.7	54.3	69.7	10.2	20.1	60.4	75.6	13.5	29.7	62.7	79.9	15.0
	UF	59.0	79.0	86.2	21.7	48.4	77.7	85.2	21.6	52.2	78.3	**89.1**	**22.8**	50.8	**79.7**	**87.1**	**21.9**
KGBD	**DG**	18.3	37.3	46.7	4.4	—	—	—	—	15.6	35.6	46.2	4.0	20.5	40.8	50.1	4.9
	UF	28.5	45.2	52.9	6.4	40.8	57.5	65.0	6.9	31.1	48.1	55.7	6.5	29.7	47.4	55.0	6.4
IAS-A	**DG**	27.9	57.1	71.5	15.6	29.6	56.5	71.6	16.0	—	—	—	—	27.5	53.6	67.9	14.4
	UF	**42.8**	**67.7**	**77.5**	23.0	34.1	59.4	72.0	18.4	41.9	66.3	75.6	24.2	37.2	61.8	72.2	23.8
IAS-B	**DG**	32.0	60.6	72.0	15.7	29.5	58.8	68.9	16.5	—	—	—	—	27.0	57.6	70.5	13.5
	UF	**45.4**	**68.8**	**81.4**	**29.0**	35.9	61.6	72.1	21.9	43.3	68.4	79.4	24.1	39.1	66.9	74.5	**24.2**
BIWI-W	**DG**	19.3	31.5	38.9	**19.6**	10.3	22.5	33.1	12.0	10.0	23.1	31.1	15.9	—	—	—	—
	UF	**21.6**	**32.3**	**40.9**	**20.7**	15.0	29.6	39.1	14.3	17.7	29.3	36.2	16.6	18.9	31.5	40.5	19.4
BIWI-S	**DG**	23.8	52.2	**69.9**	14.2	19.0	49.4	67.4	8.5	18.8	46.1	61.1	12.2	—	—	—	—
	UF	**40.4**	**62.9**	**74.2**	**16.2**	21.3	46.5	57.2	9.8	27.9	44.5	63.9	12.8	34.1	57.3	69.8	16.0

Such results imply that our model may learn skeleton representations with finer separation and enable pattern-based grouping in a specific class.

5.3 Application to Model-Estimated Skeleton Data

To verify the effectiveness of our skeleton-based approach when applied to large-scale RGB-based settings (CASIA-B), we exploit pre-trained pose estimation models [60,61] to extract 3D skeleton data from RGB videos of CASIA-B, and evaluate the performance of our approach with the estimated skeleton data. We compare our approach with representative appearance-based methods [54–58] and skeleton-based methods [20,23]. Note that as two fundamentally different data modalities, skeleton data are much smaller and less informative than appearance-based data (*e.g.*, RGB images), thus directly comparing skeleton-based methods with appearance-based methods is generally considered unfair. In our comparison, we provide results of classic and representative appearance-based methods as a performance reference.

As reported in Table 5, our approach is superior to recent skeleton-based methods SM-SGE and AGE with an evident performance gain of 7.2–50.4% top-1 accuracy and 1.1–5.6% mAP in four out of five evaluation conditions of CASIA-B, which substantiates that the proposed approach is capable of learning more discriminative skeleton representations than these methods in the case of using model-estimated skeleton data. Compared with representative classic appearance-based methods that utilize visual features, *e.g.*, RGB features and silhouettes, our skeleton-based approach still achieves the best performance in most conditions. For instance, our approach not only performs better than LMNN [54] and ITML [55] that use metric learning with different visual features (RGB and HSV colors and textures) [58], but also surpasses the score-based MLR model



(a) AGE (on BIWI) (b) SM-SGE (on BIWI) (c) Ours (on BIWI) (d) AGE (on KS20) (e) SM-SGE (on KS20) (f) Ours (on KS20)

Fig. 4. t-SNE visualization of the skeleton representations learned by AGE [20] ((a), (d)), SM-SGE [23] ((b), (e)), and our proposed SPC-MGR ((c), (f)) for the first 10 classes in BIWI and KS20 datasets. Note: Different colors indicate skeleton representations of different classes.

Table 5. Performance comparison with appearance-based and skeleton-based methods on CASIA-B. Note: "Cl-Nm" denotes the probe set under "Clothes" condition and gallery set under "Normal" condition. † refers to appearance-based methods and * represents requiring label information for training. "—" indicates no published result.

Probe-Gallery	Nm-Nm				Bg-Bg				Cl-Cl				Cl-Nm				Bg-Nm			
Methods	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
†LMNN* [54]	3.9	22.7	36.1	—	18.3	38.6	49.2	—	17.4	35.7	45.8	—	11.6	12.6	17.8	—	23.1	37.1	44.4	—
†ITML* [55]	7.5	22.2	34.2	—	19.5	26.0	33.7	—	20.1	34.4	43.3	—	10.3	24.5	36.1	—	21.8	30.4	36.3	—
†ELF* [56]	12.3	35.6	50.3	—	5.8	25.5	37.6	—	19.9	43.9	56.7	—	5.6	16.0	26.3	—	17.1	30.0	37.9	—
†SDALF [57]	4.9	27.0	41.6	—	10.2	33.5	47.2	—	16.7	42.0	56.7	—	11.6	19.4	27.6	—	22.9	30.1	36.1	—
†Score-based MLR* [58]	13.6	48.7	63.7	—	13.6	48.7	63.7	—	13.5	48.6	63.9	—	9.7	27.8	45.1	—	14.7	32.6	50.2	—
†Feature-based MLR* [58]	16.3	43.4	60.8	—	18.9	44.8	59.4	—	25.4	53.3	68.9	—	20.3	**42.6**	**56.9**	—	31.8	**53.6**	**64.1**	—
AGE [20]	20.8	29.3	34.2	3.5	37.1	56.2	67.0	9.8	35.5	54.3	65.3	9.6	14.6	33.0	42.7	3.0	**32.4**	51.2	60.1	3.9
SM-SGE [23]	50.2	73.5	81.9	6.6	26.6	49.0	59.4	9.3	27.2	51.4	63.2	9.7	10.6	26.3	35.9	3.0	16.6	36.8	47.5	3.5
SPC-MGR (Ours)	**71.2**	**88.0**	**92.8**	**9.1**	**44.3**	**66.4**	**76.4**	**11.4**	**48.3**	**71.6**	**81.6**	**11.8**	**22.4**	40.4	51.0	**4.3**	28.9	49.3	59.1	**4.6**

the proposed multi-level graphs (Id = 4) performs better than solely using single-level graph (Id = 3) with a remarkable margin of 2.1–16.7% top-1 accuracy and 0.4–5.7% mAP, which demonstrates that modeling body structure and relations at various levels with the proposed graph representations (MG) can encourage the model to learn more useful skeleton features for person re-ID. Adding FCRL further improves the overall performance in terms of both top-1 accuracy by 0.8–13.9% and mAP by 0.5–2.6% on different datasets. Such results verify our claim that combining structural and collaborative body relation learning can facilitate capturing richer features of body structure and skeleton patterns for the person re-ID task.

5.2 Visualization of Skeleton Representations

We conduct a t-SNE [59] visualization of skeleton representations for a qualitative analysis, and compare our approach with two state-of-the-art skeleton-based methods, *i.e.*, AGE [20], SM-SGE [23]. As presented in Fig. 4(c), the skeleton representations learned by our approach can form different class clusters with higher separation than AGE and SM-SGE on BIWI, which suggests the lower entropy of our representations. Interestingly, it is observed that the learned representations on KS20 are separated in small groups of the same class, as shown in Fig. 4(f), which enjoys significantly larger looseness than other two methods.

Table 4. Ablation study of our model with different components: Multi-level skeleton graphs (MG), multi-head structural relation layer (MSRL), full-level collaborative relation layer (FCRL), and skeleton prototype contrastive learning (SPC). "SG" denotes employing the single-level graph (part-level graph) and "+" indicates using the corresponding model component. "SG + MSRL" is evaluated under random model initialization without SPC.

Id	Configurations	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
		top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
1	Baseline	17.0	9.5	20.5	4.4	29.4	13.8	30.2	13.3	10.9	14.1	24.8	9.3
2	SG + MSRL	18.6	10.2	21.4	3.7	30.3	14.3	31.8	13.3	11.2	13.8	26.0	11.0
3	SG + MSRL + SPC	28.4	15.5	26.2	5.7	37.9	21.5	38.5	20.8	15.4	16.4	27.3	12.2
4	MG + MSRL + SPC	45.1	21.2	34.5	6.3	40.0	22.5	41.9	23.2	18.1	16.9	31.5	13.4
5	MG + MSRL + FCRL + SPC	59.0	21.7	40.8	6.9	41.9	24.2	43.3	24.1	18.9	19.4	34.1	16.0

further improvement and significantly outperforms existing supervised methods and other fine-tuned models ("SGELA + FT" and "SM-SGE + FT") on four of six testing sets. This demonstrates that our approach as a generic unsupervised contrastive paradigm can also be applied to more scenarios under label supervision. As here we directly fine-tune the model with a single MLP network (note that supervised learning is not the focus of this work), it is feasible to devise more effective supervised architectures to further boost its performance. In summary, considering that our approach does not require any manual annotation and can achieve highly competitive and more balanced performance with a significantly smaller size of network parameters, it can be a more general solution to skeleton-based person re-ID and related tasks.

5 Further Analysis

5.1 Ablation Study

In this section, we conduct ablation study to demonstrate the necessity of each component in the proposed approach. The skeleton sequences of concatenated joints are adopted as the baseline. As reported in Table 4, we can draw the following conclusions. The model utilizing single-level skeleton graph with MSRL (Id = 2, 3) shows higher performance than the baseline (Id = 1) that directly uses raw body-joint sequences by 0.3–8.5% top-1 accuracy and 0.7–7.7% mAP on all datasets, regardless of using SPC. Such results demonstrate the effectiveness of graph representations, as it can model richer body structural information and mine valuable body-component relations to obtain a more discriminative skeleton representation. Compared with the model without contrastive learning (Id = 2), employing SPC (Id = 3) obtains consistent re-ID performance improvement by up to 9.8% top-1 accuracy and 7.5% mAP on different datasets. This justifies that the proposed SPC is a highly effective contrastive learning paradigm, which enables the model to mine more typical and unique skeleton features of different identities from the unlabeled graph representations for person re-ID. Exploiting

sets of BIWI, both SGELA and the proposed approach obtain comparable mAP, while our SPC-MGR can achieve superior overall performance with higher top-1 (7.2-8.3%), top-5 (5.5-17.5%), and top-10 accuracy (5.4-25.8%). Finally, our approach also performs better than the latest graph-based skeleton representation learning method SM-SGE [23] by a distinct margin of 2.6-13.1% top-1 accuracy and 2.5-12.2% mAP on all datasets. In contrast to direct inter-sequence contrastive learning [21] or manually devising pretext tasks for skeleton representation learning [20,23], our approach can automatically mine most representative skeleton features by contrasting sequence-level representations (instances) and cluster-level representations (prototypes), which enables our model to learn better skeleton representations for person re-ID. Moreover, our model requires only 0.01M parameters and evidently lower computational complexity for skeleton representation learning compared with existing self-supervised and unsupervised methods, as shown in Table 1, which demonstrates its superior efficiency for person re-ID tasks.

We also compare the performance of our approach with state-of-the-art skeleton-based counterparts with the cross-view evaluation (CVE) setup of KS20. As shown in Table 3, our approach remarkably outperforms the latest self-supervised and multi-view skeleton-based methods SGELA [21] and SM-SGE [23] by an average margin of 6.5–32.1% top-1 accuracy, 12.8–41.0% top-5 accuracy, 17.5–40.0% top-10 accuracy, and 5.2–22.8% mAP on 24 out of 25 testing combinations of probe views and gallery views, which demonstrates that our model can learn more discriminative skeleton representations with better robustness against viewpoint variations for cross-view person re-ID.

Comparison with Hand-Crafted and Supervised Methods. Compared with D_{13} [27] and D_{16} [30] that extract hand-crafted geometric and anthropometric skeleton descriptors, our model achieves a significant improvement of person re-ID performance by 1.5–23.8% top-1 accuracy on KGBD, BIWI-S, and BIWI-W. Despite gaining similar performance on IAS-A and IAS-B, these methods are inferior to our approach by at least 7.3% top-1 accuracy on more challenging datasets such as KS20 and KGBD that contains more viewpoints and individuals. Furthermore, with *unlabeled* 3D skeletons as the only input, the proposed approach can obtain comparable or even superior performance to two state-of-the-art supervised methods PoseGait [26] and MG-SCR [22] on five out of six testing sets (KS20, IAS-A, IAS-B, BIWI-S, BIWI-W). Interestingly, with skeleton labels as the supervision, these methods still fail to obtain satisfactory person re-ID accuracy and even perform worse than hand-crafted methods on datasets with frequent view, shape, and appearance changes in KS20, IAS, and BIWI testing sets. This might also suggest that a limited amount of labeled skeleton data in small datasets such as IAS could reduce the ability of supervised models to learn discriminative features, while training with larger-scale skeleton data (KGBD) can encourage them to achieve better performance than conventional methods. On the other hand, when employing supervised fine-tuning with skeleton labels, the performance of our approach ("SPG-MGR + FT") gains a

Table 3. Performance comparison with state-of-the-art self-supervised and unsupervised methods with cross-view evaluation (CVE) setup of KS20 datasets. $0°, 30°, 90°, 130°$, and $180°$ denote probe or gallery sets in different views.

	Gallery	0°				30°				90°				150°				180°			
Probe		top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
0°	AGE [20]	46.7	74.2	83.5	22.5	11.0	35.7	47.5	10.0	8.1	29.9	47.5	9.2	7.5	26.7	43.5	8.4	7.0	23.0	37.4	8.2
	SGELA [21]	76.2	89.6	92.8	37.1	15.1	27.3	35.1	19.9	10.1	27.5	40.9	18.2	10.7	21.5	29.3	18.0	15.4	25.8	38.0	12.6
	SM-SGE [23]	58.4	84.7	92.2	27.7	17.2	50.0	63.3	10.8	7.2	21.9	39.1	10.5	4.4	19.4	34.7	9.3	10.0	23.8	33.1	9.4
	SPC-MGR (Ours)	**78.9**	**94.1**	**97.3**	**52.9**	**26.2**	**53.1**	**71.5**	**22.9**	**39.1**	**59.8**	**71.5**	**31.4**	**30.5**	**57.4**	**72.7**	**26.6**	**27.0**	**52.0**	**66.4**	**19.9**
30°	AGE	10.1	42.8	57.8	8.8	52.3	82.7	91.5	25.0	15.0	35.6	58.5	8.8	10.1	24.2	41.8	8.1	7.8	24.2	34.3	8.3
	SGELA	13.1	19.6	22.6	19.4	70.9	88.2	91.8	40.5	11.8	24.5	36.3	16.5	6.9	22.6	31.7	15.4	9.2	15.4	22.9	13.9
	SM-SGE	18.1	48.4	65.0	11.5	60.2	82.0	89.8	28.2	12.5	27.2	35.3	10.7	7.5	23.4	33.8	10.6	8.8	27.2	39.1	10.5
	SPC-MGR (Ours)	**39.1**	**60.2**	**69.1**	**26.2**	**75.4**	**95.7**	**96.5**	**56.7**	**40.2**	**62.5**	**72.3**	**32.4**	**28.9**	**55.1**	**66.0**	**24.9**	**18.4**	**48.1**	**66.4**	**16.1**
90°	AGE	7.5	27.3	43.2	8.7	9.0	28.5	44.1	9.3	57.4	81.4	90.7	19.2	13.8	41.1	57.1	9.0	7.8	30.0	46.0	8.3
	SGELA	9.6	19.8	29.7	16.4	10.8	15.6	20.4	17.5	48.4	75.7	86.5	31.6	17.1	35.7	43.0	22.0	13.5	23.4	31.8	21.3
	SM-SGE	19.1	33.1	48.1	12.4	23.1	40.6	57.4	11.5	72.2	89.1	92.8	24.9	20.9	48.4	69.4	12.8	19.4	36.9	51.6	11.3
	SPC-MGR (Ours)	**37.5**	**67.2**	**75.0**	**26.0**	**41.8**	**65.2**	**74.2**	**32.2**	**86.7**	**98.1**	**99.2**	**63.1**	**59.0**	**82.4**	**86.3**	**40.7**	**34.8**	**62.1**	**77.0**	**24.8**
150°	AGE	6.7	21.3	34.7	8.2	7.9	23.4	38.9	8.9	15.2	35.9	54.4	9.2	45.3	70.5	82.1	18.7	11.3	37.1	50.2	8.9
	SGELA	5.8	18.8	28.0	14.2	11.6	15.5	20.7	16.8	17.6	47.1	53.2	24.5	59.6	81.5	89.1	36.8	17.0	29.8	32.5	**23.0**
	SM-SGE	8.4	24.4	37.8	10.4	12.9	26.6	36.3	10.9	24.1	53.4	66.3	12.9	64.4	85.9	95.0	25.5	17.8	40.9	59.1	12.1
	SPC-MGR (Ours)	**28.5**	**59.8**	**71.9**	**23.3**	**25.4**	**49.6**	**65.2**	**22.3**	**57.4**	**77.0**	**85.2**	**40.8**	**77.3**	**96.5**	**97.7**	**58.6**	**35.9**	**62.5**	**79.3**	**23.0**
180°	AGE	7.9	17.7	32.6	8.1	5.2	22.4	33.4	8.3	10.5	25.6	34.0	8.2	11.6	33.1	52.9	8.8	47.1	72.4	82.6	22.6
	SGELA	14.0	29.1	39.2	21.3	11.9	20.6	25.9	17.3	18.6	37.8	49.7	19.4	22.7	45.9	55.2	20.7	**74.5**	**92.7**	95.1	38.3
	SM-SGE	5.6	20.0	30.6	8.5	6.6	22.7	31.6	8.6	13.8	34.1	45.6	9.4	10.3	37.5	56.6	10.4	51.9	79.7	87.8	25.6
	SPC-MGR (Ours)	**28.5**	**53.5**	**62.5**	**21.7**	**17.2**	**37.1**	**50.0**	**18.8**	**31.6**	**53.5**	**66.8**	**30.0**	**31.3**	**56.3**	**77.3**	**26.4**	65.2	89.5	**96.5**	**45.9**

that probe sequences are matched with the gallery sequences with correct identities using different-sized gallery candidate lists. We also report Mean Average Precision (mAP) [53] to evaluate the overall performance of our approach.

4.3 Performance Comparison

In this section, we compare our approach with state-of-the-art self-supervised and unsupervised skeleton-based person re-ID methods [20,21,23] on KS20, KGBD, IAS-Lab, and BIWI datasets with different probe settings in Table 1 and Table 2. As a reference for the overall performance, we include the latest supervised skeleton-based person re-ID methods [22,26] and representative handcrafted person re-ID methods [30,31]. For deep learning based methods, we also report their model sizes, *i.e.*, amount of network parameters, and computational complexity in Table 1.

Comparison with Self-supervised and Unsupervised Methods. As presented in Table 1 and Table 2, the proposed SPC-MGR enjoys distinct advantages over existing self-supervised and unsupervised methods on all datasets. Compared with AGE model [20] that learns skeleton features based on body-joint sequence representations, our approach consistently achieves higher person re-ID performance by a large margin of 7.2 to 37.9% top-1 accuracy and 6.0 to 12.8% mAP on different datasets, which demonstrates that the proposed multi-level skeleton graph representations with structural-collaborative body relation learning are more effective on modeling discriminative skeleton features for the person re-ID task. Our approach significantly outperforms the state-of-the-art skeleton contrastive learning method SGELA [21] by up to 25.2% top-1 accuracy and 11.0% mAP on KS20, KGBD, IAS-A, and IAS-B testing sets. On two testing

4 Experiments

4.1 Experimental Setup

Datasets . We evaluate our approach on four skeleton-based person re-ID benchmarks: *KGBD* [19], *BIWI* [27], *KS20* [50], *IAS-Lab* [51], and a large-scale RGB video based multi-view gait dataset: *CASIA-B* [52]. They collect skeleton data from 164, 50, 20, 11, and 124 different individuals respectively.

Implementation Details. The numbers of nodes in the part-level, body-level, and hyper-torso-level graphs are $n_1 = 10$, $n_2 = 5$, and $n_3 = 3$ respectively. The sequence length f is set to 6 for KS20, KGBD, BIWI, and IAS-Lab and $f = 40$ for CASIA-B following previous works for a fair comparison. The node feature dimension is $D_h = 8$ and the number of structural relation heads is $m = 16$ for KGBD and $m = 8$ for other datasets. We use $\lambda_C^{a,b} = 1$ $(a, b \in \{1, 2, 3\})$ to averagely fuse multi-level graph features. For DBSCAN, we empirically use maximum distance $\epsilon = 0.6$ (KGBD, BIWI), $\epsilon = 0.8$ (KS20, IAS-Lab), $\epsilon = 0.75$ (CASIA-B), and adopt minimum amount of samples $a_{min} = 4$ for KGBD and $a_{min} = 2$ for other datasets. We empirically set the temperature τ to 0.06 (KGBD), 0.075 (CASIA-B), 0.07 (BIWI), 0.08 (KS20, IAS-Lab) for skeleton prototype contrastive learning. Experiments with each evaluation setup are repeated for multiple times and the average performance is reported. More implementation details are provided in the appendices.

4.2 Evaluation Metrics

We compute Cumulative Matching Characteristics (CMC) curve and adopts top-1, top-5, and top-10 accuracy as the quantitative metrics, which indicate ratios

Table 2. Performance comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods on IAS-A, IAS-B, and KGBD testing sets. "+ FT" denotes employing supervised fine-tuning with labels. **Bold** refers to the best cases among self-supervised/unsupervised methods, and *italic numbers* indicate the best performers among supervised methods.

		IAS-A				IAS-B				KGBD			
Types	Methods	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
Hand-crafted	D_{13} [27]	40.0	58.7	67.6	24.5	43.7	68.6	76.7	23.7	17.0	34.4	44.2	1.9
	D_{16} [30]	42.7	62.9	70.7	25.2	44.5	69.1	80.2	24.5	31.2	50.9	59.8	4.0
Supervised	PoseGait [26]	28.4	55.7	*69.2*	17.5	*28.9*	*51.6*	*62.9*	*20.8*	*50.6*	*67.0*	*72.6*	*13.9*
	SGELA [21] + FT	18.0	32.1	46.2	13.5	23.6	42.9	51.9	14.8	43.7	58.7	65.0	7.1
	MG-SCR [22]	36.4	59.6	69.5	14.1	32.4	56.5	69.4	12.9	44.0	58.7	64.6	6.9
	SM-SGE [23] + FT	38.5	63.2	73.9	15.0	44.3	68.2	77.5	14.9	43.2	58.6	64.6	7.5
	SPC-MGR (Ours) + FT	*45.1*	*68.1*	*76.2*	*25.3*	*52.0*	*77.3*	*86.0*	*30.1*	42.5	59.6	67.1	9.0
Self-supervised/ Unsupervised	AGE [20]	31.1	54.8	67.4	13.4	31.1	52.3	64.2	12.8	2.9	5.6	7.5	0.9
	SGELA [21]	16.7	30.2	44.0	13.2	22.2	40.8	50.2	14.0	38.1	53.5	60.0	4.5
	SM-SGE [23]	34.0	60.5	71.6	13.6	38.9	64.1	75.8	13.3	38.2	54.2	60.7	4.4
	SPC-MGR (Ours)	**41.9**	**66.3**	**75.6**	**24.2**	**43.3**	**68.4**	**79.4**	**24.1**	**40.8**	**57.5**	**65.0**	**6.9**

k^{th} prototype $\boldsymbol{P}^k$, and τ represents the temperature for contrastive learning, where higher value of τ produces a softer probability distribution over prototypes and retains more similar information among clusters. The proposed SPC loss is essentially a generalized contrastive learning loss that combines multi-level skeleton graph modeling (see Sec. 3.1) and structural-collaborative relational feature fusion (see Sec. 3.2). We can theoretically formulate the objective of SPC as an Expectation-Maximization (EM) solution and extend it to other forms of contrastive paradigms. The theoretical analyses of SPC effectiveness and its relations to existing contrastive losses are provided in the appendices.

3.4 The Entire Approach

The computation flow of the proposed approach can be described as: $\boldsymbol{S} \rightarrow \mathcal{G}$ (Sect. 3.1) $\rightarrow \boldsymbol{F}$ (Sect. 3.2) $\rightarrow \boldsymbol{M}$ (Eq. 7) $\rightarrow \overline{\boldsymbol{M}}$ (Sect. 3.3) $\rightarrow \boldsymbol{P}$ (Eq. 10). For convenience, we use the embedding function $\psi(\cdot)$ to represent the multi-level skeleton graph representation encoding process, which can be formulated as $\psi(\boldsymbol{S}) = \overline{\boldsymbol{M}}$. We perform skeleton prototype contrastive learning by minimizing $\mathcal{L}_{\text{SPC}}$, so as to optimize $\psi(\cdot)$ and learn effective skeleton representations in an unsupervised manner. To facilitate better skeleton representation learning with more reliable clusters, we optimize our model by alternating clustering and contrastive learning. For the person re-ID task, we exploit the learned embedding function $\psi(\cdot)$ to encode each skeleton sequence of the probe set Φ_p into corresponding multi-level graph representation, $\{\overline{\boldsymbol{M}}^{p,i}\}_{i=1}^{N_2}$, and match it with representations, $\{\overline{\boldsymbol{M}}^{g,j}\}_{j=1}^{N_3}$, of the same identity in the gallery set Φ_g using Euclidean distance.

Table 1. Performance comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods on KS20, BIWI-Still (BIWI-S), and BIWI-Walking (BIWI-W) testing sets. "+ FT" denotes employing supervised fine-tuning with labels. The amount of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) for the deep learning based methods are also reported. **Bold** refers to the best cases among self-supervised/unsupervised methods, and *italic numbers* indicate the best performers among supervised methods.

				KS20				BIWI-S				BIWI-W			
Types	Methods	# Params	GFLOPs	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
Hand-crafed	D_{13} [27]	—	—	39.4	71.7	81.7	18.9	28.3	53.1	65.9	13.1	14.2	20.6	23.7	17.2
	D_{16} [30]	—	—	51.7	77.1	86.9	24.0	32.6	55.7	68.3	16.7	17.0	25.3	29.6	18.8
Supervised	PoseGait [26]	8.93M	121.60	49.4	80.9	*90.2*	23.5	14.0	40.7	56.7	9.9	8.8	23.0	31.2	11.1
	SGELA [21] + FT	9.09M	7.48	49.7	67.0	77.1	22.2	29.2	65.2	73.8	23.5	13.9	15.3	16.7	22.9
	MG-SCR [22]	0.35M	6.60	46.3	75.4	84.0	10.4	20.1	46.9	64.1	7.6	10.8	20.3	29.4	11.9
	SM-SGE [23] + FT	6.25M	23.92	49.8	78.1	85.2	11.7	34.8	60.6	71.5	12.8	16.7	31.0	40.2	18.7
	SPC-MGR (Ours) + FT	0.03M	0.22	*65.8*	*82.4*	87.3	*30.6*	*43.8*	*73.6*	*80.5*	*20.3*	*21.5*	*33.9*	*41.0*	*22.9*
Self-supervised/Unsupervised	AGE [20]	7.15M	37.37	43.2	70.1	80.0	8.9	25.1	43.1	61.6	8.9	11.7	21.4	27.3	12.6
	SGELA [21]	8.47M	7.47	45.0	65.0	75.1	21.2	25.8	51.8	64.4	15.1	11.7	14.0	14.7	19.0
	SM-SGE [23]	5.58M	22.61	45.9	71.9	81.1	9.5	31.3	56.3	69.1	10.1	13.2	25.8	33.5	15.2
	SPC-MGR (Ours)	**0.01M**	**0.12**	**59.0**	**79.0**	**86.2**	**21.7**	**34.1**	**57.3**	**69.8**	**16.0**	**18.9**	**31.5**	**40.5**	**19.4**

Given multi-level graph representations $(\boldsymbol{M}_1, \cdots, \boldsymbol{M}_f)$ of an input skeleton sequence $(\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f))$, we first integrate graph features into a *sequence-level* skeleton graph representation:

$$\overline{\boldsymbol{M}} = \frac{1}{f} \sum_{i=1}^{f} w_i \boldsymbol{M}_i \tag{8}$$

where $\overline{\boldsymbol{M}}$ is the multi-level graph representation of skeleton sequence $\boldsymbol{S}_{1:f}$, which incorporates structural-collaborative features and temporal dynamics of f consecutive multi-level skeleton graphs, and w_i denotes the importance of i^{th} skeleton graph representation. Here we assume that each skeleton equally contributes to representing graph features of a sequence, *i.e.*, $w_i = 1$. For clarity, we use $\overline{\mathbb{M}} = \{\overline{\boldsymbol{M}}_i\}_{i=1}^{N_1}$ to represent multi-level graph representations of skeleton sequences in the training set Φ_t, which are exploited as *skeleton instances* in the proposed SPC scheme.

Then, to gather skeleton instances $\overline{\mathbb{M}}$ that contain similar features to find the representative skeleton prototypes, we leverage the DBSCAN algorithm [49], which can discover clusters with arbitrary shapes or semantics, to perform clustering as:

$$\text{DBSCAN}(\overline{\mathbb{M}}) \longrightarrow \overline{\mathbb{M}}^1, \overline{\mathbb{M}}^2, \cdots, \overline{\mathbb{M}}^z, \overline{\mathbb{M}}^o \tag{9}$$

where $\overline{\mathbb{M}} = \overline{\mathbb{M}}^1 \cup \overline{\mathbb{M}}^2 \cup \cdots \cup \overline{\mathbb{M}}^z \cup \overline{\mathbb{M}}^o$, z is the number of clusters (*i.e.*, pseudo classes), $\overline{\mathbb{M}}^k = \{\overline{\boldsymbol{M}}_i^k\}_{i=1}^{x_k}$, $k \in \{1, \cdots, z\}$, is the cluster that contains x_k instances belonging to the k^{th} pseudo class, and $\overline{\mathbb{M}}^o = \{\overline{\boldsymbol{M}}_i^o\}_{i=1}^{x_o}$ denotes the set of outlier instances that do not belong to any cluster. We compute the centroid of each cluster, which *averagely aggregates* features of skeleton instances in the cluster, to obtain corresponding skeleton prototype:

$$\boldsymbol{P}^k = \frac{1}{x_k} \sum_{i=1}^{x_k} \overline{\boldsymbol{M}}_i^k \tag{10}$$

where $\overline{\boldsymbol{M}}_i^k \in \mathbb{R}^{(n_1+n_2+n_3)\times D_h}$ is the i^{th} skeleton instance in the k^{th} cluster, and $\boldsymbol{P}^k$ denotes the k^{th} skeleton prototype.

To focus on the typical and discriminative features of skeleton prototypes as well as facilitate learning high-level skeleton semantics from different prototypes, we propose to enhance the inherent similarity of a skeleton instance to corresponding skeleton prototype and maximize its dissimilarity to other skeleton prototypes with a skeleton prototype contrastive loss as:

$$\mathcal{L}_{\text{SPC}} = \frac{1}{N} \sum_{k=1}^{z} \sum_{i=1}^{x_k} -\log \frac{\exp\left(\overline{\boldsymbol{M}}_i^k \cdot \boldsymbol{P}^k / \tau\right)}{\sum_{j=1}^{z} \exp\left(\overline{\boldsymbol{M}}_i^k \cdot \boldsymbol{P}^j / \tau\right)} \tag{11}$$

where N represents the number of all training instances, z denotes the number of skeleton prototypes, x_k is the number of skeleton instances belonging to the

Multi-level Graph Feature Fusion. To enhance structural semantics of multiple graphs (*e.g.*, global graph patterns) and adaptively integrate key correlative features in component collaboration, we exploit collaborative relations to fuse body-component node features across different spatial levels. We update the node representation ($\widehat{\boldsymbol{v}}_i^a$) of a^{th} level graph by fusing collaborative node features ($\widehat{\boldsymbol{v}}_j^b$) learned from different graphs:

$$\widehat{\boldsymbol{v}}_i^a \leftarrow \widehat{\boldsymbol{v}}_i^a + \sum_{b=a}^{3} \left(\lambda_C^{a,b} \sum_{j=1}^{n_b} \widehat{\mathbf{A}}_{i,j}^{a,b} \, \mathbf{W}_C^{a,b} \, \widehat{\boldsymbol{v}}_j^b \right) \tag{6}$$

where $\mathbf{W}_C^{a,b} \in \mathbb{R}^{D_h \times D_h}$ is a learnable weight matrix to integrate features of collaborative node $\widehat{\boldsymbol{v}}_j^b$ of b^{th} level into a^{th} level node $\widehat{\boldsymbol{v}}_i^a$. n_b denotes the number of nodes in the b^{th} level graph, and $\lambda_C^{a,b}$ represents the fusion coefficient between a^{th} level and b^{th} level graphs, which can be adjusted according to their inherent correlations (*e.g.*, level similarity). We denote the fused l^{th} level graph features of the i^{th} skeleton as $\boldsymbol{F}_i^l \in \mathbb{R}^{n_l \times D_h}$ by concatenating all node representations. Inspired by [23], we retain graph representations of each individual level and adopt their *concatenation* to represent a skeleton as follows:

$$\boldsymbol{M}_i = [\boldsymbol{F}_i^1; \boldsymbol{F}_i^2; \boldsymbol{F}_i^3] \tag{7}$$

where $\boldsymbol{M}_i \in \mathbb{R}^{(n_1+n_2+n_3) \times D_h}$ is the multi-level graph representation of the i^{th} skeleton $\boldsymbol{S}_i$, and $[;]$ indicates the concatenation of graph features. By combining all graph-level representations that integrate structural and collaborative body relation features (see Eq. 1-6), we encourage the model to capture richer features of body structure and skeleton patterns at various levels. Compared with the previous work [22] that adopts a direct graph weighting strategy, the proposed multi-level graph features fusion strategy is *learnable* and can adaptively integrate key relational features among different-level body components to enhance body and motion semantics learning.

3.3 Skeleton Prototype Contrastive Learning Scheme

As skeletons of the same individual typically share highly similar body attributes (*e.g.*, anthropometric attributes) and unique walking patterns [29], it is natural to consider mining the most *typical* attributes or patterns to identify the same person from others. To achieve this goal and encourage the model to capture more high-level skeleton semantics (*e.g.*, class-related patterns), we propose a Skeleton Prototype Contrastive learning (SPC) scheme to focus on the most *representative* skeleton graph features (referred as ***skeleton prototypes***) of pedestrians and exploit their inherent *similarity* and *dissimilarity* with other *unlabeled* graph representations (referred as ***skeleton instances***) to learn general and discriminative representations of each individual. The SPC scheme is built based on the proposed multi-level graph representations and structural-collaborative relation learning, and enables us to learn effective representations from *unlabeled* skeleton data for person re-ID.

the same computation of Eq. 3 to learn a potentially different structural relation, as shown in Fig. 3. We *averagely aggregate* features learned by m different structural relation heads as the representation of node i as follows:

$$\widehat{\boldsymbol{v}}_i^l = \frac{1}{m}\sum_{s=1}^{m}\sigma\left(\sum_{j\in\mathcal{N}_i}(\mathbf{A}_{i,j}^l)^s(\mathbf{W}_v^l)^s\boldsymbol{v}_j^l\right) \tag{4}$$

where $\widehat{\boldsymbol{v}}_i^l \in \mathbb{R}^{D_h}$ denotes the multi-head feature representation of node i in $\mathcal{G}^l$, m is the number of structural relation heads, $(\mathbf{A}_{i,j}^l)^s \in \mathbb{R}$ represents the structural relation between node i and j computed by the s^{th} structural relation head, and $(\mathbf{W}_v^l)^s$ denotes the corresponding weight matrix to perform feature mapping in the s^{th} head. Here we use *average* rather than concatenation operation to reduce feature dimension and allow for more structural relation heads. MSRL enables our model to capture the relations of correlative neighbor nodes (see Eq. 1 and 2) and integrates key spatial features into node representations of each graph (see Eq. 3 and 4). However, it only considers the local relations of the same-level components in graphs and is insufficient to capture global collaboration between different level body components, which motivates us to propose the full-level collaborative relation layer.

Full-Level Collaborative Relation Layer. Motivated by the natural property of human walking, *i.e.*, gait, which could be represented by the dynamic cooperation among body joints or between different body components [29], we expect our model to infer the degree of collaboration (referred as ***collaborative relations***) among body-component nodes in multi-level graphs, so as to capture more unique and recognizable walking patterns from the motion of skeletons. For this purpose, we propose a *full-level collaborative relation layer* (FCRL) to capture relations between a node and all motion-related nodes of the *same level* and that between a node and its spatially corresponding *higher level* body component or other potential components. As shown in Fig. 2 and Fig. 3, we compute collaborative relation matrix $\widehat{\mathbf{A}}^{a,b} \in \mathbb{R}^{n_a\times n_b}$ $(a, b \in \{1, 2, 3\}, a \leq b)$ between the a^{th} level nodes $\mathcal{V}^a$ and the b^{th} level nodes $\mathcal{V}^b$ as following:

$$\widehat{\mathbf{A}}_{i,j}^{a,b} = \mathrm{softmax}_j\left(\widehat{\boldsymbol{v}}_i^{a\top}\widehat{\boldsymbol{v}}_j^b\right) = \frac{\exp\left(\widehat{\boldsymbol{v}}_i^{a\top}\widehat{\boldsymbol{v}}_j^b\right)}{\sum_{k=1}^{n_b}\exp\left(\widehat{\boldsymbol{v}}_i^{a\top}\widehat{\boldsymbol{v}}_k^b\right)} \tag{5}$$

where $\widehat{\mathbf{A}}_{i,j}^{a,b}$ is the collaborative relation between node i in $\mathcal{G}^a$ and node j in $\mathcal{G}^b$. Here we use the inner product of multi-head node feature representations (see Eq. 4) that retain key spatial information of nodes to measure the degree of collaboration. Compared with the previous work [22] that merely considers body relations between adjacent level graphs, FCRL can capture the global collaborative relations among both *adjacent* and *non-adjacent* graphs, and meanwhile provides more comprehensive collaboration inferences between a node and all potential motion-correlated nodes in the same graph.

may act collaboratively in various global patterns during motion [47]. To exploit such internal relations to mine rich body-structure features and unique motion characteristics from skeletons, we propose the multi-level structural relation layer (MSRL) and full-level collaborative relation layer (FCRL) to model the structural and collaborative relations of body components from multi-level skeleton graphs as follows.

Multi-head Structural Relation Layer. To capture latent body structural information and learn an effective representation for each body-component node in skeleton graphs, we propose to focus on features of structurally-connected neighbor nodes, which enjoy higher correlations (referred as ***structural relations***) than distant pairs. For instance, adjacent nodes usually have closer spatial positions and similar motion tendency. Therefore, we devise a *multi-head structural relation layer* (MSRL) to learn relations of neighbor nodes and aggregate the most correlative spatial features to represent each body-component node.

We first devise a basic *structural relation head* based on the graph attention mechanism [48], which can focus on more correlative neighbor nodes by assigning larger attention weights, to capture the internal relation $e^l_{i,j}$ between adjacent nodes i and j in the same graph as:

$$e^l_{i,j} = \text{LeakyReLU}\Big(\mathbf{W}_r^{l\,\mathsf{T}}\left[\mathbf{W}^l_v \boldsymbol{v}^l_i \| \mathbf{W}^l_v \boldsymbol{v}^l_j\right]\Big) \tag{1}$$

where $\mathbf{W}^l_v \in \mathbb{R}^{D \times D_h}$ denotes the weight matrix to map the l^{th} level node features $\boldsymbol{v}^l_i \in \mathbb{R}^D$ into a higher level feature space $\mathbb{R}^{D_h}$, $\mathbf{W}^l_r \in \mathbb{R}^{2D_h}$ is a learnable weight matrix to perform relation learning in the l^{th} level graph, $\|$ indicates concatenating features of two nodes, and LeakyReLU$(\cdot)$ is a non-linear activation function. Then, to learn flexible structural relations to focus on more correlative nodes, we normalize relations using the softmax function as follows:

$$\mathbf{A}^l_{i,j} = \text{softmax}_j\left(e^l_{i,j}\right) = \frac{\exp\left(e^l_{i,j}\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(e^l_{i,k}\right)} \tag{2}$$

where $\mathcal{N}_i$ denotes directly-connected neighbor nodes (including i) of node i in graph. We use structural relations $\mathbf{A}^l_{i,j}$ to aggregate features of most relevant nodes to represent node i:

$$\overline{\boldsymbol{v}}^l_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \mathbf{A}^l_{i,j} \mathbf{W}^l_v \boldsymbol{v}^l_j\right) \tag{3}$$

where $\sigma(\cdot)$ is a non-linear function and $\overline{\boldsymbol{v}}^l_i \in \mathbb{R}^{D_h}$ is feature representation of node i computed by a structural relation head.

To sufficiently capture potential structural relations (*e.g.*, position similarity and movement correlations) between each node and its neighbor nodes, we employ *multiple structural relation heads*, each of which independently executes

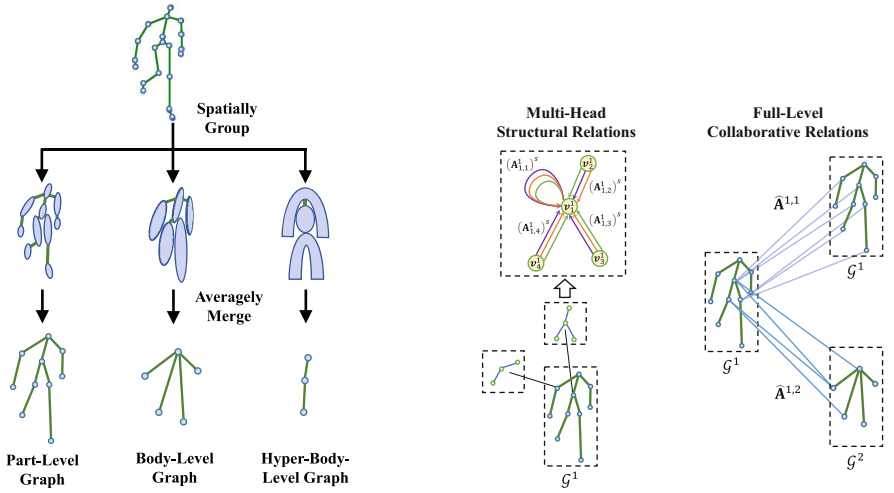


Fig. 3. Left: Three graph levels for a skeleton. We spatially divide human body into 10, 5 and 3 partitions to construct part-level, body-level, and hyper-body-level graphs, and averagely merge internal body joints into nodes. Right: Examples of multi-head structural relations in $\mathcal{G}^1$ and full-level collaborative relations among graphs ($\mathcal{G}^1$, $\mathcal{G}^1$ and $\mathcal{G}^1$, $\mathcal{G}^2$).

define a graph's adjacency matrix as $\mathbf{A}^l \in \mathbb{R}^{n_l \times n_l}$ to represent structural relations among n_l nodes. We compute the *normalized* structural relations between node i and its neighbors, *i.e.*, $\sum_{j \in \mathcal{N}_i} \mathbf{A}^l_{i,j} = 1$, where $\mathcal{N}_i$ denotes the neighbor nodes of node i. $\mathbf{A}^l$ is adaptively learned to capture flexible structural relations in the training stage.

Remarks: Compared with [22] that relies on a specific topology of original skeletons (*e.g.*, joint-level graphs), the proposed unified multi-level skeleton graphs can be viewed as *topology-independent* as they *unify* different skeleton data into an identical number of pre-defined body partitions. It can be generalized to different skeleton datasets and enables the pre-trained model to be directly transferred across different domains for generalized person re-ID (see Sect. 5.4). Besides, they can be extended to skeleton data estimated from RGB videos to learn effective person re-ID representations (see Sect. 5.3). It should be noted that the multi-level graphs can be further extended with different pre-defined body partitions. In our work, we adopt hyper-body-level, body-level, and part-level graphs since their coarse-to-fine body-component divisions match human cognition and prior knowledge of body construction [23,46].

3.2 Structural-Collaborative Body Relation Modeling

The physical connections of body structure typically endow body components in a local partition with higher correlations, while components of different parts

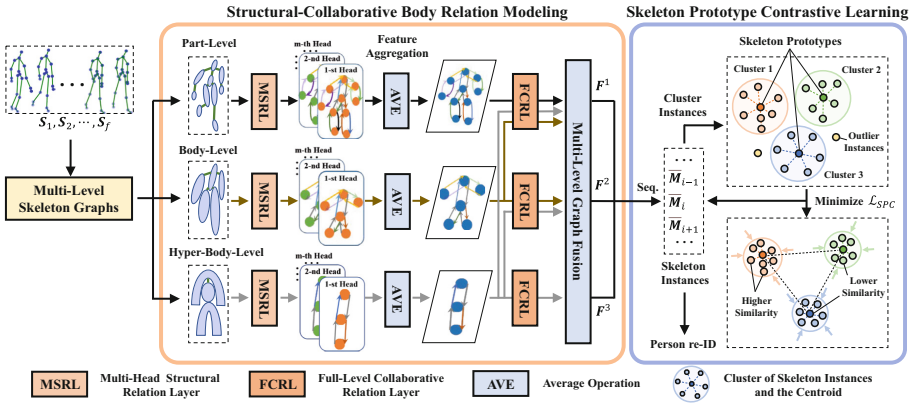


Fig. 2. Schematic diagram of SPC-MGR. Firstly, each 3D skeleton of an input sequence $\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_f$ is represented with part-level, body-level, and hyper-body-level graphs. Secondly, we employ multi-head structural relation layers (MSRL) to capture structural relations of neighbor nodes in each graph, and averagely aggregate features learned by multiple heads to obtain node representations. Then, full-level collaborative relation layers (FCRL) infer the dynamic collaborative relations among the same-level and different-level body components, which are exploited to integrate key graph features into multi-level skeleton graph representations $\boldsymbol{F}^1$, $\boldsymbol{F}^2$, and $\boldsymbol{F}^3$. Next, we perform clustering on skeleton instances, which are sequence-level ("Seq.") multi-level skeleton graph representations, to generate clusters and corresponding skeleton prototypes. Finally, during skeleton prototype contrastive learning, we enhance the similarity of instances belonging to the same prototype and maximize their dissimilarity to other prototypes by minimizing contrastive loss $\mathcal{L}_{\text{SPC}}$. The learned skeleton graph representations are exploited to perform person re-ID.

3.1 Multi-level Skeleton Graphs

Inspired by the fact that human motion can be decomposed into movements of functional body-components (*e.g.*, legs, arms) [23,46], we spatially group skeleton joints to be *higher level* body components at their centroids. Specifically, we first divide human skeletons into several partitions *from coarse to fine*. Based on the nature of body structure, we specify the location of each body partition and its corresponding skeleton joints of different sources (*e.g.*, datasets). Then, we adopt the weighted average of body joints in the same partition as the node of higher level body component and use its physical connections as edges, so as to build unified skeleton graphs for an input skeleton. As shown in Fig. 3, we construct three levels of skeleton graphs, namely *part-level*, *body-level* and *hyper-body-level* graphs for each skeleton $\boldsymbol{S}$, which can be represented as $\mathcal{G}^1$, $\mathcal{G}^2$ and $\mathcal{G}^3$ respectively. Each graph $\mathcal{G}^l(\mathcal{V}^l, \mathcal{E}^l)$ $(l \in \{1, 2, 3\})$ consists of nodes $\mathcal{V}^l = \{\boldsymbol{v}_1^l, \boldsymbol{v}_2^l, \cdots, \boldsymbol{v}_{n_l}^l\}$, $\boldsymbol{v}_i^l \in \mathbb{R}^D$, $i \in \{1, \cdots, n_l\}$ and edges $\mathcal{E}^l = \{e_{i,j}^l \mid \boldsymbol{v}_i^l, \boldsymbol{v}_j^l \in \mathcal{V}^l\}$, $e_{i,j}^l \in \mathbb{R}$. Here $\mathcal{V}^l$ and $\mathcal{E}^l$ denote the set of nodes corresponding to different body components and the set of their internal connection relations, respectively. n_l denotes the number of nodes in $\mathcal{G}^l$. More formally, we

datasets with varying skeletal topologies. Furthermore, a new full-level collaborative relation layer is devised to capture not only the cross-level relations in [22] but also more comprehensive relations among body components at the same level and non-adjacent levels, while a new learnable multi-level graph feature fusion strategy is explored to enhance graph semantics and global pattern learning.

2.2 Contrastive Learning

Contrastive learning has recently achieved great success in many self-supervised and unsupervised learning tasks [17,21,32–36]. Its general objective is to learn effective data representations by pulling closer positive pairs and pushing apart negative pairs in the feature space using contrastive losses, which are often designed based on certain auxiliary tasks (*e.g.*, similarity metrics learning). For example, Wu *et al.* [33] devise an instance-level discrimination method in the form of exemplar task [37] to perform image contrastive learning with noise-contrastive estimation loss (NCE) [38]. In [34], contrastive predictive coding (CPC) based on a probabilistic contrastive loss (InfoNCE) is proposed to learn general representations for different domains. To optimize representation learning (*e.g.*, consistency) in memory bank based contrastive methods [39–41], some recent end-to-end works [42–44] utilize all samples of the current mini-batch to generate negative instance features, while the momentum-based approach [45] further explores the use of momentum-updated encoder and queue dictionary to improve consistency of both encoder and instance features. The PCL [36] integrates both contrastive learning and clustering into an expectation-maximization (EM) framework, which is highly efficient on unsupervised visual representation learning and inspires our work for 3D skeletons.

3 The Proposed Approach

Suppose that a 3D skeleton sequence $\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f) \in \mathbb{R}^{f \times J \times D}$, where $\boldsymbol{S}_t \in \mathbb{R}^{J \times D}$ is the t^{th} skeleton with J body joints and $D = 3$ dimensions. Each skeleton sequence $\boldsymbol{S}_{1:f}$ corresponds to an ID label y, where $y \in \{1, \cdots, C\}$ and C is the number of different persons. The training set $\Phi_t = \left\{ \boldsymbol{S}_{1:f}^{t,i} \right\}_{i=1}^{N_1}$, probe set $\Phi_p = \left\{ \boldsymbol{S}_{1:f}^{p,i} \right\}_{i=1}^{N_2}$, and gallery set $\Phi_g = \left\{ \boldsymbol{S}_{1:f}^{g,i} \right\}_{i=1}^{N_3}$ contain N_1, N_2, and N_3 skeleton sequences of different persons under varying views or scenes. Our goal is to learn an embedding function $\psi(\cdot)$ that maps Φ_p and Φ_g to effective skeleton representations $\{\overline{\boldsymbol{M}}^{p,i}\}_{i=1}^{N_2}$ and $\{\overline{\boldsymbol{M}}^{g,j}\}_{j=1}^{N_3}$ *without using any label*, such that the representation $\overline{\boldsymbol{M}}^{p,i}$ in the probe set can match the representation $\overline{\boldsymbol{M}}^{g,j}$ of the same identity in the gallery set. The overview of the proposed approach is given in Fig. 2, and we present the details of each technical component below.

– We present a skeleton prototype contrastive learning (SPC) scheme based on the proposed multi-level skeleton graph representations to capture representative discriminative skeleton features and high-level class-related semantics from *unlabeled* skeleton data for person re-ID.
– Extensive experiments show that the proposed SPC-MGR outperforms several state-of-the-art skeleton-based methods on four person re-ID benchmarks, and is also highly effective when applied to skeleton data estimated from large-scale RGB videos under more general re-ID settings.

2 Related Works

2.1 Skeleton-Based Person Re-identification

Hand-Crafted Methods. Early skeleton-based works extract hand-crafted descriptors in terms of certain geometric, morphological or anthropometric attributes of human body. Barbosa *et al.* [18] compute 7 Euclidean distances between the floor plane and joint or joint pairs to construct a distance matrix, which is learned by a quasi-exhaustive strategy to extract discriminative features for person re-ID. Munaro *et al.* [27] and Pala *et al.* [30] further extend them to 13 (D_{13}) and 16 skeleton descriptors (D_{16}) respectively, and leverage support vector machine (SVM), k-nearest neighbor (KNN) or Adaboost classifiers for person re-ID. Since such solutions using 3D skeletons alone are hard to achieve satisfactory performance, they usually combine other modalities such as 3D point clouds [31] and 3D face descriptors [30] to improve person re-ID accuracy.

Supervised and Self-supervised Methods. Most recently, a few works exploit deep learning paradigms to learn gait representations from skeleton data for person re-ID in a supervised or self-supervised manner. Liao *et al.* [26] propose PoseGait, which feeds 81 hand-crafted pose features of 3D skeletons into CNN for human recognition. Rao *et al.* [20] devise a self-supervised attention-based gait encoding (AGE) model with multi-layer LSTM to encode gait features from unlabeled skeleton sequences, and then fine-tune the learned features with the supervision of labels for person re-ID. In [21], they further propose a locality-awareness approach (SGELA) that combines various pretext tasks (*e.g.*, reverse sequential reconstruction) and contrastive learning scheme to enhance self-supervised gait representation learning for the person re-ID task. The self-supervised work SM-SGE [23] utilizes a skeleton graph based reconstruction and inference mechanism to encode discriminative skeleton structure and motion features for the person re-ID task.

The most similar work to ours is [22]. Different from [22] that performs supervised skeleton representation for person re-ID, this work proposes the novel skeleton prototype contrastive learning (SPC) to achieve *unsupervised* skeleton-based person re-ID without using labels for more general settings. We for the first explore unified and generalizable multi-level (part-level, body-level, hyper-body-level) skeleton graphs to extend skeleton graph modeling to different skeleton

skeleton features that can re-identify different pedestrians under the unavailability of labels, which limits its application in many real-world scenarios.

To address the above challenges, this work for the first time proposes a generic Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation modeling (SPC-MGR) in Fig. 1 that can comprehensively model body structure and relations at various levels and mine discriminative features from *unlabeled* skeletons for person re-ID. Specifically, we first devise ***multi-level graphs*** to represent each 3D skeleton in a *unified coarse-to-fine* manner, so as to fully model body structure within skeletons. Then, to enable a comprehensive exploration of relations between different body components, we propose to model *structural-collaborative body relations* within skeletons from multi-level graphs. In particular, since each body component is highly correlated with its physically-connected components and may possess different *structural relations* (*e.g.*, motion correlations), we propose a ***multi-head structural relation layer*** (MSRL) to capture multiple relations between each body-component node and its neighbors within a graph, so as to aggregate key correlative features for effective node representations. Meanwhile, motivated by the fact that dynamic cooperation of body components in motion could carry unique patterns (*e.g.*, gait) [29], we propose a ***full-level collaborative relation layer*** (FCRL) to adaptively infer *collaborative relations* among motion-related components at both the *same-level* and *cross-level* in graphs. Furthermore, we exploit a multi-level graph feature fusion strategy to integrate features of different-level graphs via collaborative relations, which encourages the model to capture more graph structural semantics and discriminative skeleton features. Lastly, to mine effective features from *unlabeled* skeleton graph representations (referred as *skeleton instances*), we propose a ***skeleton prototype contrastive learning scheme*** *(SPC)*, which clusters correlative skeleton instances and contrasts their inherent similarity with the most representative skeleton features (referred as *skeleton prototypes*) to learn general discriminative skeleton representations in an unsupervised manner. By maximizing the similarity of skeleton instances to their corresponding prototypes and their dissimilarity to other prototypes, SPC encourages the model to capture more discriminative skeleton features and class-related semantics (*e.g.*, intra-class similarity) for person re-ID *without using any label*. The SPC is devised based on the proposed multi-level skeleton graph representations and structural-collaborative relation learning, and we experimentally and theoretically validate its effectiveness on unsupervised skeleton representation learning for person re-ID tasks.

Our main contributions are summarized as follows:

- We devise unified multi-level graphs to model 3D skeletons, and propose a novel Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation modeling (SPC-MGR) to learn an effective representation from unlabeled skeleton data for unsupervised person re-ID.
- We propose multi-head structural relation layer (MSRL) to capture relations of neighbor body components, and devise full-level collaborative relation layer (FCRL) to infer collaboration between different-level components, so as to learn more structural semantics and unique patterns.

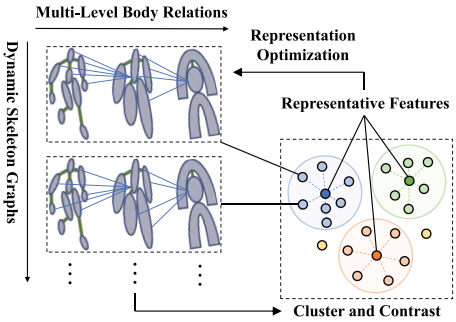


Fig. 1. Our approach constructs skeleton graphs to model multi-level body components and relations, and contrasts the clustered representative features to learn effective skeleton representations for person re-ID.

visual features such as human appearances, silhouettes and body textures from RGB or depth images to discriminate different individuals. Nevertheless, this kind of methods are often vulnerable to appearance, lighting and clothing variation in practice. Compared with RGB-based and depth-based methods, 3D skeleton-based models [18–23] exploit 3D coordinates of numerous key joints to characterize human body and motion, which enjoys smaller data size and better robustness to scale and view variation [24]. With these advantages, 3D skeleton data have drawn surging attention in the fields of person re-ID and gait recognition [20–23,25,26]. However, the way to model discriminative body and motion features with 3D skeleton data remains to be an open challenge.

To perform person re-ID via 3D skeletons, existing endeavors typically model skeleton features by two groups of methods. *Skeleton descriptor based methods* [18,19,27] manually extract certain anthropometric and geometric attributes of body from skeleton data. However, these hand-crafted methods usually require domain knowledge such as human anatomy [28], and cannot fully mine underlying features beyond human cognition. *Deep neural network based methods* [20,21,26] usually leverage convolutional neural networks (CNN) or long short-term memory (LSTM) to learn skeleton representations with sequences of raw body-joint positions or pose descriptors (*e.g.*, limb lengths). Nevertheless, these works rarely explore inherent relations between different body joints or components, which could ignore some valuable structural information of human body. Taking the human walking for example, neighbor body joints "foot" and "knee" have strong motion correlations, while they usually enjoy diverse degree of *collaboration* with limb-level components "leg" and "arm" during movement, which can be exploited to catch unique and recognizable patterns [29]. Other important flaws of this type of methods are label dependency and weak generalization ability. In practical terms, these methods usually require massive labeled data of pre-defined classes to either train the model from scratch [22,26] or fine-tune the pre-trained skeleton representations [20,21,23] to classify the known identities. As a result, they lack the flexibility to learn general and representative

Skeleton Prototype Contrastive Learning with Multi-level Graph Relation Modeling for Unsupervised Person Re-Identification

Haocong Rao[1,2] and Chunyan Miao[1,2(✉)]

[1] LILY Research Centre, Nanyang Technological University (NTU), Singapore, Singapore
{haocong001,ascymiao}@ntu.edu.sg
[2] School of Computer Science and Engineering, NTU, Singapore, Singapore
https://www.ntu.edu.sg/lily

Abstract. Person re-identification (re-ID) via 3D skeletons is an important emerging topic with many merits. Existing solutions rarely explore valuable body-component relations in skeletal structure or motion, and they typically lack the ability to learn general representations with unlabeled skeleton data for person re-ID. This paper proposes a generic *unsupervised* Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation learning (SPC-MGR) to learn effective representations from *unlabeled* skeletons to perform person re-ID. Specifically, we first construct *unified multi-level skeleton graphs* to fully model body structure within skeletons. Then we propose a *multi-head structural relation layer* to comprehensively capture relations of physically-connected body-component nodes in graphs. A *full-level collaborative relation layer* is exploited to infer collaboration between motion-related body parts at various levels, so as to capture rich body features and recognizable walking patterns. Lastly, we propose a *skeleton prototype contrastive learning scheme* that clusters feature-correlative instances of unlabeled graph representations and contrasts their inherent similarity with representative skeleton features ("*skeleton prototypes*") to learn discriminative skeleton representations for person re-ID. Empirical evaluations show that SPC-MGR significantly outperforms several state-of-the-art skeleton-based methods under different scenarios.

Keywords: Skeleton Based Person Re-Identification · Unsupervised Representation Learning · Multi-Level Skeleton Graphs · Skeleton Prototype Contrastive Learning

1 Introduction

Person re-identification (re-ID) aims at identifying or matching a target pedestrian across different views or scenes, which plays an essential role in safety-critical applications including intelligent video surveillance, security authentication and human tracking [1–10]. Conventional studies [11–17] typically utilize

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 196–218, 2024.
https://doi.org/10.1007/978-3-031-65126-7_19

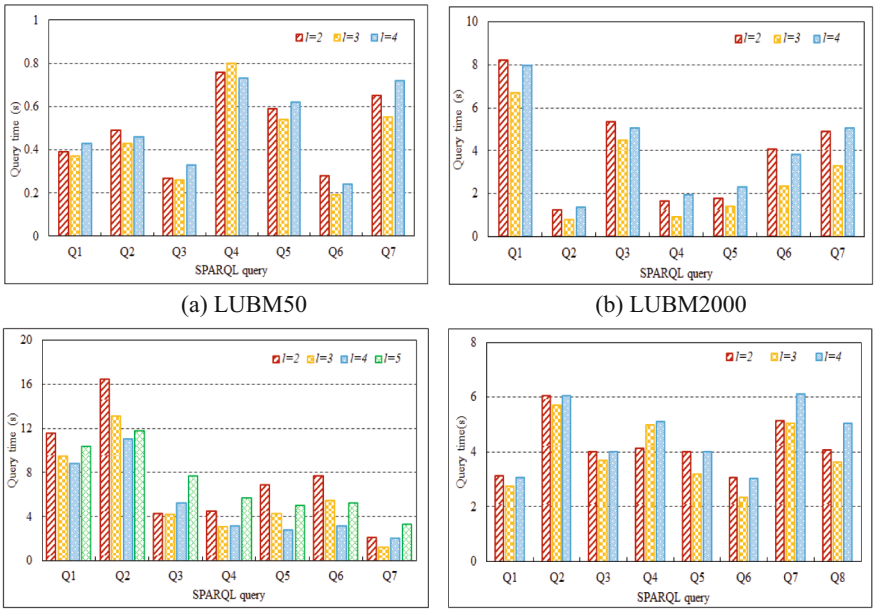
17. Udrea, O., Pugliese, A., Subrahmanian, V.S.: GRIN: a graph based RDF index. In: 22th AAAI Conference on Artificial Intelligence, British Columbia, Canada, pp. 1465–1470(2007)
18. Zou, L., Özsu, M.T., Chen, L., et al.: GStore: a graph-based SPARQL query engine. VLDB J. **23**(4), 565–590 (2014)
19. Wang, D., Zou, L., Feng, Y., et al.: S-store: an engine for large RDF graph integrating spatial information. In: 18th International Conference on Database Systems for Advanced Applications, Wuhan, China, pp. 31–47 (2013)
20. Tran, T., Ladwig, G.: Structure index for RDF data. In: Proceeding of the Workshop on Semantic Data Management (2010)
21. He, H., Wang, H., Yang, J., et al: BLINKS: ranked keyword searches on graphs. In: The ACM SIGMOD International Proceedings on Management of Data,Beijing, China, pp. 305–316 (2007)
22. Kim, K., Moon, B., Kim, H.J.: R3F: RDF triple filtering method for efficient SPARQL query processing. World Wide Web-Internet Web Inf. Syst. **18**(2), 317–357 (2015)
23. Brisaboa, N.R., Ladra, S., Navarro, G.: Compact representation of Web graphs with extended functionality. Inf. Syst. **39**(1), 152–174 (2014)
24. Schmidt, M., Guo, Y., Pan, Z., Heflin, J.: LUBM: a benchmark for OWL knowledge base systems. Web Semant. Sci. Serv. Agents World Wide Web **3**(2), 158–182 (2005)
25. Hornung, T., Lausen, G., et al: SP2Bench: a SPARQL performance benchmark.In: 25th International Proceedings on Data Engineering, Shanghai, China, pp. 222–233 (2009)
26. Uniprot RDF. http://dev.isb-sib.ch/projects/uniprot-rdf/

parallel queries. Of course, distributed query systems involve communication and load balancing issues, which will also be our future research direction.

Acknowledgments. This work is partially supported by the Scientific research project of The Educational Department of Liaoning Provincial under Grant LJ2020016 and Research Institute Project of Bohai University under Grant XK202134-3.

References

1. Ning, Z., Huang, J., Wang, X.: Vehicular fog computing: enabling real-time traffic management for smart cities. IEEE Wirel. Commun. **26**(1), 87–93 (2019)
2. Pirrò, G.: Building relatedness explanations from knowledge graphs. Semant. Web **10**(6), 963–990 (2019)
3. Ning, Z., Kwok, R., Zhang, K., et al.: Joint computing and caching in 5G-envisioned Internet of Vehicles: a deep reinforcement learning-based traffic control system. IEEE Trans. Intell. Transp. Syst. **22**(8), 5201–5212 (2020)
4. Feng, J., Meng, C., Song, J., et al.: SPARQL query parallel processing: a survey. In: IEEE International Conference on Big Data, Boston, USA, pp. 444–451 (2017)
5. Wang, X., Ning, Z., et al.: Offloading in Internet of Vehicles: a fog-enabled real-time traffic management system. IEEE Trans. Ind. Inform. **14**(10), 4568–4578 (2018)
6. Neumann, T., Weikum, G.: The RDF-3X engine for scalable management of RDF data. VLDB J. **19**(1), 91–113 (2010). https://doi.org/10.1007/s00778-009-0165-y
7. Weiss, C., Karras, P., Bernstein, A.: Hexastore: sextuple indexing for semantic web data management. In: 34th International Conference on VLDB, Auckland, New Zealand, pp. 1008–1019 (2008)
8. Mulay, K., Kumar, P.S.: SPOVC: a scalable RDF store using horizontal partitioning and column oriented DBMS. In: 4th International Workshop on Semantic Web Information Management, Scottsdale, USA, pp. 1–8 (2012)
9. Atre, M., Chaoji, V., Zaki, M.J., et al.: Matrix "Bit" loaded: a scalable lightweight join query processor for RDF data. In: 19th International Conference on World Wide Web, Raleigh, USA, pp. 41–50 (2010)
10. Matono, A., Pahlevi, S.M., Kojima, I.: RDFCube: a P2P-based three-dimensional index for structural joins on distributed triple stores. In: The International Conference on Databases, Information System, and Peer-to-Peer Computing, Seoul, Korea, pp. 323–330 (2005)
11. Yuan, P., Liu, P., Wu, B., et al.: TripleBit: a fast and compact system for large scale RDF data. In: 39th International Conference on VLDB, Trento, Italy, pp. 517–528 (2013)
12. Harris, S., Gibbins, N.: 3store: efficient bulk RDF storage. In: 1st International Workshop on Practical and Scalable Semantic Systems,Florida, USA, pp. 1–15 (2003)
13. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: a generic architecture for storing and querying RDF and RDF schema. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 54–68. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-48005-6_7
14. Carroll, J.J., Dickinson, I., Dollin, C., et al.: Jena: implementing the semantic web recommendations. In: 13th International World Wide Web Conference on Alternate Track Papers & Posters,New York, USA, pp. 74–83 (2004)
15. Abadi, D.J., Marcus, A., Madden, S.R., et al.: SW-store: a vertically partitioned DBMS for Semantic Web data management. VLDB J. **18**(2), 385–406 (2009)
16. Ning, Z., Xia, F., Ullah, N., Kong, X., Xiping, Hu.: Vehicular social networks: enabling smart mobility. IEEE Commun. Mag. **55**(5), 16–55 (2017). https://doi.org/10.1109/MCOM.2017.1600263



(a) LUBM50 (b) LUBM2000

(c) Uniprot (d) SP^{2}Bench

Fig. 4. Comparision of query process time on different values of parameter *l*

5 Conclusions

Aiming at the frequent self joins in triple based retrieval and the semantic association characteristics reflected by chain structure information in SPARQL complex query. This paper proposed a bit index structure based on path (PathBit) for large scale RDF Graph. PathBit created predicate path tree (IPT) to filter complete path sets associated with SPARQL query and designed a k^2-tree index (k^{2}TIP) according to the hierarchy of each predicate path tree. k^{2}TIP realized fast association matching of known predicate path triples. Meanwhile, the compression mechanism is used to implement the compressed storage and retrieval algorithm of triples. In addition, two auxiliary indexes: SP and OP are added to assist predicate path retrieval. In the experiment, we compare PathBit with the three existing index storage schemes. The experimental results show that PathBit is very effective for complex queries, especially for queries with long paths. And with the expansion of data scale, PathBit has higher retrieval advantages. At the same time, the storage space of the compressed storage method used in this paper is 0.96 times less than that of RDF-3X in four datasets under the parameter of combined prefix, and has certain advantages over Triplebit.

In addition, with distributed data processing become mainstream, distributed indexing and querying have become research hotspots. The index structure of PathBit can be divided into several sub trees, and the leaf nodes of each sub tree are related to a complete set of paths. Allocating these sub trees to various computing nodes can achieve

4.4 Parameter Analysis

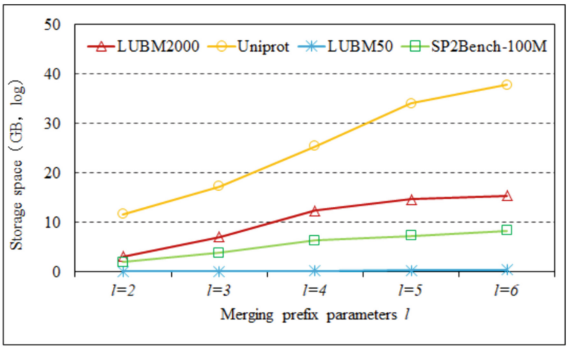


Fig. 3. Comparison of storage space on different values of parameter l

In the process of predicate path merging, the length of common prefix edge is controlled by parameter l. When the length of the common prefix is greater than or equal to l, the edges with the common prefix are merged. This part will test the influence of the parameter l on the storage space.

The range value of l is from 2 to 6. Figure 3 shows that the storage space of all datasets increases with the increase of l. The reason is that the larger the value of l, the smaller the probability of having a common prefix, and the fewer replica nodes that can be merged. In addition, Fig. 3 also shows that with the gradual increase of l value, the change of storage space will be smaller and smaller. When l is set to 4 or 5, the storage space tends to be stable for LUBM and Sp^{2}bench datasets. However, for UniProt, l varies from 5 to 6, because the predicates in UniProt are larger than the other two datasets, which makes the length of common prefix between paths longer.

Figure 4 shows a comparison of query performance. The experimental results show that l has an optimal value, but this value is not directly proportional to the value of l. As shown in Fig. 4, the optimal value of l is 3 for LUBM and Sp^{2}bench datasets, and 4 for UniProt dataset. There are two main reasons. First, when l value is too small, a large number of predicate path are merged, resulting in more candidate paths in the path template matching, which affects the final path matching efficiency. On the contrary, when the value of l is too large, the candidate set becomes smaller and the number of copies increases, which also reduces the query efficiency. Considering the storage space and query performance, the query performance is the best when l takes the storage space to be stable.

The size of Unirpot dataset is 700 million. When the merging common prefix parameter l is set to 3 and the current hardware environment is used to execute the query, Bitmat cannot get the query result. Therefore, Fig. 2(c) only lists the query time comparison of RDF-3X, Triplebit and PathBit. Similarly, after PathBit decomposes the query into query paths, many join operations between triples are decomposed into each path set to complete separately, and the intermediate result set involved in join becomes smaller, especially Q1, Q2, Q5, Q6 and Q7 contain long paths. Combined with the auxiliary indexes, the query performance is superior to Triplebit and RDF-3X. However Q3 and Q4 are star queries, so the overall performance is not as good as long-path retrieval.

The same index is also used to execute queries on SP^2Bench dataset. As with LUBM dataset, the merging common prefix parameter l is also set to 3. Figure 2(d) shows the execution time of each query. Since most of SP^2Bench standard data queries are star structured and query design pays more attention to the use of query operators, the overall query efficiency takes less time, but the filtering and merging operations of the results take a long time. Figure 2(d) depicts the time taken to execute a basic query. Among them, Q1, Q2, Q3 and Q4 are star queries, which are comparable to Triplebit, but better than RDF-3X. However, Q5 and Q7 contain the long path queries, especially Q7, so the query efficiency is significantly improved.

4.3 Comparison of Storage Space

In this part, we compare the storage space of PathBit, RDF-3X and Triplebit. Here, the storage space refers to the space consumed by storing datasets and indexes. Since Bitmat does not contain dictionary tools, the comparison results don't include Bitmat. The merging common prefix parameter l is also set to 3. Table 4 lists the space consumed of different datasets. It can be seen that the storage space of PathBit on all datasets is lower than the other three indexes. As explained earlier, RDF-3X needs to create 6 cluster indexes and 9 clustered indexes. As we all know, the high efficiency of RDF-3X is at the cost of storage space. The dictionary tool used by PathBit is the same as Triplebit. Due to the small number of predicates in LUBM and SP^2Bench and the high merging rate of predicate path, the number of copies is greatly reduced. Therefore, PathBit is better than Triplebit on these two datasets. But the storage space on UniProt dataset is higher than Triplebit.

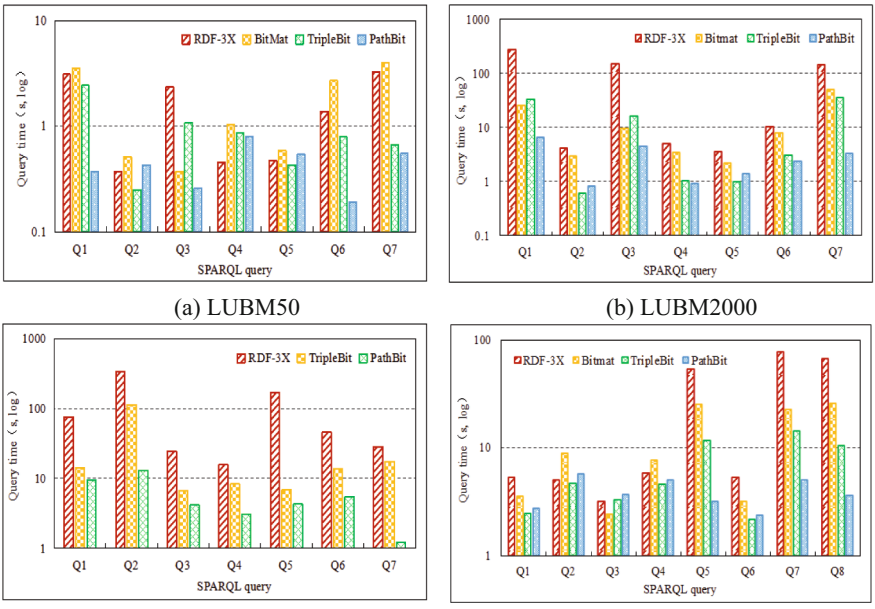
Table 4. Comparision of storage space (GB)

	RDF-3X	TripleBit	PathBit
LUBM50	0.35	0.28	0.19
LUBM2000	13.95	8.74	7.11
Uniprot	33.89	15.19	17.28
SP^2Bench	7.28	4.17	3.88

of query efficiency. The intermediate result in this paper refers to the number of triples matched with the query and the data loaded into memory during the query. Because of the compression method and the direct search in the compressed form, PathBit loads more query data in the same memory and reduce the I/O cost.

It also shows that PathBit is very effective in retrieving large data sets. When the size of LUBM increases from 50 to 2000, the minimum change of query time on RDF-3X is 7.5 times, and the maximum change is 90.83 times, especially for complex queries Q1 and Q3. However, the maximum change of PathBit was only 18.11 times. The reason is that RDF-3X query needs to load more indexes into memory, and at the same time, it also needs to decompress. Therefore, the I/O is larger. Bitmat and Triplebit are both based on triple mode.

In the face of complex queries, they need to join and merge triple more times, so the query performance is lower than that of PathBit. Q2, Q4 and Q5 are star queries. The semantic relevance of predicate path information obtained by star structure is relatively low, but the subject set meeting the conditions is obtained by auxiliary index SP. Combined with subject set and hierarchical edge index, a large number of unrelated triples can be filtered, and the scale of merging results can be reduced. Therefore, the execution efficiency on LUBM 2000 dataset is still better than RDF-3X and Bitmat, which is equivalent to Triplebit. Because the LUBM 50 dataset is small, the index can load memory at once, so RDF-3X retrieval is the highest.



(a) LUBM50 (b) LUBM2000

(c) Uniprot (d) SP²Bench

Fig. 2. Comparision of query performance

Table 2. Connection types of triple pattern2

Category	Type
1	$s\ ?p_1\ ?x \bowtie ?x\ ?p_2\ o$
2	$?s\ ?p_1\ ?x \bowtie ?x\ ?p_2\ o$
3	$?s\ ?p_1\ ?x \bowtie ?x\ ?p_2\ ?o$

4 Experiments

In this section, PathBit indexing scheme is tested on synthetic and real datasets.

4.1 Datasets and Setting

Table 3. Test datasets

Data set	Vertex	Triple	Predicate
LUBM50	1,706,230	6,888, 642	18
LUBM2000	66,059, 204	276, 345,040	18
SP2Bench	56,125,032	113,246,165	22
Uniprot	139,942,781	687,025,165	84

In the experiment, two synthetic datasets LUBM and Sp^2Bench were selected. The LUBM features a university domain, and the SP^2Bench dataset features a DBLP domain [24, 25]. In our experiments, we also use a protein dataset Uniprot [26] (Table 3).

PathBit index is written in C++ and compiled with GCC. We select the optimization level of O2. The experiment runs on a server with Intel Xeon 2.00GHz processor and 20GB memory. Considering the influence of warm cache on experimental error, each query is executed five times, and the arithmetic average is taken as the final experimental result.

4.2 Comparison of Query Performance

In the experiment, LUBM data set generates 81 predicate paths. If the merging common prefix parameter l is set to 3, we obtain 26 predicate path trees. Figure 2(a) and (b) show the query execution time of SPARQL. The query time of Q1, Q3, Q6 and Q7 are better than the other indexes. These four queries have longer join paths than the star queries Q2, Q4 and Q5. Using the path association information to search can filter a large number of unrelated triples and narrow the retrieval range. Hierarchical path index decomposes the connection between triples into smaller ones, which reduces the connection size of triples and improves the matching efficiency. Intermediate results are also an affecting factor

According to Theorem 4, given a SPARQL query G_q, G_q is decomposed into a complete path set $QCPath = \{R_1^q, R_2^q, \ldots R_n^q\}$. If G_q is a subgraph of G, then $\forall R_i^q \in QCPath$, there must exist at least one complete path R_i, satisfying R_i^q is a subpath of R_i. Therefore, a SPARQL query need to be decomposed into multiple search paths from the source vertex to the sink vertex. The decomposition principle is to follow the full coverage of vertices and edges. That is, starting from any source vertex, if the decomposed full path already contains all the edges in the query, the decomposition ends. Because the complete path decomposition already includes all possible complete paths, there must be a complete path corresponding to it.

According to whether the predicate path contains a constant, the decomposed search path can be divided into two categories: constant predicate path and variable predicate path. Constant predicate refers to the path containing one or more known predicates, while variable predicate refers to the path in which all predicates are unknown.

The analysis result shows that most of predicate paths of SPARQL queries are constant predicate path. For constant predicate paths, the retrieval is performed on the IPT index to obtain the candidate complete paths containing known predicates. Then according to the connection relationship of adjacent predicates, they are divided into six types as shown in Table 1. When performing the retrieval, the type of adjacent predicate is judged from the source vertex in turn, and is executed in the order from low to high.

Table 1. Connection Types of triple pattern1

Category	Type
1	$s\,p_1\,?x \bowtie ?x\,p_2\,o$
2	$?s\,p_1\,?x \bowtie ?x\,p_2\,o$
3	$?s\,p_1\,?x \bowtie ?x\,p_2\,?o$
4	$s\,p_1\,?x \bowtie ?x\,?p_2\,o$
5	$?s\,?p_1\,?x \bowtie ?x\,p_2\,o$
6	$?s\,?p_1\,?x \bowtie ?x\,p_2\,?o$

When the search path is a variable predicate path, the adjacent predicate connections can be divided into three types as shown in Table 2. Since k^2TIP is only applicable to the case that there is a constant predicate in the retrieval path. For the variable predicate path, not only IPT is invalid, but k^2TIP is also invalid. In order to ensure the validity of the index, we design two auxiliary indexes, namely SP and OP, to solve this problem. SP and OP store all predicates corresponding to each subject or object, respectively. SP and OP indexes adopt the compressed representation and retrieval method proposed in Triplebit, which will not be explained in detail here.

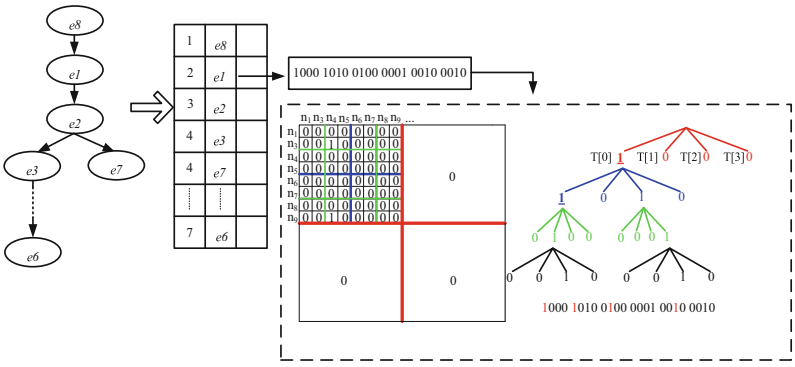


Fig. 1. Example of k^2TIP

node of the root node in k^2 tree. If the bit element of the sub matrix contains 1, then the corresponding sub node of k^2 tree is 1, otherwise the sub node is 0. After the first level node is created, the matrix corresponding to the node with the value of 1 will continue to be divided in the same way until the sub matrix is 0 or there are only k^2 bits in the sub matrix.

The generated k^2-tree connects nodes corresponding to 0 or 1 from top to bottom and left to right to form a compressed bit string. Figure 1 describes the generation process of bit matrix, k^2-tree and compressed bit string of a predicate path tree. For a triple set of a predicate, it is usually necessary to search all the objects corresponding to a known subject, or to search all the subjects corresponding to the known subject. As shown in Fig. 2 shows all objects associated with the subject n_3, where the id of n_3 is 2 and the value of k is 2. In order to get the all objects associated with n_3, we have to find all the columns with cell 1 in the row of n_3 in the matrix. The specific steps are as follows:

Step 1: starting from the root node of k^2-tree, set its position *pos* as 0, and its four sub nodes corresponding to the four sub matrices respectively. The sub matrix corresponding to the first two sub nodes intersect with the row of n_3, while the other two child nodes has no association with n_3. So we only need to consider the first two sub nodes. Set the two sub matrices to T[0] and T[1]. T[1] sub matrix is 0, which means that there is no object associated with n_3 in T[1]. In order to get the number of the column in which the object associated with n_3, it is necessary to record the starting position of the corresponding column of the incidence sub matrix when locating the incidence sub matrix. For example, the starting position of T[0] sub matrix is 0.
Step 2: in the compressed bit string, the starting position *pos* of sub node corresponding to T[0] is 4. The corresponding bit of four sub matrices of T[0] is '1010'. The id of n_3 indicates that the sub matrix associated with it is T[0][0] and T[0][1], where T[0][0] is 1 and T[0][1] is 0. And the starting position of T[0][0] sub matrix is 0.
Step 3: according to this method, continue to search down until the leaf node. If the leaf node is 1 and satisfies the association with n_3, then the column corresponding to the node is the object associated with n_3.

bit of the bit string corresponds to a unique predicate. Supposing $E(ppt_i)$ is the predicate set of a predicate path tree, where ppt_i is the ith predicate path tree in *PPtree*. If $\exists pre \in ppt_i$, set the *eId* (*pre*) bit of the bit string to 1, where *eId* (*pre*) represents the id of the predicate *pre*.

The establishment of IPT is based on the bottom-up process. Each leaf node of IPT corresponds to a predicate path tree, and each path template tree corresponds to a full path set. Non-leaf nodes of IPT are obtained by performing a logical 'OR' operation on their sons.

3.3 Retrieval of Predicate Path Tree

When retrieving the query predicate path tree on the IPT index tree, we encode the query path tree in the same way according to Definition 8. If the query predicate path tree includes predicate variables, the code of predicate variable is set to 0. Then use the top-down method to search for the match paths in IPT index tree. If the query predicate path tree and the node of IPT meet the matching principle of Definition 9, the search continues, otherwise the subtrees corresponding to the unmatched node on IPT index tree are pruned.

Definition 9 (Matching Principle). Given a bit string of predicate path tree bit string *ppt** and a query path bit string *qtt**, if *ptt** matches *qtt**, if and only if the logical 'and' operations satisfy AND(*ptt**, *qtt**) = *qtt**.

3.4 Match of Complete Path

Retrieved results in IPT are a candidate set of complete path. In order to get the final query results, it is necessary to accurately match the candidate path set. In this section, we will create a k^2-tree index (k^2TIP) for the corresponding complete path collection according to the hierarchy of each predicate path tree. The k^2TIP adopts two stage compression modes. Figure 1 shows the k^2-tree index structure. The k^2TIP index contains each edge in the predicate path template tree and its corresponding hierarchical ID information, and each edge points to a storage area, which is used to store the triple set associated with the edge in the whole path. Since all triples of this set have a common predicate, we use k^2-tree [23] structure to store triples for each triple set, and compress triples on this basis.

Each predicate in the predicate path tree corresponds to a triple set, and these triples share the same predicate. In order to reduce the storage space, we compress each triple set, and the compression method adopts k^2-tree.

k^2-tree first uses a two-dimensional bit matrix to establish the corresponding relationship between subject and object. If there is a corresponding relationship between the subject and the object, the corresponding bit is set to 1, otherwise it is 0. A large number of subjects and objects are not related, so the bit matrix is a sparse matrix. Therefore, we divide the bit matrix into k^2 sub matrices, and each sub matrix corresponds to a sub

Proof: Suppose there is no complete path $Ri \in CPath$, satisfying R_i^q is a subpath of R_i. (1) If the source vertex v_0 and sink vertex v_m of the complete query path R_i^q in G_q is also the source and sink vertices of G, then this will conflict with the hypothesis, because there must be a complete path between v_0 and v_m. (2) If the source vertex v_0 and sink vertex v_m of R_i^q in G_q is not the source and sink vertices of G, then there must be a source vertex v_s, which v_s to v_0 is reachable and there is also a sink vertex v_e, which v_m to v_e is also reachable. That is to say, there is at least one complete path from v_s to v_e and v_0 to v_m is a subpath of the whole path, which contradicts the hypothesis. To sum up, the hypothesis does not hold, that is, there is at least one complete path $R_i \in CPath$, satisfying R_i^q is a subpath of R_i.

Definition 5 (Predicate Path). Given an RDF graph G, according to the Definition 4, G is decomposed into a set of complete paths. We use $CPath = \{R_1, R_2, \ldots, R_m\}$ to represent it. For any path R, Extract the edge information of the path to construct a summary path $E(R_i) = \{e_1, e_2, \ldots, e_m\}$, then $E(R_i)$ is called predicate path.

Definition 6 (Isomorphism Path). If the complete paths have the same predicate path, they are called to be isomorphic paths.

Obviously, after decomposing the RDF graph into the set of complete path, a large number of vertices and edges repeatedly appear in different complete paths, which puts a lot of pressure on data storage. However, many complete paths have similar predicate path, that is, many vertices information is the same. If these complete paths are divided into the same class, the number of copies of vertices will be reduced. So we define the following two conditions to merge predicate path.

Definition 7 (Predicate Path Tree). If two or more predicate paths meet the following conditions: (i) Two or more predicate paths have a common prefix and the length of the edge of the common prefix is greater than or equal to a threshold. (ii) A predicate path is the suffix of another predicate path. We will merge these paths into one predicate path tree. The predicate path tree is denoted as *PPtree.*

3.2 Index of Predicate Path Tree

Each predicate path tree corresponds to a complete path set. For a SPARQL query, it is decomposed into several query paths in the same way. We can obtain the complete path set corresponding to each query path by retrieving the predicate path tree. In order to quickly locate the complete path set of the query paths, we create an index based on predicate path tree (IPT). The establishment process of IPT mainly include three steps: the first is to code the predicate path tree, the second is the construction process of IPT and the last is how to retrieve IPT.

Definition 8 (Encoding Predicate Path Tree). Assign a unique id to each predicate in L in order. Obviously, the maximum id is the number of elements in L, which is represented by ℓ. The encoding of predicate path tree is a bit string of length ℓ, and each

filtering operations are used to filter out irrelevant data in the input triples. TripleBit [11] vertically divides the triple matrix based on predicates, and sorts triples with the same predicate in the order of subject or object. During the query process, two index structures was introduced to minimize the cost of index selection. In summary, path, compression and index tree are very effective techniques for improving query efficiency and reducing storage space.

3 PathBit

PathBit includes an index based on predicate path tree (IPT) and a k^2-tree index (k^2TIP) according to the hierarchy of each predicate path tree. IPT is in charge of the filter of complete path set, which related to the retrieval path. k^2TIP according to the hierarchy of each predicate path tree to realize fast association matching of known predicate path triples. Meanwhile, the compression mechanism is used to implement the compressed storage and retrieval algorithm of triples. In addition, two auxiliary indexes: SP and OP are added to assist predicate path retrieval.

3.1 Complete Predicate Path

An RDF database is a set of RDF triples, we use $T = \{t \mid t \in S \times P \times O\}$ to describe the dataset, where S, P, O are the set of subjects, predicates and objects, respectively.

Definition 1 (Path). Given an RDF $G = (V, E, L)$, a path is a set of ordered vertices, denoted by $R = (v_0 v_1 v_2 \ldots v_m)$, $\forall k \in [0, m-1]$, $<v_k, v_{k+1}> \in E$.

Definition 2 (Complete Path). For any path R in RDF graph, if v_0 is a source vertex and v_m is a sink vertex, we say that R is a complete path. We use $CPath = \{R_1, R_2, \ldots, R_m\}$ *to* denote a set of RDF complete paths.

Theorem 3. Given an RDF G, $\forall v \in V$ and $e(u, v) \in E$ must belong to at least one complete path.

Proof: (1) Assuming SV is the set of source vertices. For any vertex v in graph G, there are two states. The first is $v \in SV$. If $v \in SV$, because any complete path starts from a source vertex, v must exist in a complete path. The second is $v \notin SV$. If $v \notin SV$, then there must be a source vertex s, so that s to v can be reached, that is, the vertex v belongs to a complete path whose source vertex is s. If s doesn't exist, then v becomes the source vertex, which conflicts with the condition. Therefore, for any vertex v in set V must belong to at least one complete path. (2) For any edge $e(u, v) \in E$ in graph G, if it does not belong to any complete path, then the two vertices u or v do not exist in any complete path, which is in contradiction with that any vertex $v \in V$ belongs to at least one complete path. Therefore, any edge $e(u, v) \in E$ belongs to at least one full path.

Theorem 4. Given a SPARQL query G_q, according to the Definition 4, G_q is decomposed into a set of complete query paths. We use $QCPath = \{R_1^q, R_2^q, \ldots R_n^q\}$ to represent it. If G_q is a subgraph of G, then $\forall R_i^q \in QCPath$, there must exist at least one complete path R_i, satisfying R_i^q is a subpath of R_i.

2 Related Work

In order to improve the retrieval efficiency, researchers have conducted extensive research on the storage and index of RDF. This paper analyzes the current research status from three different perspectives.

RDF storage and index technology based on relationships utilizes relational database query technology to convert SPARQL queries into SQL to realize data retrieval. 3-Store [12] and Sesame [13] all use triple tables. Due to all data exists in a large table, SPARQL queries are easy to result in many self joins and decrease the query efficiency. Jena2 [14] uses an attribute table, which greatly reduces self joins and merge operations. But not all objects have the same properties, which leads to a large number of empty values. In addition, a large number of multi-valued attributes can also generate more multi-valued dependencies. Therefore, the attribute table is not a universal storage model. SW-store [15] decomposes triples based on the predicate, storing triples with the same predicate in the same table. For the two columns table, a subject based clustered index can be created to achieve rapid subject localization. This scheme not only reduces the merging operation of the same predicate, but also avoids the control problems caused by the attribute table.

Triple index scheme is a combination and permutation of S, P and O. RDF-3X [6] stores all permutations and combinations of subject, predicate and object on a B+ tree, respectively. Moreover, RDF-3X also combines two or single elements to directly form a clustered index. Similar to RDF-3X, Hexastore [7] also establishes six indexes based on the triple table. The difference is that in the establishment process of index, Hexastore considers the order relationship between the subject, predicate and object. Meanwhile, Hexastore will reduce the redundancy of memory by sharing index lists. SPOVC [8] creates five index types based on subject, predicate, object, object data types, and triple classes. Each index type was horizontally segmented according to certain rules, which is effective for the query of range or rule expressions.

Bitmat [9] and RDFcube [10] map S, P and O into a three-dimensional space to form a three-dimensional matrix. Each element in the matrix corresponds to a triple. Bitmat is a memory based bit matrix primarily used to handle concatenation operations in triple pattern. Although these two index types utilize bit technology to achieve high compression of triples, they face large-scale data, especially Bitmat, which makes it difficult to load the index into memory at once.

RDF itself is a directed graph, so SPARQL query can be seen as a sub graph matching problem. GRIN [17] indexes RDF graphs with a balanced binary tree. By utilizing the distance conditions, it can quickly filter the data that does not meet the criteria. But GRIN index has poor scalability. Zou et al. [18, 19] proposed VS-tree and VS*-tree index to handle precise and wildcard SPARQL queries. PIG [20] (Parameterized Index Graph) index corresponds to a set of vertices with similar or identical neighborhood structures in the original data graph. PIG first retrieve edges that are homomorphic to the edges in the query graph to form a set of candidate edges, and then perform join operations in the set of candidate edges. He et al. [21] proposed a two-layer index scheme (BLINKS) for searching the top-k keywords on a graph, which only supports searching on node labeled directed graphs. In order to reduce redundant intermediate results, RP-index [22] creates a path based index to index the RDF graph in-edge. During the executive process,

The rapid increase of RDF data brings great challenges to traditional data storage, index and query. The triple table, vertical partition and attribute table use an alternative relational storage mode and mature management mechanism of relational database to accelerate data retrieval. However, these relational data models could not fully reflect the logical structure of RDF data [5].

Some native storage systems, such as RDF-3X [6], Hexastore [7] and SPOVC [8], store multiple copies of data according to different combinations of subject, predicate and object to assist in generating better query plan. Although the query efficiency is improved, these systems are at the expense of storage space. Bitmat [9] and RDFcube [10] use three-dimensional matrix to store triples, and divide the three-dimensional matrix into two-dimensional matrices along a certain dimension. For each two-dimensional matrix, D-gap compression method is used for row compression storage. However, in the face of large-scale data, it is difficult for Bitmat to load all indexes into memory at one time. Triplebit [11, 12] reduces the storage scale of RDF-3X, and only stores two combinations of subject and object (SO) and object and subject (OS) based on predicate. At the same time, Triplebit establishes the corresponding index according to the predicate and realizes the compressed storage. For a SPARQL query, Triplebit generates the corresponding query plan according to certain heuristic rules, and dynamically modifies the query plan to reduce the intermediate results.

All these storage systems view triple mode as retrieval unit to realize data retrieval. Through certain query plan and optimization technology, these storage systems reduce intermediate results, and realize fast connection of intermediate results. These indexes are based on triples and do not consider the structure and semantics of RDF graphs. As mentioned above, triple patterns in SPARQL queries have certain connection relations, which not only reflect the structural information of RDF graphs, but also reflect certain semantic relations. This paper proposes a bit index structure based on path (PathBit) for large scale RDF Graph. The major contributions include:

(1) Taking the complete path from source to sink point in RDF graph as the structure object, we create the bit index based on predicate path tree to realize the retrieval and filtering mechanism, and reduce the connection scale of intermediate results in triple pattern matching;
(2) For each complete path tree, we will create a k^2-tree index (k^2TIP) according to the hierarchy of each predicate path tree to realize fast association matching of known predicate path triples;
(3) We use k^2-tree compression mechanism to implement the compressed storage and retrieval algorithm of triples. At the same time, two auxiliary indexes: SP and OP are added to assist predicate path retrieval.

The other parts of the paper are summarized as follows: part two introduces the relate works; part three describes the design scheme of PathBit in detail; finally, experiments verify the performance of PathBit and draw the conclusion.



PathBit: A Bit Index Based on Path for Large-Scale Knowledge Graph

Yonglin Leng(✉), Peiyi Qu, Ying Guo, and Chaoliang Xi

College of Information Science and Technology, Bohai University, Jinzhou 121000, China
lengyonglin@qq.com

Abstract. As the latest achievement of symbolism, knowledge graph is an important cornerstone of artificial intelligence. In order to better manage the knowledge graph, RDF triples have been used to represent knowledge graph. The rapid growth of data brings great challenges to knowledge graph storage and quick retrieval. Among them, self joins, high storage cost and intermediate results are the main problems. In this paper, we propose a bit index structure based on path (PathBit) for large scale knowledge graph. PathBit includes an index based on predicate path tree (IPT) and a k^2-tree index (k^2TIP) according to the hierarchy of each predicate path tree. IPT is in charge of the filter of complete path set. k^2TIP according to the hierarchy of each predicate path tree to realize fast association matching of known predicate path triples. Meanwhile, the compression mechanism is used to implement the compressed storage and retrieval algorithm of triples. In addition, two auxiliary indexes: SP and OP are added to assist predicate path retrieval. Finally, we conduct a series of experiments on two representative datasets and compare the results with RDF-3X, Bitmat and TripleBit. Results indicate that PathBit can achieve better response time on complex queries and has greater advantages in storage space compared with RDF-3X and Bitmat.

Keywords: Knowledge Graph · Index · Predicate Path · Compressed storage

1 Introduction

As the supporting foundation of AI, knowledge graph shows more and more value in semantic search, intelligent question answering, data analysis, natural language processing, vision understanding and IoT. In order to better manage the knowledge graph, RDF triples have been used to represent knowledge graph. The rapid growth of RDF data also bring great challenges to query. SPARQL is the most widely used query language in RDF data query [1, 2]. A SPARQL query includes many triple patterns, which can also be described as a directed query graph. The query graph generally consists of four basic sub graphs: star, chain, ring and tree topology [3, 4]. The basic sub graph has a lot of connections. These connection relations can be divided into chain and star relations. The chain relation refers to the subject of triple pattern is the object of another triple pattern. Star relation refers to a group of triple patterns with the same subject or object. In these basic sub graphs, the chain relation is an important structure in SPARQL. Because the ring and tree query all contain chain structure.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 181–195, 2024.
https://doi.org/10.1007/978-3-031-65126-7_18

Reliability and Scalability

2. Mahmoud, M.S., Hamdan, M.M., Baroudi, U.A.: Modeling and control of cyber-physical systems subject to cyber attacks: a survey of recent advances and challenges. Neurocomputing **338**, 101–115 (2019)
3. Deng, Z., Xiong, C., Cai, M.: An autonomous transportation system architecture mapping relation generation method based on text analysis. IEEE Trans. Comput. Soc. Syst. **9**(6), 1768–1776 (2022)
4. Zhou, Z., Cai, M., Xiong, C., et al.: Construction of autonomous transportation system architecture based on system engineering methodology. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 3348–3353. IEEE (2022)
5. You, L., He, J., Wang, W., et al.: Autonomous transportation systems and services enabled by the next-generation network. IEEE Netw. **36**(3), 66–72 (2022)
6. Graja, I., Kallel, S., Guermouche, N., et al.: A comprehensive survey on modeling of cyber-physical systems. Concurr. Comput. Pract. Experience **32**(15), e4850 (2020)
7. Tantawy, A., Abdelwahed, S., Erradi, A., et al.: Model-based risk assessment for cyber physical systems security. Comput. Secur. **96**, 101864 (2020)
8. Schranz, M., Di Caro, G.A., Schmickl, T., et al.: Swarm intelligence and cyber-physical systems
9. Deka, L., Khan, S.M., Chowdhury, M., et al.: Transportation cyber-physical system and its importance for future mobility. Transp. Cyber-Phys. Syst. 1–20 (2018)
10. Hussain, M.D.M., Beg, M.M.S.: Using vehicles as fog infrastructures for transportation cyber-physical systems (T-CPS): fog computing for vehicular networks. Int. J. Softw. Sci. Comput. Intell. (IJSSCI) **11**(1), 47–69 (2019)
11. Lin, J., Yu, W., Zhang, N., et al.: Data integrity attacks against dynamic route guidance in transportation-based cyber-physical systems: modeling, analysis, and defense. IEEE Trans. Veh. Technol. **67**(9), 8738–8753 (2018)
12. Hou, Y., Zhao, Y., Wagh, A., et al.: Simulation-based testing and evaluation tools for transportation cyber–physical systems. IEEE Trans. Veh. Technol. **65**(3), 1098–1108 (2015)
13. Zhang, L., Jiang, S., Huang, K., et al.: Knowledge graph-based network analysis on the elements of autonomous transportation system. In: 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 536–542. IEEE (2021)
14. Deng, Z., Xiong, C., Cai, M.: Research on theoretical model and construction method of the physical object for autonomous transportation system. In: CICTP 2022, pp. 647–655 (2022)
15. Deng, Z., Cai, M., Xiong, C.: An architecture integrity simulation evaluation method for an autonomous transportation system based on an information-triggered collaboration mechanism. IEEE Intell. Trans. Syst. Mag. https://doi.org/10.1109/MITS.2023.3272501
16. Younis, O., Moayeri, N.: Employing cyber-physical systems: dynamic traffic light control at road intersections. IEEE Internet Things J. **4**(6), 2286–2296 (2017)
17. Guzmán, J.A., Núñez, F.: A cyber-physical systems approach to collaborative intersection management and control. IEEE Access **9**, 99617–99632 (2021)
18. Kamal, M.A.S., Tan, C.P., Hayakawa, T., et al.: Control of vehicular traffic at an intersection using a cyber-physical multiagent framework. IEEE Trans. Ind. Inf. **17**(9), 6230–6240 (2021)

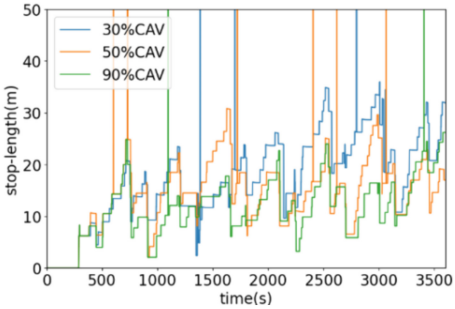


Fig. 6. Comparison of queue length under Webster adaptive control under different CAV penetration rates

of research on the application and modeling problems of CPS in the transportation industry, and identifies the key problems in the development of TCPS: the lack of complete and effective theoretical guidance and the lack of scenario-level TCPS modeling.

In order to solve this problem, this paper, with the help of the theoretical framework of ATS, starts from the five categories of ATS: service, technology, demand, function and technology, and focuses on the connection of function, transmission of data flow, and connection of entity to form the three categories of functional, logical and physical architectures respectively, and analyzes the typical traffic scenarios according to this idea to form a complete scenario architecture.

Subsequently, under the guidance of intersection scenario architecture, this paper draws on the design of cyber layer and physical layer of information-physical system, and adopts Netlogo, a multi-intelligence approach software, to simulate physical objects, information flow and information interaction pairs in the scenario architecture respectively, and constructs the physical layer of intersection scenario; adopts Python language for joint simulation, and designs the flow change process of information flow in the cyber layer. After the model was successfully constructed, it was simulated and the simulation results were analyzed qualitatively for functional integrity and quantitatively for traffic indicators.

However, the current study still has certain shortcomings. First, in the physical layer of the simulation model, simulation software as well as variables are still used to simulate the real world in the physical layer, which cannot really simulate the variability and complexity of the real world. Second, in the cyber layer of the information out process, there are still areas that can be improved. The next step is to combine the work of this paper with real-world intersection data, compare the differences between this model and the actual data, and use it to continuously adjust the model to form a more accurate and more reflective real-world intersection TCPS model.

References

1. Chen, H., Cai, M., Huang, K., et al.: Classification and evolution analysis of key transportation technologies based on bibliometrics. Sci. Program. **2021**, 1–13 (2021)

Table 3. Traffic participants and usage information flows in similar literature

Author of the literature	Traffic participants in the model	The flow of information that the model leverages
Younis O	Cars and Signal Lights	Vehicle coordinates and speed information, traffic flow information
Guzman J A	Cars and Signal Lights	Vehicle coordinates and speed information, traffic flow information
Kamal M A S	Cars and Signal Lights	Vehicle coordinates and speed information, traffic flow information
This article	Cars and Signal Lights, Roadside infrastructures, traffic operation center, meteorological information center	Vehicle coordinates and speed information, traffic flow information, road surface information, weather condition information, road congestion information

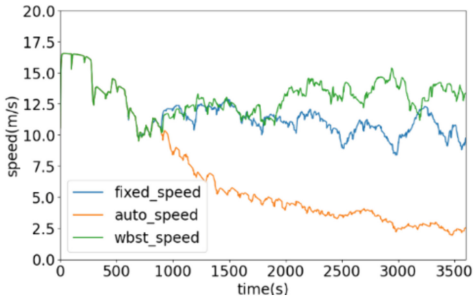


Fig. 5. Comparison of average vehicle speed of three types of phase control methods with 30% CAV

It can be found that CAV has a significant improvement in queue length at higher penetration rates, but the improvement is not significant enough, which may be related to the fact that the control method adopted lacks vehicle-road cooperative driving and cannot fully utilize the potential of CAV traffic.

5 Conclusion

This paper reviews the key technologies that have helped the rapid development of the transportation industry in recent years, and provides an overview of CPS, which has gradually emerged with the development of various technologies, especially the current state

4.4 Simulation Conclusion

1. Integrity Analysis

During the simulation, the real-time operation of CPS can be simulated, and the integrity of the operation of information flow such as weather information, vehicle information, and infrastructure information can be properly demonstrated respectively. The integrity test of the control schemes such as vehicle speed limit and no left turn can operate normally. The following figure shows some of the information flows involved in the intersection scenario and the results of real-time information access during the simulation (Table 2).

Table 2. Simulation test results of perceptual information flow in scenario architecture

Information Flow	Form of embodiment	Test results
Environment information	Every 60 s, the Weather Center perceives environmental information	Normal
Traffic flow basics	Average speed, queue length	Normal
Traffic flow information	Traffic flow	Normal
Lane monitoring data	Zoning according to roadside facilities (Location and time of infrastructure failures)	Normal
Road guardrail monitoring data		
Road sign marking monitoring information		
Vehicle driving condition monitoring information	Whether the vehicle is faulty (Fault number and time)	Normal
Vehicle location and movement information	Vehicle location information	Normal

The following table compares the intersection TCPS functions designed in this paper with similar literature [16–18] (Table 3):

2. Traffic Indicator Analysis

The model proposed in this paper can also be applied to traditional traffic microsimulation, and the diversity of agents in it can help support traffic flow simulation in future heterogeneous mixing phases.

The following results were obtained by simulating different signal control schemes for CAV penetration of 30% at a traffic volume of 360veh/h and analyzing the average vehicle speed when different signal control schemes were used (Fig. 5):

It is easy to see that the traffic Indicator based on Webster's adaptive control method performs better.

Then compare the queue length using the same Webster's adaptive control method for different CAV penetration rates, as shown below (Fig. 6):

Different types of agents possess different variables and model the differences in information flow with changes in variables. The following table briefly describes some of the variables belonging to each agent and the meaning of these variables (Table 1).

Table 1. Agents and some of their variables

Agent type	Variable	Meaning
Cars	Xcor, ycor	The location of the vehicle
	v	Vehicle speed
	a	Vehicle acceleration
	Auto-type	Decide whether the vehicle is CAV or HDV
	Driver-type	Decide on the type of HDV driver
	Normal	Determine the condition of the vehicle's interior
Weather Center	Weather	Weather conditions collected every minute
Roadside facilities	Congestion	Congestion on roads near roadside facilities
Signal Lights	Sign	Determines the phase of the signal lights
Transportation Operations Center	Control	Decide on the control measures to be taken by the transportation operations center

4.3 Simulation Experiments

After defining the cyber layer and physical layer of the scenario separately, the corresponding simulation experiments can be taken to analyze the practicality and effectiveness of the model.

The simulation experiments focus on two parts: the examination of the scenario model on the integrity of information flow and physical objects described by the ATS architecture, and the comparative analysis of various traffic flow metrics, including average speed and queue length for various CAV penetrations under different phase control methods. The simulation time step is taken as 0.1 s, and the simulation length is 3600 time steps.

response" according to the different ways of processing and utilizing information. These four steps realize the initial processing of data through built-in data calculation formulas; the generation and comparison of advantages and disadvantages of decision solutions through built-in decision algorithms; and the effective use of data and effective feedback to the physical world through built-in control solutions. In the actual simulation process, the pynetlogo module built in Python is used to co-simulate with the Netlogo used for physical layer simulation, and the communication, computation and control functions of data are simulated by different operations respectively (Fig. 4).

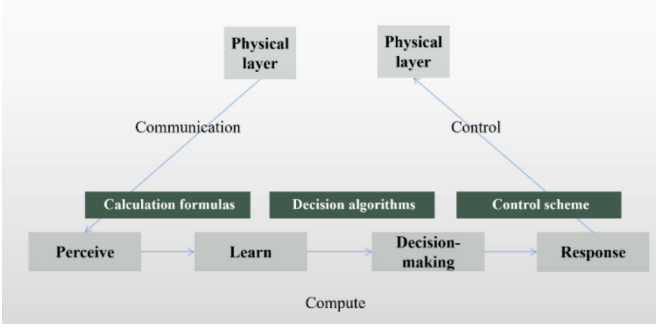


Fig. 4. Schematic diagram of the cyber layer structure

4.2 Scenario Overview

In the actual simulation experiment, the intersection scenario applied to the common four inlet lane intersection, where each inlet lane contains three lanes of left turn, straight ahead, and right turn, and the scenario contains five major categories of traffic participants: vehicles, signal lights, roadside infrastructures, weather center, and traffic operation center.

Vehicles are divided into two types of Connected-Automated Vehicle (CAV) and human driving vehicles (HDV), which take three different categories of follow-the-road models, IDM, CAC and CACC, according to their own vehicle types, where HDVs are divided into aggressive, normal and conservative types according to the characteristics of the drivers, and take different reflection times during the driving process according to the driver types. Vehicles will also monitor their own conditions in real time and take countermeasures in case of abnormal conditions, with vehicle arrivals conforming to Poisson distribution and a flow rate of about 360 vehicles per hour; Signal control methods include four-phase fixed-cycle, adaptive phase control based on the traffic volume of the previous cycle, and adaptive phase control based on Webster's timing method; The Roadside infrastructures will sense road congestion and road conditions in real time, and use them to impose speed limit notices and other measures on vehicles; the weather center will collect weather information at regular intervals and distribute it to vehicles and drivers; the traffic operation center can take common intersection control measures such as speed limit and no left turn according to the actual situation.

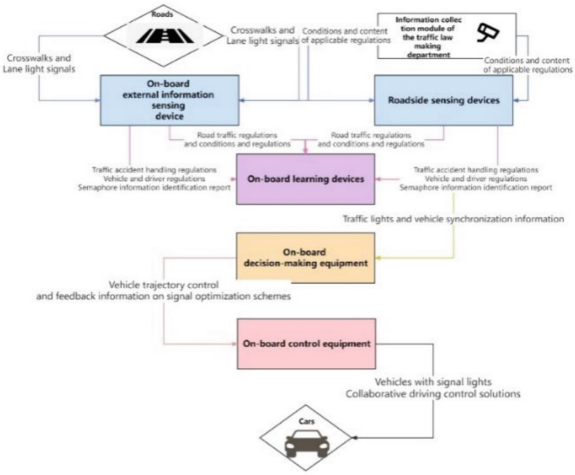


Fig. 3. Physical architecture of vehicle and signal light coordination sub-service in the intersection scenario

4 Modeling Based on ATS Scenario Theory

After the introduction of the theoretical framework of ATS, it is necessary to consider how to carry out scenario-oriented TCPS construction under the guidance of relevant theories, and to simulate and test the model after its successful construction in order to judge whether the model has good completeness and practicality.

4.1 Model Design

During the construction of the model, the structural relationships within the physical layer and the cyber layer need to be designed separately. First is the physical layer, which helps to examine the completeness and richness of the scene model with the help of the scenario architecture theory of ATS. The scenario architecture contains several kinds of physical objects that interact and influence each other, and the multi-agent simulation software Netlogo is used for the construction of the physical layer for this feature. In the mapping process of the simulation software to the scenario architecture, the basic elements in these three types of scenario architectures are mapped into the simulation software as the physical objects, the variables and global variables of the agents as the information flow, and the interaction between the agents and the data transmission process as the information interaction process, respectively, to realize the design of the physical layer.

The cyber layer is mainly concerned with the whole process of information from generation to being perceived, processed and utilized, so as to realize the effective application of real-time data in the cyber physical system, and therefore the cyber layer is designed as shown in the figure below. Information is generated in the physical world and reaches the cyber layer through V2X and other communication technologies. The cyber layer is divided into four steps: “perception → learning → decision making →

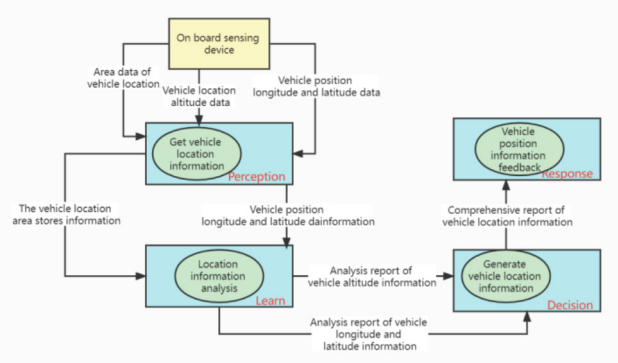


Fig. 2. Logical architecture of vehicle location aware service

they are clustered into information streams, and the information streams are connected in the physical objects with the help of the basic properties of the physical objects, and finally the physical architecture is generated as shown in the figure below.

3.3 Architectures of the Scenarios

ATS has built and analyzed the physical architecture for five typical scenarios, specifically divided into MaaS scenario, Electric Bus Operation scenario, Cargo Multimodal transportation scenario, Highway Formation scenario, and Intersection self-driving vehicle pedestrian avoidance scenario. Similar to the physical architecture of sub-services, the scenario architecture is also composed of three types of elements: physical objects, information flow and information interaction pairs.

In terms of concrete implementation, the demands and components contained within the scenario are analyzed by collecting a large number of definitions related to the scenario, and this is used as a guide to gradually obtain the services and functions that the scenario needs to provide, as well as the technologies needed to achieve them; subsequently, the corresponding three types of architectures correspond to the required sub-services, and on the basis of these materials, the physical architecture of the scenario is constructed with the help of a collaborative mechanism of elements and architectures [14, 15]. As an example, the intersection self-driving vehicle pedestrian avoidance scenario contains 27 sub-services and 117 sub-functions (Fig. 3).

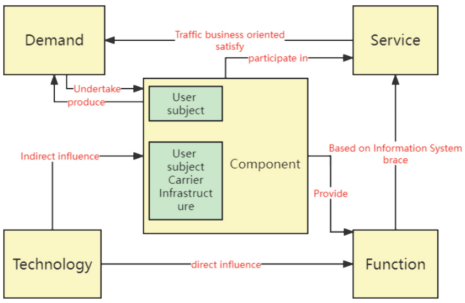


Fig. 1. Contact diagram of five types of ATS elements

3.2 Three Types of Architectures

To better explain the correlation of functions in each ATS service and to guide the construction of realistic infrastructure, it is necessary to form corresponding functional, logical and physical architectures for each service.

The ATS functional architecture serves as the initial architecture to describe the linkage between functions and to determine the logical sequence for implementing the functions in order to guide the subsequent architecture. Specifically, through the implementation logic of "sense, learn, decide, respond", a number of functions involved in each sub-service are classified, and the functions are divided into sub-functions according to the implementation process, and finally the sub-functions are connected in the order of their implementation in the service, and finally the functional architecture is formed.

ATS logical architecture is based on the understanding of the traffic semantics of the service, which is realized by organizing the information system functions, and plays the role of connecting "traffic" and "information functions", which is manifested in two aspects. On the one hand, it can express the hierarchical and progressive relationship between functions, that is, the further expression of "perception-learning-decision-response", and form a hierarchical system between functions, which provides a theoretical basis for architectural reconstruction, integration and optimization in any scenario. For this purpose, we layered perception into acquisition and identification, learning into fusion and analysis, decision making into generation of solutions and selection of optimal solutions, and corresponding into execution and feedback. On the other hand, the logical architecture must also clarify the input-output relationship of each function, and thus express the process of service implementation (Fig. 2).

The physical architecture is the carrier for the transformation of the logical architecture to the real transportation entity, the framework view that guides the planning and construction of the transportation system, and the ultimate embodiment of the basic theory of ATS. Firstly, the system functions in ATS need to be mapped to real traffic entities, and for this purpose, according to the current traffic system construction, the theoretical model of "physical object" is proposed, and its properties are described in terms of ontological attributes and connectivity attributes, the former including object categories, autonomy, and traffic information participation methods, and the latter including access capability and mapping logic [14]. Subsequently, by analyzing the types of data streams,

the transportation industry, which refers to the construction idea of CPS to establish the mechanism of sensing, communication, computation, and application of traffic information, so as to achieve the improvement of traffic efficiency and the efficient control of vehicles and infrastructure [9]. The current research on TCPS is divided into several aspects, some scholars focus on the computation and communication time problems of TCPS, and compare the advantages of TCPS through the time and efficiency of information transmission [10]; some scholars focus on the data security and related problems of TCPS, and propose various methods to guarantee the data security of the system [11]; some scholars focus on the modeling problems of TCPS, abstractly model the traffic events existing in the real world, and analyze the advantages and problems of TCPS through simulation.

The modeling of TCPS includes traffic modeling at the physical layer and network modeling at the cyber layer [12]. Due to the lack of real-world actual data of TCPS, the modeling of the physical layer of TCPS often relies on common microscopic traffic simulation software such as Sumo and Vissim. However, TCPS, as a complex fusion system, is concerned with the impact of a wide variety of information flows on the actual traffic, but most current studies rely on microscopic traffic simulation software, which can only simulate common traffic participants such as vehicles and signals, and cannot achieve effective simulation of the physical layer. Research on modeling the cyber layer for TCPS can be divided into two categories, one that models the cyber layer by considering vehicles as communication nodes and analyzing the communication metrics of multi-node networks, and the other that focuses on different decision and control algorithms to build the cyber layer of CPTS by modeling out the generation, processing, and analysis of data.

In general, there are still few studies on modeling TCPS as a trend of traffic system development, and there are drawbacks such as low completeness and lack of theoretical guidance, thus this paper decides to model and analyze the intersection scenario of TCPS with the help of the complete theoretical framework of ATS.

3 ATS Theoretical Framework

3.1 Five Types of Elements

ATS has been designed using object-oriented design methodology with 5 categories of basic elements of ATS. The capabilities possessed by the transportation system are decoupled into a number of mutually independent basic units, which are services, the transportation tests and requests made to the transportation system are summarized as requirements, the factors that facilitate the capabilities and evolution of the transportation system are summarized as technologies, the units that realize the capabilities of the transportation system are summarized as functions, and the participating roles of the transportation system are summarized as components, which are related as shown in the following figure [13] (Fig. 1).

for the common intersection scenarios in transportation system, the physical world and cyber space under this scenario are designed respectively, and the intersection scenario model based on ATS theory is constructed through the definition of information transmission and application methods, and simulation analysis is conducted for the completeness of the scenario and each traffic index of the scenario to form a complete TCPS scenario modeling technology system, which is conducive to guiding the development of future traffic information physical systems.

2 Related Works

2.1 CPS and Modeling

CPS enables the perception and control of the physical world through the integration of the physical world and the in cyber space, using advanced perception, communication, and computing technologies, with strong real-time capabilities. In recent years, countries around the world have been investing a lot of efforts in CPS research, and in 2022, the Chinese Natural Science Foundation listed CPS as one of 115 "priority areas for development".

The key step of CPS from abstract architecture to concrete model is the modeling of CPS, and the problem is the focus and difficulty of CPS research and has attracted the attention of a large number of scholars worldwide. The modeling process is limited by the characteristics of discrete CPS cyber layer, continuous physical layer, containing more elements, and the integration of multiple industrial fields, which cannot use the traditional modeling method and needs to take into account multiple aspects. The current solutions to the problem fall into several categories: First, the traditional discrete system modeling methods are borrowed to build discrete systems, mainly including formal modeling and high-level language modeling. Among them, formal modeling includes formal inference modeling, extended Petri net modeling, time automaton modeling and other methods, while high-level language modeling includes AADL, modelica, UML, etc., and reaches the unification of the overall system through the discretization of continuous events; Second, drawing on the traditional continuous system, model the continuous system in CPS with the help of traditional continuous system modeling methods, such as parametric model, Newtonian mechanics, etc., and then reasonably embed the discrete events into the continuous system to achieve the unification of the two; The third approach is hybrid modeling, where CPS modeling of hybrid tools is achieved by modeling the physical and cyber layers separately and disposing the interfaces between them rationally [7]. In addition, some scholars try to adopt emerging technologies such as group intelligence [8] and data-driven to solve the problem, but they are limited by the maturity of technology development and need further research and improvement.

2.2 TCPS and Modeling

In recent years, with the increasing traffic demand and the deep application of traffic data, a physical system of traffic information combining 3C (communication, computation, and control) technology and traffic elements has become a hot research topic in

technology infrastructures, and the improvement of service quality due to technological developments, but also in stimulating the emergence of more demanding transportation demands [1].

Specifically, the advent of next-generation data communication technologies makes larger data transfers and very small communication delays possible, which allows transportation systems to obtain the state of traffic participants (people, vehicles, roads and environment) in near real-time or quasi-real-time conditions, greatly enhancing the sensing capabilities of transportation systems. Meanwhile, with the widely application of AI, ML, and IC technologies, the computing capacity of end-users and edge devices is fully utilized, making it possible for end-users to obtain their own status with high precision in most cases, which not only reduces the amount of information transmission, but also supports more personalized and intelligent system services. In addition, the development of distributed system management technology, optimization theory and other technologies make the collaboration process among the agents more robust and efficient, and also play a strong guarantee in terms of system scalability and security. The development of these three types of technologies has led to the efficient implementation of the three main functions of sensory and communication, computing and processing, decision and control of data in transportation system, respectively, which are the main components of concern for transportation cyber physical system (TCPS) [2]. The transportation cyber physical system regards the real traffic world as the physical entity layer, and transmits all kinds of traffic data generated by the physical entity layer, such as traffic flow, vehicle speed, parking delay to the cyber space in real time, and generates some kinds of control schemes in real time through many kinds of data processing and decision algorithms, and synchronizes them to the physical world to implement control, so as to realize the effective utilization of traffic information.

At the same time, emerging technologies are transforming transportation systems from "passive" to "active" in meeting the demands of transporting people and goods, which has given birth to the concept of autonomous transportation systems (ATS). ATS realizes transportation by self-organized operation and autonomous service. Its operation logic is autonomous perception, autonomous learning, autonomous decision making and autonomous response, which is essentially to reduce human intervention in transportation system and enhance the autonomous capability of transportation system, specifically in the four aspects of active response to traffic demand, automatic operation of vehicles, active control of infrastructure and active adaptation of external environment. With the help of system engineering modeling theory, a set of theoretical framework of transportation system from transportation elements to specific guiding structures has been constructed with "autonomy" as the core in the study of ATS [3–5].

Among the studies targeting TCPS, simulation techniques try to reproduce the implementation logic of TCPS and are one of the main techniques related to TCPS. However, due to the synergistic requirements of TCPS for multiple domains and the lack of effective theoretical guidance, the current research on TCPS simulation technology is more focused on the study of modeling methods and mainly stays at the level of individual events [6].

To this end, this paper first analyzes the development status of TCPS, and then introduces the theoretical framework of ATS from elements to architecture to scenarios. Next,



Cyber Physical System Modeling and Analysis in Typical Scenarios Based on the Theory of Autonomous Transportation System

Zi-Sheng Zhou[1,2], Ming Cai[1,2], Zhuo-Lin Deng[1,2], and Chen Xiong[1,2(✉)]

[1] School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510006, China
xiongch8@mail.sysu.edu.cn

[2] Guangdong Key Laboratory of Intelligent Transportation Systems (ITS), Guangzhou 510006, China

Abstract. With the rapid development and application of sensing, computing and controlling technologies in the transportation industry, the construction of transportation cyber physical system (TCPS) relying on these three types of technologies has gradually become a research hotspot in the transportation system. However, the modeling of TCPS suffers from a lack of theoretical guidance and a single modeling hierarchy. To this end, this paper introduces the theoretical framework system of Autonomous Transport System (ATS) as the theoretical basis for TCPS scenario modeling, and sorts out the theoretical framework of ATS from five types of elements to three types of architectures to typical scenarios. Then, taking the modeling of TCPS in intersection scenario as an example, the physical layer of the model is constructed by mapping the physical objects, information flow, and information interaction pairs in the ATS scenario architecture, and the cyber layer of the model is designed by the service implementation logic of ATS, and through the process of data generation and application to achieve the integration of the two layers of applications. After the model was constructed and simulations were implemented, the functional integrity of the scenario reflected by the model was analyzed qualitatively, and the results of specific traffic indicators under different parameters were analyzed quantitatively to verify the integrity and validity of the model.

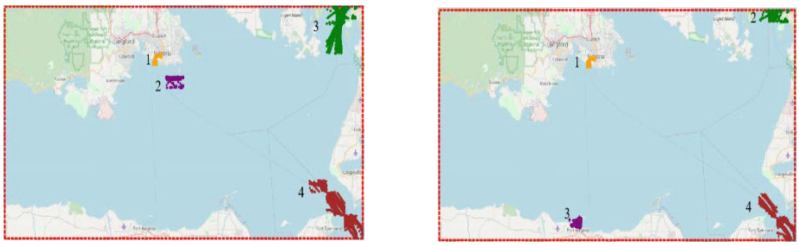
Keywords: Cyber-physical system · traffic system modeling · autonomous transportation system

1 Introduction

With the rapid development of information and intelligent technology, the new generation of communication technology, Internet of Things, big data, AI, mobile Internet, energy management and Intelligent & connected vehicle technology are gradually used in the field of intelligent transportation system. This is manifested not only in the rapid growth of passenger volumes, the orderly construction and replacement of information

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 165–177, 2024.
https://doi.org/10.1007/978-3-031-65126-7_17

2. Ning, Z., et al.: Blockchain-enabled intelligent transportation systems: a distributed crowd-sensing framework. IEEE T Mob. Comput. **21**(12), 4201–4217 (2021)
3. Ning, Z., et al.: Intelligent resource allocation in mobile blockchain for privacy and security transactions: a deep reinforcement learning based approach. Sci. China Inf. Sci. **64**(6), 162303 (2021)
4. Ning, Z., et al.: Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE T Mob. Comput. **22**(5) (2021)
5. Ning, Z., Chen, H., Ngai, E.C., Wang, X., Guo, L., Liu, J.: Lightweight imitation learning for real-time cooperative service migration. IEEE T Mob. Comput. (2023)
6. Ferreira, M.D., Campbell, J., Purney, E., Soares, A., Matwin, S.: Assessing compression algorithms to improve the efficiency of clustering analysis on AIS vessel trajectories. Int. J. Geogr. Inf. Sci. **37**(3), 660–683 (2023)
7. Egala, B.S., Pradhan, A.K., Badarla, V., Mohanty, S.P.: Fortified-chain: a blockchain-based framework for security and privacy-assured internet of medical things with effective access control. IEEE Internet Things **8**(14), 11717–11731 (2021)
8. Xiong, L., Xiong, X., Zhang, F., Chen, H.: Unsupervised Deep Embedding Clustering for AIS Trajectory. In: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 2283–2286 (2022)
9. Deebak, B.D., Al-Turjman, F.: Smart mutual authentication protocol for cloud based medical healthcare systems using internet of medical things. IEEE J. Sel. Area Commun. **39**(2), 346–360 (2020)
10. Ali, A., et al.: Security, privacy, and reliability in digital healthcare systems using blockchain. Electronics **10**(16), 2034 (2021)
11. Bureva, V., Popov, S., Traneva, V., Tranev, S.: Generalized net model of cluster analysis using CLIQUE: clustering in quest. Int. J. Bioautom. **23**(2), 131 (2019)
12. Wang, S., Li, Y., Xing, H.: A novel method for ship trajectory prediction in complex scenarios based on spatio-temporal features extraction of AIS data. Ocean Eng. **281**, 114846 (2023)
13. Ganesh, E.N., Rajendran, V., Ravikumar, D., Kumar, P.S., Revathy, G., Harivardhan, P.: Detection and route estimation of ship vessels using linear filtering and ARMA model from AIS data. Int. J. Oceans Oceanogr. **15**(1), 1–10 (2021)
14. Hartawan, I.P.N., Widyantara, I.M.O., Karyawati, A.A.I.N.E., Er, N.I., Artana, K.B., Sastra, N.P.: AIS data pre-processing for trajectory clustering data preparation. In: 2021 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), Bali, Indonesia, 2021, pp. 1–5. (2021)



(a) Areas with frequent ship movements on January 1

(b) Areas with frequent ship movements on January 2

Fig. 4. Visualization of the map of the frequent movement area of ships with a single spatiotemporal granularity

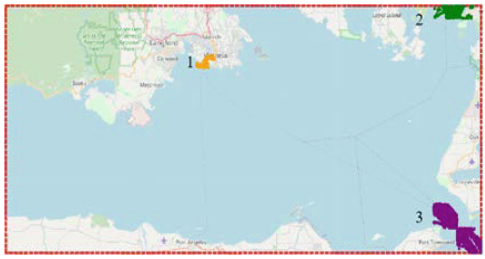


Fig. 5. Visualization of the map of the frequent movement area of ships with a multi-temporal and spatial granularity based on two-day data from January 1st to 2nd

spatial-temporal granularity shows great differences on different dates, which proves the effectiveness of the method proposed in this paper.

4 Conclusion

In this paper, we propose a method for extracting frequent ship moving areas based on grid density peak clustering, which solves the problem that grid density peak clustering methods need to manually select cluster centers. To learn more fine-grained spatial-temporal information, we consider frequently active regions of both spatial and temporal information. We fuse the frequently active regions with single spatiotemporal granularity on the timeline to obtain frequent active regions with multiple spatiotemporal granularities, which makes the extracted frequent active regions more accurate. In simulation experiments, we evaluate the effectiveness of the proposed ship frequent activity area extraction method and compare it experimentally with other methods. The results show that our method can more accurately and effectively extract the areas with frequent ship activities.

References

1. Wang, X., Ning, Z., Guo, L., Guo, S., Gao, X., Wang, G.: Mean-field learning for edge computing in mobile blockchain networks. IEEE T Mob. Comput. **22**(10), 5978–5994 (2022)

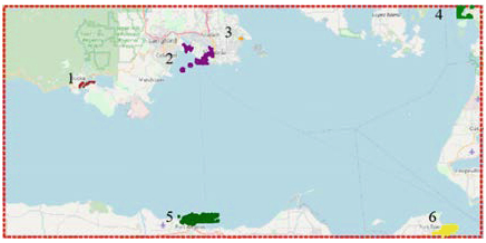


Fig. 2. Map visualization of frequent activity areas by using CLIQUE method

area 2, but it is identified as a frequent activity area. However, some areas among the screened-out areas have very high ship densities, such as grid 18, grid 38, grid 399, grid 379 and grid 358, which represent areas with significantly higher ship density than other areas., but was identified as an infrequently active area.} Fig. 3 shows the frequently active regions extracted by grid density peak clustering. Compared with Fig. 2, the density of frequent active areas extracted in Fig. 3 is in the forefront, which shows the effectiveness of the advanced selection threshold method proposed in this paper, which can correctly screen out high-density grids.



Fig. 3. Map visualization of frequent activity areas by using AGDPC method

Figure 2 and Fig. 3 only consider the spatial information of the frequent activity area extraction method, and can only obtain the frequent activity area in the entire large time period, but cannot obtain the frequent activity area in different time periods. In order to extract frequent activity areas more accurately, this paper firstly extracts the frequent activity areas of ships with single spatio-temporal granularity in different time periods in the same area under the given time granularity and spatial granularity. On this basis, on the time axis, if there is an intersection between frequent ship activity areas in adjacent time periods, the frequent activity areas in adjacent time periods are merged. Otherwise, the fusion of the next time period is performed until the time span is traversed.

The frequently active regions extracted at a single spatio-temporal granularity using the grid density peak clustering method are shown in Fig. 4. And the frequent movement area of ships with multiple spatial and temporal granularities is shown in Fig. 5. It can be seen that the frequent ship activity areas under multiple spatial-temporal granularity tend to be consistent on different dates, and the propagation activity area of a single

3.1 Experimental Parameter Settings

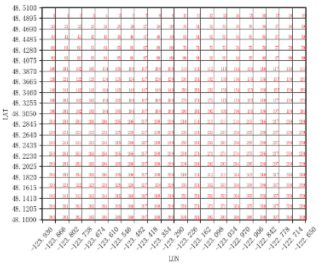
The parameter setting of the comparison method in the experiment is selected through manual tuning, as shown in Table 1.

Table 1. .

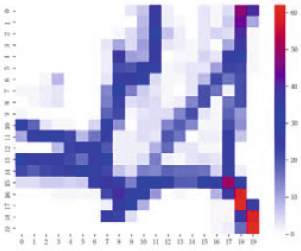
Parameter	Value
Meshing	20*20
Density threshold	200
Time granularity	1(day)

3.2 Results and Analysis

We first divide the experimental data into 20*20 grid areas. The number of ships in each grid area and the heat map are shown in Fig. 1. Subsequent experiments will be compared with Fig. 1.



(a)Meshing and grid number



(b) heat map of ship distribution in grid

Fig. 1. The attributes of the research area (a) Meshing and grid number, (b) heat map of ship distribution

The extraction of frequent ship activity areas at a single spatio-temporal granularity refers to extracting frequent ship activity areas in different time periods within the same research area, given the time granularity and spatial granularity. We divide the time range into several uniform equal parts, and divide the space range into m*m grids. An improved grid density peak clustering algorithm is used to automatically select the cluster center and extract its frequent activity areas at a single spatio-temporal granularity. Figure 2 shows the frequent activity areas of ships extracted by CLIQUE. It extracts 6 frequently active regions. However, compared with Fig. 1, it can be seen that the ship density in some of the six frequent activity areas is very low, such as grid 107 in area 1, area 3, and

identified. First, count the number of times that each mesh object in the cluster is the nearest higher-density mesh object to other mesh objects:

$$nt_j = \sum_{i=1}^{n} z\left(j - \underset{j:\rho_j > \rho_i}{arg\,min}\left(d_{ij}\right)\right) \tag{1}$$

In the formula, $z(x) = \begin{cases} 1, x = 0 \\ 0, other \end{cases}$. d_{ij} is expressed as the Euclidean distance between the grid object i and the grid object j. ρ is the local density of the mesh object. Since it is difficult for a point located on the boundary of a cluster to become the closest high-density mesh object to other mesh objects, when is 0, the mesh object is usually located on the boundary area, so the core area of the cluster can be defined as:

$$c_{core}^{k} = \left\{x_i | \rho_i > \max\left(\rho_j\right), x_i \in c^k, x_j \in c^k \& nt_j = 0\right\} \tag{2}$$

In the formula, c^k represents the clusters obtained by clustering, c_{core}^{k} represents the core region of the class cluster c^k, $\max(\rho_j)$ is the maximum function. In these cluster core areas, although their density is larger than their neighbors, from the overall data distribution, some areas have relatively few ships and should not be considered frequent activity areas. In order to obtain the frequent activity areas of ships that meet the actual situation, these core areas need to be further screened. In order to reduce human participation, this paper automatically selects the density threshold $d_{th} = \max_{ntj=0}(\rho_j)$ according to the distribution characteristics of grid density in various clusters, and selects the grids whose grid density exceeds the threshold in various clusters:

$$area_{fre} = \left\{x_i | \rho_i > d_{th}, x_i \in c_{core}^{k}\right\} \tag{3}$$

By merging adjacent high-density mesh objects, the ship frequent activity area can be obtained. However, the ship frequent activity area extracted by this method ignores the time information. In fact, the areas of ship activities are different at different times. In this paper, by fusing ship frequent areas with single spatio-temporal granularity on the time axis, more accurate ship frequent areas with multiple spatio-temporal granularities are obtained.

3 Experiment and Analysis

In order to validate the proposed method, this paper uses two common frequent activity region detection methods for comparison. The first is the classic grid clustering method Clustering In QUEst(CLIQUE) [9], which uses the number of data points in the grid as the grid density to extract areas with frequent ship activities; the second is the advanced grid density peak clustering method proposed in this paper. This experiment selected AIS data of ships in the sea area of 122°35′W−123°55′W, 48°06′W−48°30′N from January 1, 2019 to January 3, 2019, and the data comes from the open source website https://marinecadastre.gov/ais/.

such as Maritime Mobile Service Identity (MMSI), vessel position, and speed. By analyzing and mining AIS data, extracting frequent activity areas of vessels can provide technical support for research in detecting abnormal vessel behavior, predicting port traffic flow, voyage planning [11], and recognizing maritime target intentions.

At present, clustering algorithms have been widely used in the research of object hotspots region extraction [12]. Wang et al. [13] developed a rapid clustering model of trajectories based on hierarchical modeling. Each ship state establishes its trajectory similarity model and performs recursive clustering of ship trajectories from top to bottom, avoiding the cumbersome calculation of existing ship clustering models, high time complexity, difficult parameter adjustment process, and other shortcomings. In [14], a spectral clustering algorithm is used to cluster the sub-trajectory segments to identify representative ship maneuvering behavior trajectories. Hartawan et al. [11] suggested that a typical motion model of ships in the area could be obtained based on AIS data by DBSCAN clustering of ship trajectory segments combined with track similarity measure and extraction of typical trajectories.

The above methods only focus on the spatial information of moving objects and ignore the time information, which will result in the identification of an area with no ships or only a few ships in a certain interval as an area where ships are frequently active. In this regard, this paper proposes an advanced grid density peak clustering method (AGDPC). This method can extract frequent ship motion areas at multiple time granularities while using spatial clustering and considering ship time information.

2 Advanced Grid Density Peak Clustering

Traditional grid density peak clustering requires manual determination of cluster centers, which can easily lead to inaccurate clustering results. To address this problem, this paper uses the boxplot method and the elbow method to automatically determine the cluster center and number of clusters. First, in order to solve the problem that the local density and relative distance of grid objects affect the selection of cluster centers due to different dimensions, this paper first uses the minimum and maximum normalization method to map the value range of grid objects to the local density and relative distance of grid objects. Perform normalization processing between 0 and 1, and preselect the cluster center set. Then, use the box plot method to calculate its upper and lower bounds and quartiles to obtain its box plot distribution. Grid objects with higher local density and relatively far distance is further filtered according to the box plot distribution as a cluster set. At this time, the cluster center candidate set may contain more cluster centers than the actual situation, causing the classification of clusters to be too detailed. Because there may be grid objects in the cluster set that have a high local density but a small relative distance, or a grid object that has a small local density but a large relative distance, it is necessary to further screen the cluster center candidate set. This paper uses the elbow method to filter the cluster set. By finding the inflection point of the cluster center, the candidate points before the inflection point are used as the cluster center, and the remaining candidate points are assigned to the same cluster as its nearest high-density neighboring grid object., complete the cluster analysis of ship AIS data and obtain the ship activity area.} Based on the above ideas, the core regions of the clusters can be



Extraction of Frequently Active Areas of Ships Based on Advanced Grid Density Peak Clustering

Xuanrui Xiong[1], Han Shen[1], Lanke Zhu[2,3], and Jianbo Zheng[2,3](✉)

[1] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
xiongxr@cqupt.edu.cn, s210131206@stu.cqupt.edu.cn

[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{1120200314,jianbo.zheng}@smbu.edu.cn

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

Abstract. The cognition of the frequent activity areas of ships based on AIS data is of great significance in reducing port navigation risks and improving the efficiency of ships entering and leaving ports. Traditional extraction methods only consider spatial information and ignore the impact of temporal information on clustering results, resulting in inaccurate extraction of frequently active areas. We propose an advanced grid density peak clustering method (AGDPC) to extract frequently active areas, which can advanced select cluster centers and density thresholds to solve the problem that grid density peak clustering methods cannot advanced select cluster centers. The improved grid density peak clustering method is used to extract frequent ship motion regions under a single spatial-temporal granularity according to a given spatial-temporal granularity. Then, we fuse multiple ship frequent activity areas to obtain multi-temporal and spatial granularity ship frequent activity areas. Experimental results show that this method can extract frequent motion are-as more accurately than traditional methods, and better reflect the ship's navigation rules.

Keywords: Trajectory clustering · Grid density peak clustering · Frequent activity areas extraction · AIS data

1 Introduction

The mining and analysis of existing data is one of the important means to predict and evaluate the future situation of objects. With the development of machine learning and deep learning, data mining and analysis techniques are widely used in economics, edge computing [1–3], blockchain [4–7] and other fields [8–10]. The Automatic Identification System (AIS) is a type of ship navigational system that contains essential information

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 158–164, 2024.
https://doi.org/10.1007/978-3-031-65126-7_16

7. Faragallah, O.S., et al.: Efficiently encrypting color images with few details based on RC6 and different operation modes for cybersecurity applications. IEEE Access **8**, 103200–103218 (2020)
8. Farooq, M.U., Waseem, M., Mazhar, S., Khairi, A., Kamal, T.: A review on internet of things (IoT). Int. J. Comput. Appl. **113**(1), 1–7 (2015)
9. Ghadirli, H.M., Nodehi, A., Enayatifar, R.: An overview of encryption algorithms in color images. Signal Process. **164**, 163–185 (2019)
10. Gokhale, P., Bhat, O., Bhat, S.: Introduction to IoT. Int. Adv. Res. J. Sci. Eng. Technol. **5**(1), 41–44 (2018)
11. Grichi, M., Abidi, M., Jaafar, F., Eghan, E.E., Adams, B.: On the impact of interlanguage dependencies in multilanguage systems empirical case study on java native interface applications (JNI). IEEE Trans. Reliab. **70**(1), 428–440 (2020)
12. Hwang, S., Lee, S., Kim, J., Ryu, S.: Justgen: effective test generation for unspecified JNI behaviors on JVMs. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pp. 1708–1718. IEEE (2021)
13. Jihui, Y., Qinian, Z., Zhenhao, Z.: Cloud storage and security technology based on the the internet of things. ZTE Technol. J. **18**(6), 12–16 (2012)
14. Lee, S., Lee, H., Ryu, S.: Broadening horizons of multilingual static analysis: semantic summary extraction from c code for JNI program analysis. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, pp. 127–137 (2020)
15. Li, B., Feng, Y., Xiong, Z., Yang, W., Liu, G.: Research on AI security enhanced encryption algorithm of autonomous IoT systems. Inf. Sci. **575**, 379–398 (2021)
16. Liu, F., Koenig, H.: A survey of video encryption algorithms. Comput. Secur. **29**(1), 3–15 (2010)
17. Lu, H., Jin, C., Helu, X., Zhu, C., Guizani, N., Tian, Z.: AutoD: intelligent blockchain application unpacking based on JNI layer deception call. IEEE Netw. **35**(2), 215–221 (2020)
18. Manikandan, V., Amirtharajan, R.: On dual encryption with RC6 and combined logistic tent map for grayscale and DICOM. Multimed. Tools Appl. **80**(15), 23511–23540 (2021)
19. Ming, H., Jun, J., Xiaohu, C., Guohua, C.: Technology and security of internet of things. Comput. Secur. **4**, 49–52 (2011)
20. Raghavan, B., Casado, M., Koponen, T., Ratnasamy, S., Ghodsi, A., Shenker, S.: Software-defined internet architecture: decoupling architecture from infrastructure. In: Proceedings of the 11th ACM Workshop on Hot Topics in Networks, pp. 43–48 (2012)
21. Ren, W., et al.: Privacy-preserving using homomorphic encryption in mobile IoT systems. Comput. Commun. **165**, 105–111 (2021)
22. Sajjad, M., et al.: Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities. Futur. Gener. Comput. Syst. **108**, 995–1007 (2020)
23. Wang, Y., Liang, X., Hei, X., Ji, W., Zhu, L.: Deep learning data privacy protection based on homomorphic encryption in AIoT. Mob. Inf. Syst. **2021**, 1–11 (2021)
24. Xiaoqiang, Z., Mengmeng, W., Guiliang, Z.: Research on the new development of image encryption algorithms. Comput. Eng. Sci. **34**(5), 1–6 (2012)
25. Xu, J., Zhao, B., Wu, Z.: Research on color image encryption algorithm based on bit-plane and chen chaotic system. Entropy **24**(2), 186 (2022)
26. Zhu, H., Peng, Y., Xu, H., Tong, F., Jiang, X.Q., Mirza, M.M.: Secrecy enhancement for SSK-based communications in wireless sensing systems. IEEE Sens. J. **22**(18), 18192–18201 (2022)



Fig. 8. Data of the encrypted AI model file, garbled and unreadable (partial display)

5 Conclusions

In this paper, a model privacy protection solution is proposed to help protect the privacy of model owners and data owners in complex IoT application environments. The improved RC6-plus, JNI middleware approach can more effectively protect model privacy in the IoT environment. Experimental results demonstrate the reasonableness and effectiveness of the approach. Considering AI model applications are increasingly used in new smart IoT software, hardware application products, and other services, and the arithmetic performance of cloud-based servers far exceeds the hardware configuration on the IoT side, this research can be further extended to model encryption and decryption of cloud-based server applications and data privacy protection during remote data transmission of IoT.

References

1. Ahmed, H.E.D.H., Kalash, H.M., Allah, O.F.: Encryption efficiency analysis and security evaluation of RC6 block cipher for digital images. In: 2007 International Conference on Electrical Engineering, pp. 1–7. IEEE (2007)
2. Alenezi, M.N., Alabdulrazzaq, H., Mohammad, N.Q.: Symmetric encryption algorithms: review and evaluation study. Int. J. Commun. Netw. Inf. Secur. **12**(2), 256–272 (2020)
3. Ashraf, A.M., Elmedany, W.M., Sharif, M.S.: Secure IoT data transmission at physical layer using rc6 encryption technique. In: 2022 9th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 307–315. IEEE (2022)
4. Baños-Gonzalez, V., Afaqui, M.S., Lopez-Aguilera, E., Garcia-Villegas, E.: IEEE 802.11 ah: a technology to face the IoT challenge. Sensors **16**(11), 1960 (2016)
5. Dong, B., Shi, Q., Yang, Y., Wen, F., Zhang, Z., Lee, C.: Technology evolution from self-powered sensors to AIoT enabled smart homes. Nano Energy **79**, 105414 (2021)
6. Fang, K., Yang, B., Zhu, H., Lin, Z., Wang, Z.: Two-way reliable forwarding strategy of RIS symbiotic communications for vehicular named data networks. IEEE Internet Things J. (2022)

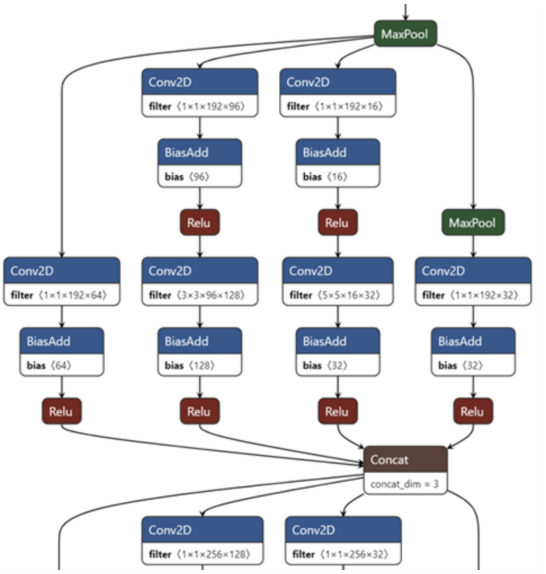


Fig. 7. Structure diagram of AI model before encryption (partial display)

RC6-plus algorithm will use the same algorithmic process but execute the algorithmic steps opposite to the encryption direction. The wheel keys are applied opposite to recover the original message. Therefore, the decryption process of the RC6-plus algorithm takes the same amount of time as the encryption process.

It is important to note that the key length and algorithm parameters used in the RC6-plus encryption and decryption process must be the same to perform the decryption correctly. If a different key length or parameters are used in the decryption phase than in the encryption phase, the decryption process may fail or get an incorrect message.

In addition, this paper also designs the AI-IoT software decoupling architecture so that the API interface layer is decoupled from the JNI layer. The user can only see the outer interface call function but does not know the principle of the internal algorithm [7,9,16], as shown in Fig. 6. The forward inference neural network of the AI model is rewritten using the miniCaffe C++ language, and the source code is obfuscated so that even standard model visualization cracking tools cannot read or write to the model. With the above source code encryption process, the project source code can be compiled to generate .so files, making the system more private, as shown in Fig. 7 and Fig. 8.

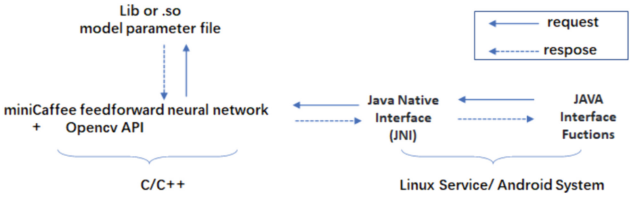


Fig. 6. AI-IOT software decoupling architecture

When the Key length equals 64 bits, the Digital dataset RC6 average encryption time equals 127 ms, while the Digital dataset - RC6-plus average encryption time only takes 89 ms; when the Key length increases to 208 bits, the Digital dataset RC6 average encryption time equals 407 ms, while the Digital dataset - RC6-plus average encryption time only takes 323 ms. When the Key length increases to 208 bits, the Digital dataset RC6 average encryption time equals 407 ms, while the Digital dataset - RC6-plus average encryption time is only 323 ms.

When the key length is 64 bits, the average encryption time of an image data set using the RC6 algorithm is 1896 milliseconds, while the average encryption time using the RC6-plus algorithm is only 1186 milliseconds, and the encryption time of the RC6-plus algorithm is significantly less than that of the RC6 algorithm; when the key length is increased to 208 bits, the advantage of RC6-plus algorithm is more prominent, and its average encryption time is 5137 ms, which takes 299 ms less than the RC6 algorithm. For audio data sets, the average encryption time of the RC6-plus algorithm is only 191 milliseconds when the key length is 64 bits. For digital data sets, the average encryption time of the RC6-plus algorithm is only 464 milliseconds when the key length is 208 bits.

From the results of the analysis of the above experimental data, we draw several conclusions:

1. The encryption time increases as the value of the Key length increases;
2. The encryption time will vary depending on the complexity of the data;
3. The RC6-plus algorithm can be used on text datasets, digital datasets, image datasets, or audio datasets. The encryption performance of the algorithms on the audio data set is significantly better than RC6.
4. These data results show that the RC6-plus algorithm has a shorter encryption time than the RC6 algorithm for different data sets and critical lengths. The advantage is undeniable for longer key lengths.

The experimental data of encryption strength and encryption time of the RC6-plus algorithm do not include the experimental data of decryption of the RC6-plus algorithm. In practical applications, the decryption time and strength are usually related to factors such as the length of the key used for encryption, the data set, and the algorithm version. Usually, the decryption process of RC6-plus is similar to the encryption process, but the order of the keys is reversed. Therefore, the same key and algorithm parameters should be used in the decryption phase as in the encryption phase. In the decryption process, the

Table 2 shows the experimental data of the performance tests on RC6 and RC6-plus encryption algorithms for text and numeric datasets, respectively, and Fig. 4 shows the visual line graph corresponding to Table 2. The data are recorded in the table.

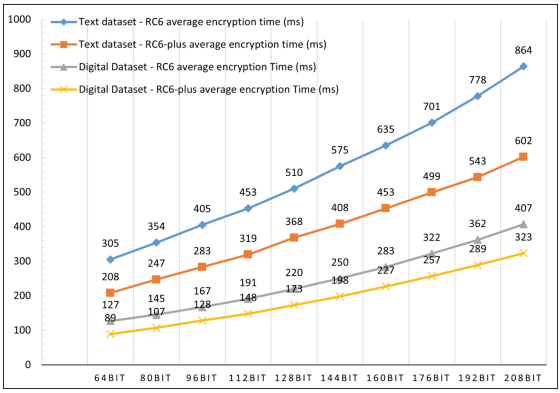


Fig. 4. Visual line chart of performance comparison of encryption algorithms for text and digital datasets

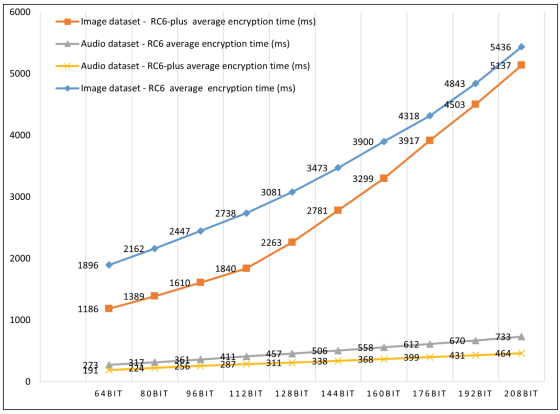


Fig. 5. Performance comparison of image and audio dataset encryption algorithms

When the Key length is equal to 64bit, the Text dataset RC6-plus average encryption time is equal to 305 ms, while the Text dataset RC6-plus average encryption time is only 208 ms, RC6-plus encryption time is less than RC6; When the Key length increases to 208 bits, the advantage of RC6-plus is more apparent, and the average encryption time of RC6-plus is 602 ms, which is 262 ms less than that of RC6.

Table 2. Performance comparison of encryption algorithms for text and digital datasets

Key Length	Average encryption time (ms) of Dataset - Algorithm			
	Text - RC6	Text - RC6-plus	Digital - RC6	Digital - RC6-plus
64bit	305	208	127	89
80bit	354	247	145	107
96bit	405	283	167	128
112bit	453	319	191	148
128bit	510	368	220	173
144bit	575	408	250	198
160bit	635	453	283	227
176bit	701	499	322	257
192bit	778	543	362	289
208bit	864	602	407	323

from web crawlers. In this experiment, 1000 copies of each data type were sampled randomly from these original data sets as the original data for the experiments in this paper. The average encryption time of each data type is taken after the experiment. The test data will be listed in a table, where each row represents a test key length, and each column represents the encryption time and encryption strength of each test item in the dataset.

In this paper, we experimentally compare the encryption time and strength of the traditional RC6 algorithm and RC6-plus algorithm on text, digital, image, and audio datasets, as shown in Table 2, Fig. 4, Table 3, and Fig. 5.

Table 3. Performance comparison of encryption algorithms for image and audio datasets

Key Length	Average encryption time (ms) of Dataset - Algorithm			
	Image - RC6	Image - RC6-plus	Audio - RC6	Audio - RC6-plus
64bit	1896	1186	273	191
80bit	2162	1389	317	224
96bit	2447	1610	361	256
112bit	2738	1840	411	287
128bit	3081	2263	457	311
144bit	3473	2781	506	338
160bit	3900	3299	558	368
176bit	4318	3917	612	399
192bit	4843	4503	670	431
208bit	5436	5137	733	464

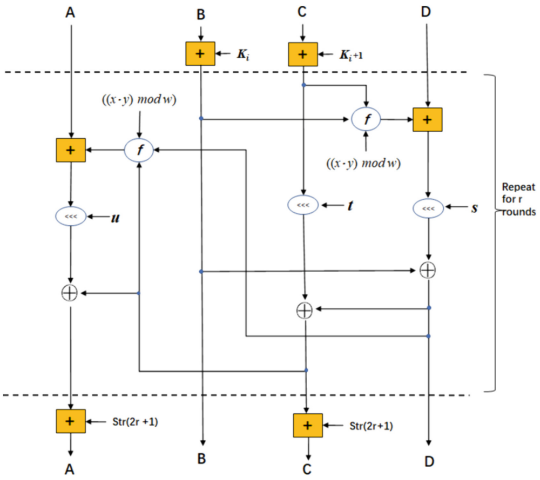


Fig. 3. Principle of wheel function of RC6-plus algorithm

Table 1. Raspberry Pi 4 Model B configuration parameters

Configuration items	Specification
CPU	Broadcom BCM2711, quad-core Cortex-A72(ARM v8), 64-bit SoC at 1.5 GHz
Memory	LPDDR4 SDRAM, 4 GB, MicroSD card slot
Network	Gigabit Ethernet, dual-band 802.11ac wireless card (with optional Bluetooth 5.0)
USB	2 * USB 3.0 and 2 * USB 2.0 ports
Video/Audio Output	2 * micro-HDMI ports supporting up to 4K resolution
Audio	Stereo output, stereo MIC, support Bluetooth audio output
GPIO	40 * GPIO pins
Operating System	Raspberry Pi OS (Debian-based operating system)

4 Experimental Results and Analysis

In this work, we tested the performance of the proposed scheme on a real Raspberry Pi-based IoT device. To evaluate the performance of the proposed scheme, we exclude the time consumed on the communication channel as it heavily depends on the network traffic. The experimental setup focuses only on the performance tests for the time used for encryption, RC6-plus operation, and decryption. The RC6-plus instances are all running on the latest Raspberry Pi (Raspberry Pi 4 Model B) with the system parameters configured, as shown in Table 1.

The following experiments are based on improved RC6 cryptographic algorithms with different bit cell lengths (64bit 208bit), using six datasets provided, each testing ten RC6 algorithms with different key lengths. The datasets are derived from the homebrew program Automatic Random Sequence, COCO dataset, ImageNet, CIFAR, MNIST, PASCAL VOC, SQuAD, Labeled Faces in the Wild, UCI Machine Learning Library, and with the addition of some data

and 4 round constants K_i in each round, and 4 outputs A', B', C', D', then the computation process of the round function can be expressed as

$$B \leftarrow B + K_i \tag{15}$$

Pass the new value B calculated in Eq. 1 into the f function:

$$D \leftarrow D + f(B, C) \tag{16}$$

$$D \leftarrow D <<< s \tag{17}$$

$$D \leftarrow D \oplus B \tag{18}$$

The new value D is calculated and rounded with C:

$$C \leftarrow C + K_i + 1 \tag{19}$$

$$C \leftarrow C <<< t \tag{20}$$

$$C \leftarrow C \oplus D \tag{21}$$

Immediately afterward, the new values C and D are rounded with A:

$$A \leftarrow A + f(C, D) \tag{22}$$

$$A \leftarrow A <<< u \tag{23}$$

$$A \leftarrow A \oplus C \tag{24}$$

After following the above cryptographic round, a new set of output values is obtained:

$$A' \leftarrow A, B' \leftarrow B, C' \leftarrow C, D' \leftarrow D \tag{25}$$

where $i = 1, 2, ..., r, s, t, u$ are the computed parameters, and f(x, y) denotes the improved RC6-plus algorithm. A nonlinear function is introduced in the program for converting inputs to outputs. Specifically, the function $f(x, y)$ can be defined as

$$f(x, y) = (x \oplus y) <<< ((x \cdot y) \mod w) \tag{26}$$

This function is a nonlinear function that converts the inputs B and C into an output word D. The class Similarly, another nonlinear function $f'(x, y)$ can be defined as

$$f'(x, y) = (x \oplus y) <<< (w - ((x \cdot y) \mod w) \tag{27}$$

This function converts inputs C and D into an output word A to further disrupt data flow during encryption. The above is the basic structure schematic of the wheel function in the RC6-plus algorithm, in which more complex nonlinear functions are introduced to enhance the security of the encryption algorithm. In practical applications, the wheel function can be adjusted and optimized according to specific needs to improve the strength and performance of the encryption algorithm. RC6-plus is similar to encryption and decryption, which is not repeated here in this paper, and the wheel function of the RC6-plus algorithm can be viewed as shown in Fig. 3.

The detailed process of RC6-plus algorithm improvement has the following steps:

1. Adjust the number of rounds r and w-bit word length to balance security and algorithm performance
   First, we assume the original RC6 algorithm uses r-rounds and w-bit word lengths. For each word A, B, C, D, we can calculate its output using the following equation:

$$A' = ((A \oplus B) <<< s) \oplus K_0 \tag{9}$$

$$B' = ((B \oplus C) <<< t) \oplus K_1 \tag{10}$$

$$C' = ((C \oplus D) <<< u) \oplus K_2 \tag{11}$$

$$D' = ((D \oplus A) <<< v) \oplus K_3 \tag{12}$$

   where $\oplus$ denotes the XOR operation, $<<<$ denotes a circular shift, K_i is a round constant, and s, t, u, and v are parameters that need to be calculated based on the w-bit word length. The formulae for these parameters are as follows:

$$w = 32 \rightarrow r = 20, s = 7, t = 2, u = 13, v = 8 \tag{13}$$

$$w = 64 \rightarrow r = 20, s = 35, t = 5, u = 31, v = 16 \tag{14}$$

   Suppose we want to improve the algorithm to increase its performance. In that case, we can reduce the number of rounds r or decrease the w-bit word length to reduce the amount of computation for encryption. However, this will also reduce the security of the algorithm.
2. Modify the generation method of the RC6-plus algorithm wheel constant K_i
   Increase the randomness and complexity of wheel constant generation to enhance the strength of the encryption algorithm. The wheel constant Ki of the RC6 algorithm is derived from a specific key. If this key can be guessed or leaked, then the security of the encryption is threatened. Therefore, we need to enhance the randomness and complexity of the wheel constant generation to improve the strength of the algorithm. We use more complex key derivation algorithms or introduce more wheel constants to increase the randomness and complexity of encryption.
3. Optimization of operations in the encryption wheel
   The original RC6 algorithm uses relatively simple arithmetic, so we must introduce more complex nonlinear functions and improve the algorithm using iso-or and circular shifts. Simple attack methods can break this simple arithmetic, so we can optimize the arithmetic in the encryption wheel by introducing more complex nonlinear functions to increase the strength of the algorithm.
   With the above three improvements, we can improve the security and performance of the RC6 algorithm to make it more suitable for practical applications, and the improved RC6 algorithm is noted as RC6-plus. Suppose the RC6-plus algorithm uses r rounds of encryption, with 4 inputs A, B, C, D,

2. In the algorithm of RC6, Str(i) represents the subkey word, while "<<<" and ">>>" represent the controlled left rotation and right rotation, respectively. The symbol controls the amount of rotation, followed by the number of rotations is controlled by the lowest 5 bits of the number following the symbol. In addition, to meet the fixed grouping bit length requirement, RC6 also uses a quadratic function B*(2B+1) to strengthen the diffusion property, which is very different from most other encryption and decryption algorithms. The graphical representation of the wheel function is shown below (in Fig. 2), where $f(x)$ represents the following nonlinear invertible function:

$$f(x) = (x * (2x + 1)) <<< lg(w) \tag{8}$$

3. RC6 is a high-performance, highly flexible group iterative cipher whose compact and transparent architecture makes it widely used in monolithic microcontrollers. In addition, RC6 performs even better in application scenarios such as fingerprint recognition and POS machines. The data-dependent cyclic nature of RC6 can significantly improve encryption efficiency while its memory requirements are relatively low, and the highly integrated internal cache technology can significantly reduce production costs.
Although the RC6 algorithm is designed for simplicity and efficiency, it still has some shortcomings, such as the lack of performance of the nonlinear function f because the bit diffusion of f is unidirectional. The diffusion speed is slow, and the average computation is w/2 = 16 additions due to the use of multiplication in f, so the nonlinear function becomes the bottleneck of the operation speed. In addition, RC6 has significant differences in the encryption and decryption algorithms, which is also a drawback. To address these issues, some improvements were made to RC6 as follows.

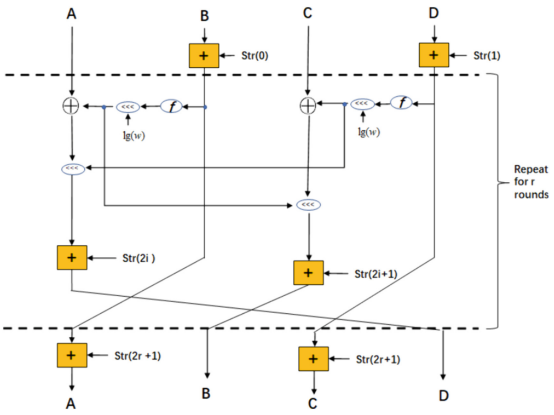


Fig. 2. Principle of RC6 algorithm wheel function

transmitted to RC6-plus for encryption and decryption via JNI. If the secret key is decrypted successfully, the gate control valve (GS, Gate Switch) is opened, allowing the data received by the I/O to reach the feedforward neural network layer through the JNI layer to complete the inference, which is often accompanied by the process of model loading. The final result is again transmitted to the terminal display device through the I/O interface.

3.2 Key Algorithm Study

In this section, we analyze the traditional RC6 algorithm in detail and find that RC6 has certain defects in the application of this project [7,9,22]. Based on this, we propose the RC6-plus encryption algorithm with flexible control of the number of round bits and successfully apply it to our intelligent IoT terminal AI project, achieving good results.

The detailed process of the RC6 algorithm has the following steps:

1. RC6 is one of the AES candidate algorithms. It is an improved version of the RC5 algorithm. (w, r, b) in RC6-$w/r/b$ denotes the operation word length, the number of iteration rounds, and the user master key length, respectively. Usually, we choose the arithmetic word length w = 32 bits (bit). The plaintext packet length is 4 characters (128 bits). RC6 consists of input encryption, r rounds of iteration, and output transformation.
   Input encryption:

$$(A, B, C, D) = (A, B + Str(0), C, D + Str(1)) \tag{1}$$

   Round r iteration:

$$t = [B * (2B + 1)] <<< lg(w); \tag{2}$$

$$u = [D * (2D + 1)] <<< lg(w); \tag{3}$$

$$A = [A \oplus t <<< u] + Str(2i); \tag{4}$$

$$C = [C \oplus u <<< t] + Str(2i + 1); \tag{5}$$

$$(A, B, C, D) = (B, C, D, A); \tag{6}$$

   Output transformation:

$$(A, B, C, D) = (A + Str(2r + 2), B, C + Str(2r + 3), D); \tag{7}$$

2. A new improved RC6 (Rivest cipher 6) algorithm is proposed for enhanced model encryption, denoted as RC6-plus;
3. An improved design scheme between AI-IOT software layers is proposed to further enhance security.

3 Proposed Method

This section focuses on two parts; the first part is an explanation of the model training steps and the architecture diagram of the AI-IOT model inference system; the second part introduces the traditional RC6 while proposing the RC6-plus algorithm and its improvement process; the reader will learn about the method of AI model security protection research for IoT terminals described in this paper.

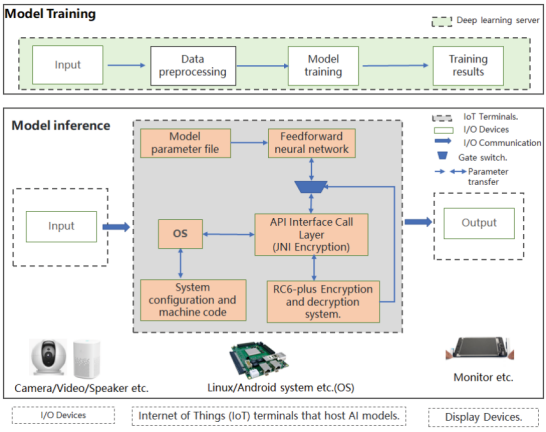


Fig. 1. Server model training and AI-IOT model inference system architecture

3.1 System Architecture Design

As the architecture is shown in Fig. 1, the model training task is done on the deep learning GPU server, the model inference is made on the smart IoT terminal, and the model can run on Linux or Android operating systems. The basic steps of model training (feedback neural network) include data input, data preprocessing, feedback neural network calculation, and parameter storage. Among them, model inference (feedforward neural network) mainly includes input, model parameter loading, and model computation inference. The data on the input side can come from the camera, microphone, or locally stored data, and the output receiver can be the display or stored in the database. This paper focuses on the software system architecture approach to the smart IoT terminal. The operating system (OS, Operate System) extract is coded by the connected terminal machine and

3. Encryption algorithm research
   Encryption algorithms protect secure communication between IoT endpoints and AI models. Researchers have recently researched encryption algorithms to ensure data security during the communication process. For example, researchers have proposed trusted authentication techniques, secure transmission, and data encryption algorithms [2,3,7,15,18,19,24,25].
4. Multi-level security mechanism design
   In order to enhance the security between IoT terminals and AI models, the researchers proposed a multi-level security mechanism design. This approach will strengthen security in several aspects, such as model management, model usage, and model storage, to maximize the security of the IoT terminal system [2,7,9,11,12,14,16–18].

The above are only some of the research contents related to the research of AI model security protection methods for IoT terminals. However, the research in this area is still in its initial stage, and further in-depth research and exploration are needed in many aspects. Therefore, it is of significant theoretical and application value to research the AI model security protection method for IoT terminals [7,11–13,19,24].

Based on this, this research will draw on domestic and international research results and methods to explore a more comprehensive, detailed, and feasible AI model security protection method from various aspects, such as data security, model protection, and encryption algorithms. The research aims to solve the challenges faced by AI model security on IoT endpoints, protect the security and privacy of devices and data, and provide useful academic and practical references for the innovative development of related fields [2,3,7,18,25].

This paper first introduces the basic concepts of IoT and AI models and analyzes the existing AI model security threats and the limitations of existing solutions. Secondly, the paper also proposes a security protection method based on encryption and access control and details the implementation process and related technical details of the method.

Finally, the paper verifies the effectiveness and feasibility of the proposed method through experiments. The experimental results show that the method has high security and scalability and can provide more comprehensive protection for AI models in IoT applications.

Overall, the application of AI-IoT offline identification devices can effectively improve the efficiency of data sampling, reduce the cost of manual work, enhance accuracy, and is suitable for data collection and processing tasks in various complex environments. The development of this technology plays an essential role in developing the modern manufacturing industry, promoting industrial upgrading, and enhancing national strength. The research results of this paper are of great significance to the development and popularization of IoT applications and provide a helpful reference for AI safety.

The main contributions of this work can be summarized as:

1. An efficient system architecture of AI model is proposed for IoT terminals;

Internet of Things (AI-IOT) offline identification device can realize real-time processing and identification of data by deploying the model on the body of the device, and the data processing method that does not depend on the transmission environment reduces the risk of data transmission interruption and solves the problem of its online real-time identification instability, which has high practical value. In archaeology, exploration, and geological examination, AI-IOT can process and comprehensively analyze a large amount of data. Based on the characteristics of offline processing technology, this AI-IOT can also work autonomously through unitized design and intelligent scheduling technologies to further improve work efficiency. Conventional monitoring systems may have signal interruptions and analysis errors in bad weather, such as rain, wind, and lightning. At the same time, AI-IOT can deploy AI models on the equipment to realize real-time monitoring and grasp the comprehensive situation of the machine, environment, and other parameters, and conduct intelligent analysis and processing of this, increasing the application areas of artificial intelligence [8,21]. At the same time, the application of AI-IOT is becoming increasingly widespread and will face many challenges, including physical security, identity identification issues, external attacks, vulnerability issues, data validation, model security, program update mechanism, and communication security. In particular, AI models involve a large amount of data and privacy; once attacked and maliciously changed, it is easy to affect the stability and reliability of the whole system, and people start to pay attention to the impact of read and write operations of smart IoT terminal device data on the privacy and security of AI models. Therefore, how to protect the AI models of IoT terminals has become a critical research direction [1,20,21,23].

This paper is organized as follows: Sect. 2 reviews the current research status of AI-IOT; Sect. 3 reviews an efficient AI model application system architecture for IoT terminals and crucial algorithm research (RC6, RC6-plus); Sect. 4 focuses on the research content experimental results and analysis process. Finally, conclusions are drawn in Sect. 5.

2 Related Work

At present, AI model security protection for IoT terminals in academia and industry is actively carrying out relevant research, mainly around the following aspects:

1. Data security protection
   For the security threats and risks faced by the data security of IoT terminals, researchers have proposed a series of data security protection techniques. For example, encryption technology is used for secure data storage, integrity protection, and access control mechanisms [6,8,10,21,23].
2. AI model protection
   In response to the security risks faced by AI models, researchers have developed a series of protection techniques for AI models. For example, AI models are encrypted, compressed, and cut to increase the security of the models; a reliable AI algorithm framework is used to ensure the reliability of model training and inference [1,13,20,23,26].



Artificial Intelligence Model Based Security Protection Method for IoT Applications

Xiaolong Luo[1], Xiaoli Chen[2(✉)], Jie Wei[1], Liang Zhang[2], Luping Xu[2], and Bijun Zhao[2]

[1] Zhejiang Water Conservancy Information Publicity Center, Hangzhou 310009, China
[2] Zhejiang Ponshine Information Technology Co., Ltd., Hangzhou 311100, China
chenxiaoli@ponshine.com

Abstract. The issue of model privacy security is increasingly affecting the application systems of Artificial Intelligence Internet of Things (AI-IOT) terminals, where it is challenging to protect the privacy of the underlying AI models. In this paper, we propose a security protection RC6-plus algorithm based on cryptography and access control for AI model security in IoT applications. Specifically, the proposed method effectively protects the privacy of crucial algorithms in the program by encrypted storing the model parameters, as well as storing and code obfuscating the neural network structure and parameters of the AI model independently while adding the isolation treatment of the JNI communication layer. The results of the experiments verify the effectiveness of the proposed method.

Keywords: Model privacy security · internet of things · artificial intelligence model · encryption technology · access control

1 Introduction

With the continuous development of society and technology, Internet of Things (IoT) technology has been widely used in various fields. The IoT is gradually becoming an indispensable part of people's life and work because of its intelligence and connectivity. In recent years, IoT and Artificial Intelligence (AI) technologies are becoming more and more closely integrated, and more and more AI technologies are being applied to IOT end devices. Such devices also have the ability of offline recognition, which can realize various applications in highland, deep sea, remote areas, archaeology, exploration, and geological examination [4,5,10].

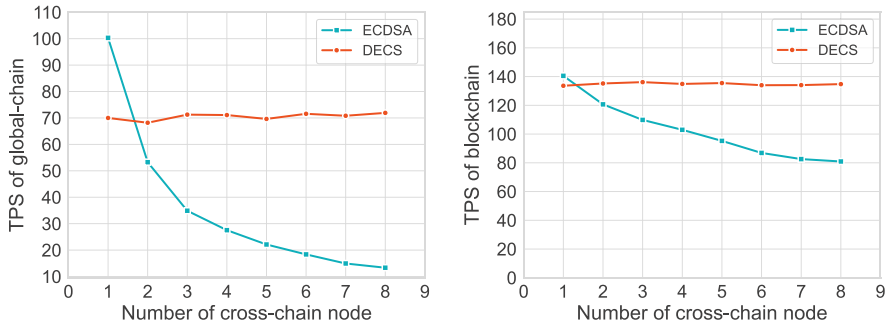
In the plateau, deep sea, and other particular environments, conventional networking equipment may have the problem of unstable network transmission, leading to interruption in the data transmission process. The Artificial Intelligence

X. Chen—This work was funded by the Zhejiang Provincial Department of Water Resources Science and Technology Plan Project, Zhejiang, China (Project no. RC2238).

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 143–157, 2024.
https://doi.org/10.1007/978-3-031-65126-7_15

16. Zhang, H., Lao, L., Shu, C., Xiao, B.: Analysis of the communication traffic model for permissioned blockchain based on proof-of-work. In: IEEE International Conference on Communications (ICC 2021). IEEE International Conference on Communications, IEEE (2021). https://doi.org/10.1109/ICC42927.2021.9500333, Telus; Huawei; Ciena; Nokia; Samsung; Qualcomm; Cisco; Google Cloud, iEEE International Conference on Communications (ICC), ELECTR NETWORK, JUN 14-23, 2021
17. Zhou, J., Feng, G., Wang, Y.: Optimal deployment mechanism of blockchain in resource-constrained IoT systems. IEEE Internet Things J. **9**(11), 8168–8177 (2022). https://doi.org/10.1109/JIOT.2021.3106355

5. Dai, H.N., Zheng, Z., Zhang, Y.: Blockchain for internet of things: a survey. IEEE Internet Things J. **6**(5, SI), 8076–8094 (2019). https://doi.org/10.1109/JIOT.2019.2920987
6. Deng, L., Chen, H., Zeng, J., Zhang, L.-J.: Research on cross-chain technology based on sidechain and hash-locking. In: Liu, S., Tekinerdogan, B., Aoyama, M., Zhang, L.-J. (eds.) EDGE 2018. LNCS, vol. 10973, pp. 144–151. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94340-4_12
7. Genc, Y., Afacan, E.: Design and implementation of an efficient elliptic curve digital signature algorithm (ecdsa). In: Chakrabarti, S., Paul, R., Gill, B., Gangopadhyay, M., Poddar, S. (eds.) 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1026–1031 (2021). https://doi.org/10.1109/IEMTRONICS52119.2021.9422589. IEEE; Inst Engn & Management; IEEE Vancouver Sect; IEEE Toronto Sect; SMART; Univ Engn & Management , iEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), ELECTR NETWORK, APR 21-24, 2021
8. Ghosh, B.C., Bhartia, T., Addya, S.K., Chakraborty, S.: Leveraging public-private blockchain interoperability for closed consortium interfacing. In: IEEE INFOCOM 2021 - IEEE Conference on Computer Communications, pp. 1–10 (2021). https://doi.org/10.1109/INFOCOM42981.2021.9488683
9. Hope-Bailie, A., Thomas, S.: Interledger: Creating a standard for payments. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 281–282. WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2016). https://doi.org/10.1145/2872518.2889307
10. Kang, E.S., Pee, S.J., Song, J.G., Jang, J.W.: A blockchain-based energy trading platform for smart homes in a microgrid. In: Proceedings of 2018 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 472–476 (2018). IEEE; Nagoya Inst Technol, 3rd International Conference on Computer and Communication Systems (ICCCS), Nagoya, JAPAN, APR 27-30, 2018
11. Lu, Q., Xu, X.: Adaptable blockchain-based systems a case study for product traceability. IEEE Softw. **34**(6), 21–27 (2017). https://doi.org/10.1109/MS.2017.4121227
12. Makhdoom, I., Abolhasan, M., Abbas, H., Ni, W.: Blockchain's adoption in IoT: the challenges, and a way forward. J. Netw. Comput. Appl. **125**, 251–279 (2019). https://doi.org/10.1016/j.jnca.2018.10.019
13. Sun, Y., Yi, L., Duan, L., Wang, W.: A decentralized cross-chain service protocol based on notary schemes and hash-locking. In: 2022 IEEE International Conference on Services Computing (SCC), pp. 152–157 (2022). https://doi.org/10.1109/SCC55611.2022.00033
14. Syta, E., et al.: Keeping authorities honest or bust with decentralized witness cosigning. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 526–545 (2016). https://doi.org/10.1109/SP.2016.38
15. Tschorsch, F., Scheuermann, B.: Bitcoin and beyond: a technical survey on decentralized digital currencies. IEEE Commun. Surv. Tutor. **18**(3), 2084–2123 (2016). https://doi.org/10.1109/COMST.2016.2535718



(a) TPS of global-chain vs. Number of cross-chain nodes.

(b) TPS of entire blockchain vs. Number of cross-chain nodes.

Fig. 5. Blockchain performance comparison.

6 Conclusion

In this paper, we present a novel decentralized and efficient cross-chain scheme tailored for IoT systems. Our constructed system demonstrates excellent adaptability to the diverse and intricate nature of IoT environments. Furthermore, our proposed cross-chain solution effectively addresses the pervasive issue of excessive centralization within current cross-chain technologies. The incorporation of BLS signature-based data verification alleviates the burden of high storage requirements associated with the ECDSA scheme. Experimental results show that our scheme has better decentralization and efficiency, and the blockchain throughput has a good performance. We hope our scheme can be used as a high-performance tool for IoT systems to provide efficient, secure protection for IoT data.

References

1. Ahluwalia, S., Mahto, V, R., Guerrero, M.: Blockchain technology and startup financing: a transaction cost economics perspective. Technol. Forecast. Soc. Change **151** (2020). https://doi.org/10.1016/j.techfore.2019.119854
2. Back, A., et al.: Enabling blockchain innovations with pegged sidechains. **72**, 201–224 (2014)
3. Boudguiga, A., et al.: Towards better availability and accountability for IoT updates by means of a blockchain. In: 2017 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 50–58 (2017). https://doi.org/10.1109/EuroSPW.2017.50
4. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. IEEE Access **4**, 2292–2303 (2016). https://doi.org/10.1109/ACCESS.2016.2566339

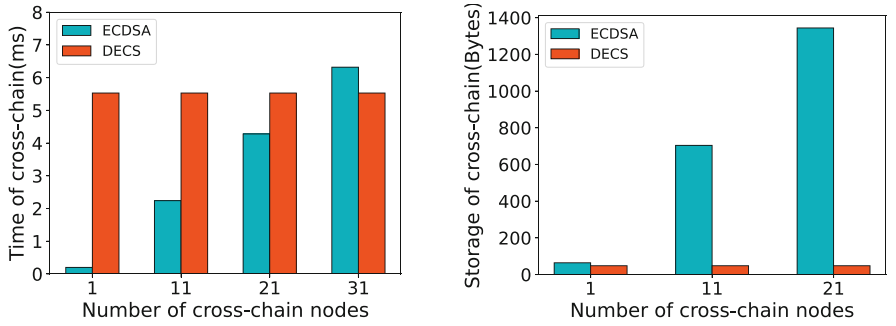
Table 1. Time cost for each step of the BLS signature.

Notation	Description	execution time
T_z	Take a random number in T_z	0.0183 ms
T_{G_1}	Point multiplication in G_1	0.5482 ms
T_{G_2}	Point multiplication in G_2	0.6671 ms
T_a	Point addition in G_1	0.3030 ms
T_p	Polynomial calculation time	0.0053 ms
T_h	Hashing with SHA256	0.0027 ms
T_e	Verification of Bilinear Functions in T_z	3.9822 ms

Figure 4a shows the time cost required by the two schemes. The experiment starts with the most basic notary scheme, which means that only 1 node is involved in the cross-chain. We can see that at this point the ECDSA scheme significantly outperforms the DECS scheme. However, as the number of cross-chain nodes increases, the time spent by the ECDSA scheme keeps increasing. This is because the ECDSA scheme is unable to aggregate signatures. The verifier needs to verify all the signatures one after another. On the other hand, the DECS scheme can keep the verification time at 5.53 milliseconds due to the aggregation of signatures.

Figure 4b illustrates the storage cost required by both schemes. Similar to the time cost. The ECDSA scheme has increasing storage space as the number of cross-chain nodes. In contrast, the DECS scheme only has one MPK for each cross-chain process, regardless of the number of cross-chain nodes. So the storage cost remains constant. Since this scheme uploads the public key into the global-chain, the storage size will significantly affect the overall performance of the blockchain. This gives the DECS scheme a more obvious advantage.

Blockchain Performance: Figure 5a and Fig. 5b illustrates the TPS of the blockchain under both scheme. Where Fig. 5a shows the TPS of the global-chain. As the number of cross-chain nodes increases. The data size that needs to be stored on the global-chain increases for the ECDSA scheme. The throughput of the blockchain also keeps decreasing. The same effect is seen in Fig. 5b. Figure 5b shows the test performed for the entire blockchain network, which includes two local-chains and one global-chain. As we can be seen from the figure, the performance of the ECDSA scheme keeps decreasing. While the performance of the DECS scheme remains stable in both experiments. When the number of cross-chain nodes reaches 8, the DECS scheme significantly outperforms the ECDSA scheme.



(a) Time cost of cross-chain vs. number of closs-chain nodes

(b) Storage cost of cross-chain vs. number of closs-chain nodes

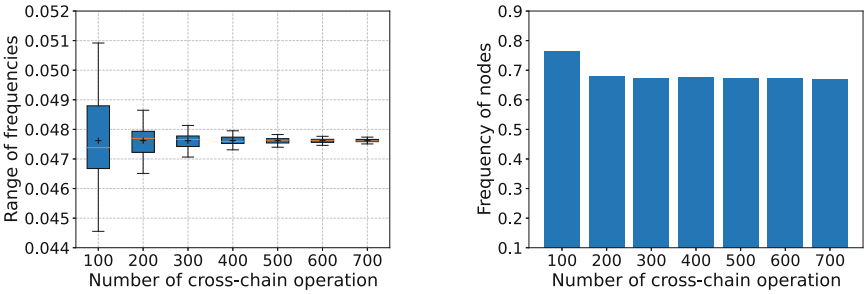
Fig. 4. Time and Storage cost of cross-chain.

the local-chains has 9 nodes and the other has 12 nodes. The two local-chains are independent of each other. The blockchain network adopts the Raft consensus, with a maximum blocking time of 7 s, a maximum number of 100 transactions, and a maximum block size of 100M. At the same time, the experiment used the mainstream testing tool Caliper of the Hyperledger Fabric to test blockchain performance.

Frequency of Nodes Participating in Cross-chain: The first experiment targets the Reputation-Based Node Selection Mechanism to test whether the selection of cross-chain nodes is decentralized. We initiate a cross-chain request to a local-chain of 12 nodes. And set the number of cross-chain nodes to 8, and the reputation parameter $\alpha = 0.5$, to test the frequency of each node participating in the cross-chain.

The experimental results are shown in Figs. 3a and Figs. 3b. Figures 3a is a box plot of the frequency of all nodes participating in cross-chaining. It can be seen that there is still an obvious gap in the frequency of the nodes when completing 100 cross-chaining. And as the number of cross-chaining increases. The gap in frequency gradually decreases. Each node can be selected evenly. Figures 3b count the 8 nodes with the lowest network latency. And calculate the sum of their frequencies. It can be seen that at 100 experiments, the frequency is the highest at 0.76. With the increase in the number of experiments, the value gradually decreases and reaches a stable value of 0.67. This number is the ratio of the number of nodes to the number of all nodes. It shows that our scheme is able to decentralize the selection of nodes participating in cross-chain.

Time and Storage Cost of Cross-chain: We compare the DECS proposed in this paper with the ECDSA-based notary scheme [9]. Compare the time and storage cost used to validate cross-chain data for both. First, We performed a step-by-step test of BLS signatures. The experimental results are shown in the Table 1. Each step was experimented with 100 times and averaged.



(a) Range of all node frequencies vs. Number of cross-chain

(b) Frequency of the 8 nodes with the lowest latency vs. Number of cross-chain

Fig. 3. Frequency of nodes participating in cross-chain.

Reputation Update Process: We use smart contracts to automatically update the reputation scores of nodes. The core idea of this contract is to consider the frequency of participation in cross-chaining and network latency of each node, achieving a balance between frequency and network latency. After the node participant in the cross-chain, its called coefficient c_i is updated. The smaller the c_i is, the more likely it is to be selected as a participant for cross-chain. The specific description is as follows:

1. After the blockchain network is initialized, the smart contract awards cross-chain participation status to the nodes with high trustworthiness. And initialize their called coefficients $c_i = c_(init)$.
2. Cross chain initiator s_r is for each strong node v_i Calculate reputation p_i. And sort the obtained values in ascending order, determine the priority of the signature, select nodes in order to form a set N, send (t, n) threshold signature requests, and send unselected messages to nodes outside the set N.
3. After collecting c_i, t_i from these nodes, the smart contract computes the reputation p_i for each node v_i. And it sorts the obtained values in ascending order. And select the nodes in order to combine them into a set N. A (t, n) threshold signature request is sent to the nodes in the set, and an unselected message is sent to nodes outside the set N.
4. Node $v_k(k \in N)$ that receives threshold signature request performs the reputation update algorithm of the Eq. (11). And the unselected nodes $v_l(l \in Vandl \notin N)$ perform the reputation update algorithm of the Eq. (12).

5 Performance Analysis

We tested our scheme on the HyperLedger Fabric, using Intel (R) Xeon (R) Silver 4214 CPU with 256 GB RAM with 16TB hard drive. We built a blockchain network with 21 nodes. It includes two local-chains and one global-chain. One of

If the final calibration determines that $e(\sigma, G) = e(h(M), MPK)$ is equal, then the returned output is true, and the opposite is false.

The above six steps describe the cross-chain process of the BLS-based threshold signature scheme. The scheme utilizes a polynomial to hide the master private key in the BLS signature method and generates the private keys of the participants from it. Then a specific number of individual participant signatures on the data are integrated using an elliptic curve bilinear mapping function. The master signature is recovered using Lagrange interpolation. Validation is then performed to determine the validity of the transaction.

4.2 Reputation-Based Node Selection Mechanism in Cross-Chain

Parameter Settings: Before describing this mechanism, We first define the following.

Definition 1: Assuming that there are m nodes within a localized chain. One of the nodes is denoted as v_i. where $1 \leq i \leq m$. Each node maintains a called coefficient c_i. And the initial value is c_Init, which satisfies the following equation:

$$c_{init} = \frac{1}{m} \tag{8}$$

Definition 2: A node that is not part of the current local-chain is selected as a cross-chain requester. It records the network latency t for each node in $V = \{v_1, v_2, ..., v_m\}$. The time factor r_i is then computed for each v_i. r_i satisfies the following equation:

$$r_i = \frac{t_i}{\sum_{j \in V} t_j} \tag{9}$$

Definition 3: Knowing the called coefficient c_i and the time coefficient r_i of a node v_i, the cross-chain initiator determines the weight parameter $\alpha(\alpha \in [0, 1])$, for which it can compute the prestige value p_i, which satisfies the following equation:

$$p_i = \alpha c_i + (1 - \alpha) r_i \tag{10}$$

Definition 4: After each cross-chain completion, each selected node v_i needs to be updated with the following equation:

$$c_i = c_i + \frac{1}{mn} (i \in N) \tag{11}$$

And the unselected node v_i also needs to update with the following equation:

$$c_i = c_i - \frac{1}{m(m-n)} (i \in V and i \notin N) \tag{12}$$

After the computation is completed, KGC sends the threshold encryption private key to the selected nodes through a secure channel.

Signature: $Sign(SK, m, t, h) \rightarrow \sigma_i$

Each participant i receives the private key and signs the data $m = \{0,1\}^*$. First, participant i hashes the data and maps the resulting hash digest to G_1: $h(M) \in G_1$, and $M = m||d$, with d equal to 0, 1, 2, ... This is because if the value obtained by hashing m is mapped directly onto the curve, there is a 50% probability that it will not map to a particular point. Therefore, the value of d is increased until the point is successfully mapped. The signature σ_i of participant i will be as shown in Eq. (5:)

$$\sigma_i \leftarrow x_i \times h(M) \in G_1 \tag{5}$$

Aggregation: $Aggregate(\sigma) \rightarrow \sigma$

The node A_j with the highest reputation score is responsible for aggregating signatures. It collects signatures from n participants. When it receives a signature group σ generated from the set Q combined by t participants. and the individual signatures within the signature group are verified. A_j can obtain the signature of the master private key MSK on the data based on Lagrange interpolation.

$$\begin{aligned}
\sigma &= \sum_{i \in Q} \sigma_i \prod_{j \in Q, j \neq i} \frac{j}{j-i} \\
&= \sum_{i \in Q} (h(M)) \times x_i \prod_{j \in Q, j \neq i} \frac{j}{j-i} \\
&= h(M) \times \sum_{i \in Q} x_i \prod_{j \in Q, j \neq i} \frac{j}{j-i} \\
&= h(M) \times x
\end{aligned} \tag{6}$$

Response: After signature aggregation, A_j packages the cross-chain data and MSK in Cross-chain.response. Then A_j sends the response data to B_i via A_{Cross}.

Verification: $Verify(MPK, \sigma, h(M), G, e) \rightarrow true|false$

After the B_i obtains the complete signature obtained by the endorsement initiator, the signature can be verified based on the properties of the bilinear function. The specific principle of verification is shown in the following equation:

$$\begin{aligned}
e(\sigma, G) &= e(x \times h(M), G) \\
&= e(h(M), x \times G) \\
&= e(h(M), MPK)
\end{aligned} \tag{7}$$

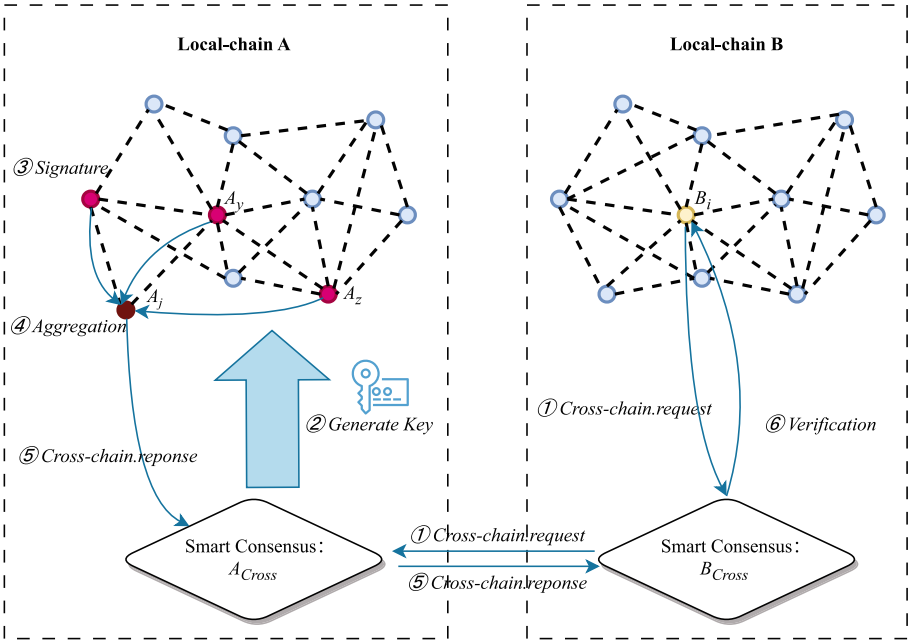


Fig. 2. Cross-chain process of DECS.

SK, and the threshold public key set PK. After the generation is complete, the generation center discloses the master public key MPK and the user group's public key PK. The key generation center randomly selects a random integer as shown in the Eq. (2)

$$x \leftarrow Z_p^* \tag{2}$$

At the same time, KGC set it as the master private key $MSK = x$, and obtains the master public key according to Eq. (3), where G is a randomly selected point from the elliptic curve:

$$v \leftarrow x \times G \in G_2 \tag{3}$$

After this, v will be set as the master public key $MPK = v$. Subsequently, the key generation center randomly selects $t-1$ elements $a_1, a_2, a_3, \ldots, a_{t-1} (a_i \in Z_p^*)$, and set $a_0 = x$, thus constructing a polynomial of order $t-1$:

$$f(x) = \sum_{i=0}^{t-1} a_i x^i \tag{4}$$

Then, KGC calculates x for each participant i $x_i = f(i)$ and set the corresponding threshold private key $sk_i = x_i$. Calculate v_i also according to Eq. (3). And set $pk_i = v_i$. It also calculates for n participants to obtain $SK = x_1, x_2, x_3, \ldots, x_n$ and $PK = \{v_1, v_2, v_3, \ldots, v_n\}$.

4 Cross-Chain Scheme

In this section, we introduce the cross-chain scheme proposed in this paper. It includes a cross-chain interaction process based on BLS signatures and a Reputation-Based Node Selection mechanism.

4.1 Cross-Chain Process Based on BLS Signature

The verification process based on the BLS signature consists of BLS signature algorithm and Shamir based secret sharing algorithm. BLS signature is used to avoid excessive storage cost of the final combined uplink signature, while Shamir secret sharing algorithm is used to achieve a certain number or more node endorsements for specific cross-chain transactions on the basis of BLS signature, thereby ensuring security.

As shown in Fig. 2, our cross-chain process consists of 6 stages: Request, Key Generation, Signature, Aggregation, Response, and Verification. Next, we will describe the 6 stages in detail.

Request: The request stages consists of 2 steps: main steps. First, node B_i of B initiates a request $Cross-chain.request$. Smart contract B_cross receives and parses the request. It will obtain the address of the other party of the cross-chain. After that, it forwards the request to A.

Key Generation: This stage also consists of 2 steps: Setup and Generation.

step1: $Setup(\lambda) \rightarrow \{G_1, G_2, G_T, e, g_1, g_2, p, h\}$

In the setup step, the smart contract B_{Cross} first parses the request. Determine the number of nodes participating in the cross-chain: n. And select the n nodes with the highest reputation. This includes A_j, A_x, A_y, and A_z where A_j is the node with the highest score. After that, B_Cross sends the parameters to the key generation center(KGC).

Then KGC obtains the relevant security parameters p of the algorithm and the elliptic curve bilinear mapping functions $e : G_1 \times G_2 \rightarrow G_T$. Multiplicative Cyclic group G_1, G_2 and its generator g_1, g_2, hash function h for mapping points onto elliptic curves. Specifically, Algorithm Setup(λ) inputs security parameters λ, and outputs as Two multiplicative Cyclic groups of a prime p: G_1, G_2 and a Bilinear map $e : G_1 \times G_2 \rightarrow G_T$. And g_1, g_2 as the generator of G_1, G_2. At the same time, a hash function is used in the scheme to map the hash digest of the signature data to the elliptic curve. These parameters are publicly available in the network:

$$h : \{0,1\}^* \rightarrow G_1 \tag{1}$$

step2: $Generate(G_2, p, t, n) \rightarrow \{MSK, MPK, SK, PK\}$

After initialization of the relevant parameters, KGC generates the master private key MSK, the master public key MPK, the threshold private key set

such as identity information of personal computers, user access information in gateways, etc. Due to the complexity and large number of IoT devices and the heterogeneous nature of IoT data. Adding all IoT devices to the same blockchain will be a very difficult behavior. In this regard, we divide the devices according to their types and organize the IoT devices of the same type to join their respective local-chains.

2. **Local-chain:** The local-chain connects all IoT devices with similar features, such as personal computers and mobile phones that belong to the same smart device. The whole system will have multiple local-chains. It mainly accomplishes the information storage function of the system and is responsible for the data recording, identity granting and world state updating tasks within the local-chains. Taking the first local-chain in Fig. 1 as an example, it is responsible for storing user information, personal wallet, and other data uploaded by cell phones and computers onto the blockchain.
   Moreover, it also records the cross-chain requests and reputation scores of the devices, and these results will be recorded on the local-chain in a summarized form. IoT devices can selectively add one or more local-chains according to their geographic locations and device characteristics.
3. **Global-chain:** There is one and only one global-chain in the system. All blockchain nodes are to be added to the global-chain. The global-chain has two main functions. One is to store the local-chain data abstracts and provide validation function for the local-chain data. When a blockchain node submits data to the local-chain, it can choose to upload the abstracts to the global-chain. Unlike the local-chain which stores a large amount of data, the global-chain only needs to access a small amount of summary information. Second, it provides information records during cross-chain interaction. During cross-chain interaction, some parameters and key information of data exchange will be submitted to the global-chain to ensure security.
4. **Cross-chain module:** The cross-chain module is an important component in linking all local-chains. In this module, we propose a decentralized cross-chain verification scheme. The scheme consists of two parts. The first one is a reputation-based node selection mechanism. We calculate the reputation value of a node based on its network latency and the number of times it has been invoked. Multiple nodes with a good reputation are selected to participate in cross-chain verification. This solves the centralization problem under the notary schemes. It can effectively avoid a single point of failure and improve cross-chain security.
   The second is the data verification technology based on BLS signature. In the cross-chain process, the selected nodes will utilize the BLS signature to sign and verify the data. After that, the node with the highest reputation utilizes BLS to aggregate the signatures of each node to jointly complete the verification of the data. The specific cross-chain process will be introduced in the next section.

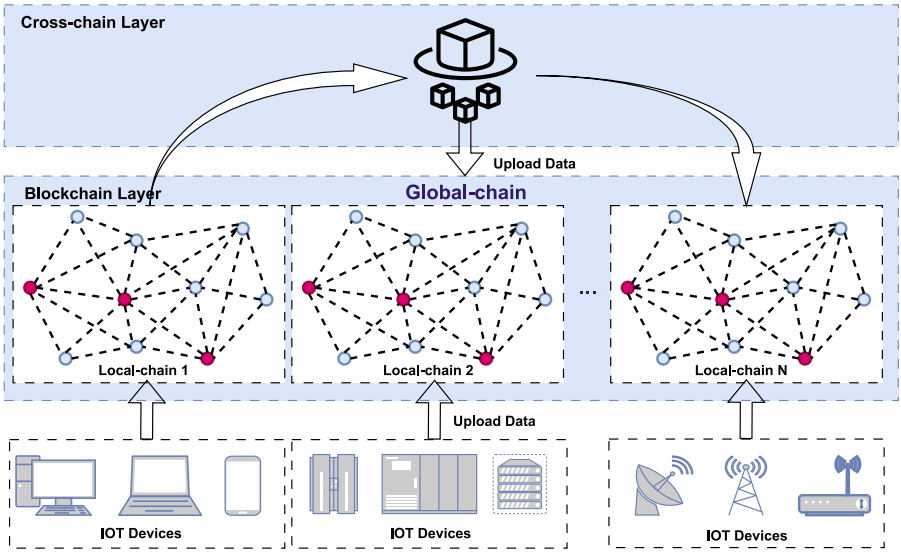


Fig. 1. System model.

which causes it to be more centralized. Once the node itself carries malicious intent or suffers an attack. It will not be able to guarantee the authenticity and trustworthiness of the cross-chain data. However, the cross-chain purpose of this scheme lies in asset transfer, without realizing a universal cross-chain approach.

Sun et al. proposed a decentralized cross-chain scheme [13], which combines a notary mechanism and hash-locking. By setting up multiple notaries and establishing an election mechanism, the degree of centralization of the cross-chain process is reduced. However, the cross-chain purpose of this scheme lies in asset transfer, without realizing a universal cross-chain approach. Ghosh et al. [8] proposed a decentralized gateway architecture that connects private blockchains to end users. The gateway employs a collective signature technique [14] to verify the data. However, this method allows only one-way communication and does not verify the identity of the requester.

3 System Architecture

In this section, we introduce the architecture and components of the system. As shown in Fig. 1, we build a multi-chain network structure in HyperLedger Fabric: including multiple local-chains, a global-chain, and a module that provides cross-chain functionality. IoT devices join the local-chains according to their different features and participate in the whole blockchain system. Next, we will introduce each component of the system in detail.

1. **IoT device:** IoT devices include all networked devices that have a need to participate in the blockchain. They upload important data to the blockchain,

the process of using BLS signatures to verify data in Sect. 4. The experiment results are stated in Sect. 5. Finally, Sect. 6 concludes this article.

2 Related Work

In this section, we introduce the research on blockchain in IoT and cross-chain technology.

2.1 Blockchain in IoT

Existing work applies blockchain to IoT. Blockchain can enhance the security of IoT systems. With encryption and digital signatures by cryptographic keys [7], IoT data can be protected through blockchain. And the smart contract carried by the blockchain can automatically update the firmware of IoT devices and close the vulnerabilities that are susceptible to attacks [4]. Moreover, IoT data stored on the blockchain data can be identified and verified anywhere and anytime. For example, the work of Lu et al. [11] develops a blockchain-based product traceability system that provides traceability to suppliers and retailers. In this way, the quality and originality of products can be checked and verified. Boudguiga et al. [3] proposed a decentralized mechanism to push updates to IoT devices using blockchain. The blockchain is used to record transactions for software updates pushed to the device to prevent malware updates on the device. In this case, there is no need for a trusted agent to deliver the updates as the updates propagated to the device through the blockchain have guaranteed integrity.

The diversity and heterogeneity of IoT devices pose a huge challenge for blockchain [12]. Optimization of blockchain networks is one way to address performance. In 2021, Zhou et al. [17]. proposed an optimization mechanism in resource-constrained IoT systems. They improve the performance of the system by dynamically adjusting optimal block assignments. Zhang et al. [16]. analyze the IoT traffic by establishing a blockchain network that matches the scale of IoT. And they implement a lightweight Bitcoin-like blockchain based on PoW to solve the problem of high traffic load and network congestion.

2.2 Cross-Chain Technology

Cross-chain technology was first applied to the exchange of assets. It is mainly used for the conversion of Bitcoin and Ethereum. In 2014, adam et al. propose sidechain [2], which is a blockchain system independent of bitcoin. Sidechain can access the Bitcoin network and interact with the Bitcoin ledger to enable asset transfers. As a separate blockchain, the technical solutions and consensus mechanisms adopted by sidechain are not restricted by the main chain. The notary schemes [9] is currently the most widely used cross-chain scheme. It set a trusted node in the blockchain system, which is responsible for completing cross-chain operations. However, the notary scheme uses a fixed node for cross-chaining,

during the transaction process [15]. So it leads to a significant gap in blockchain performance compared to centralized systems.

Cross-chain technologies are seen as an effective way to address blockchain scalability and performance. It can join different devices to separate blockchain networks and organize multiple blockchains through cross-chain technologies. The performance of blockchains can be effectively improved. Currently, the mainstream cross-chain technologies include sidechains/relays [2], notary schemes [9], and hash-locking [6]. However, all these technologies have shortcomings. Among them, hash-locking and sidechains/relays are mainly used for asset transfer rather than information interaction. This poses the problem that they focus more on security when crossing chains. And reduce the performance requirements. The notary schemes, on the other hand, suffer from the problem of centralization. In general, the notary mechanism is only responsible by fixed nodes in the cross-chain. The trustworthiness of the cross-chain data is completely guaranteed by the node's own credit. If the node carries malicious intent or it receives an attack, the security of the entire cross-chain process cannot be guaranteed.

In this paper, We propose an decentralized and efficient cross-chain scheme (DECS). First, We construct a multi-chain architecture consisting of multiple local-chains and a global-chain. And we add IoT devices to the local-chain network. Mutually independent local-chains can perform block transactions in parallel. It improves the blockchain and performance. Meanwhile, we propose a scheme in which multiple nodes jointly participate in cross-chain data verification. We design a calculation method for node reputation. It is calculated based on the network latency and invocation frequency of the nodes so that the selection of each node is as average as possible. This reduces the degree of centralization in cross-chain and effectively addresses the shortcomings of the notary schemes. During the cross-chain process, multiple nodes with the highest reputation are selected. They are responsible for verifying the data by using BLS signatures. Our major contributions in this article are summarized as the following aspects:

- We propose an Decentralized and Efficient Cross-chain Scheme in IoT systems. We adopt a multi-chain architecture to adapt the diversity of types of IoT devices and enhance the scalability of the blockchain.
- We design a cross-chain scheme based on BLS signatures and implement a smart contract that automatically updates node reputation. The scheme reduces the degree of centralization of the cross-chain process while guaranteeing the accuracy of cross-chain data.
- We implement and evaluate our scheme on the HyperLedger Fabric, and the results shows that our scheme can effectively improve system performance, and the selection of nodes is uniform and fair in the cross-chain process.

The rest of this paper is structured as follows. In Sect. 2, we introduce the related work of blockchain. In Sect. 3, we described the system architecture and cross-chain interaction mechanisms. And we introduced how to select cross-chain nodes based on reputation. Furthermore, we provide a detailed introduction to



DECS: A Decentralized and Efficient Cross-Chain Scheme in IoT System

Ying Gao(✉), Peihao Zhang, Qiaofeng Pan, and Xianfeng Qiu

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, People's Republic of China
gaoying@scut.edu.cn

Abstract. The rapid development of blockchain technology has demonstrated great potential to revolutionize various application domains, especially in the context of the Internet of Things (IoT). However, the sheer number and diversity of IoT devices pose significant challenges to the scalability and security of blockchain systems. To address these issues, cross-chain technology has emerged as a promising solution. Nevertheless, existing cross-chain solutions suffer from high centralization and inefficiency. Our approach involves constructing a multi-chain network capable of connecting different types of IoT devices, along with designing a collaborative cross-chain mechanism engaging multiple participating nodes. This collective participation diminishes the centralization inherent in the cross-chain process. Specifically, we propose a data verification method based on BLS signature. It aggregates cross-chain data signatures across multiple chains, which leads to a reduced storage burden on the blockchain. Furthermore, we introduce a reputation update algorithm that leverages network latency and cross-chain operation metrics to automatically update node reputation scores via smart contracts. Experimental results demonstrate that our solution achieves better decentralization and efficiency.

Keywords: blockchain · IoT · cross-chain · BLS signature

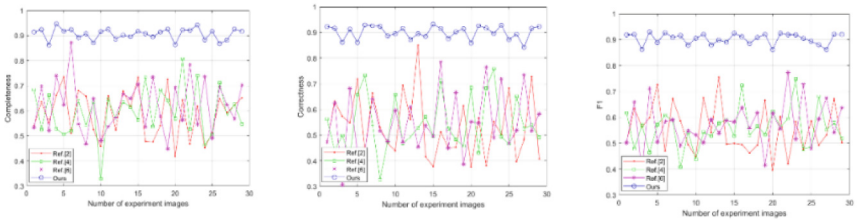
1 Introduction

Blockchain is gradually extending from small-scale applications to multiple fields, showing a bright development prospect. It has been applied to finance [1], food traceability [1], smart home [10], IoT, and other industries. In particular, IoT has a very good application prospect with blockchain due to its own decentralized features [5]. However, another feature of IoT is the large number and diversity of devices. With a large number of devices connected to the blockchain system, higher requirements are placed on the performance of the blockchain system. Unlike centralized systems, blockchain requires all nodes to reach a consensus

This work is supported by the Guangzhou Science and Technology Program key projects (202103010005).

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 128–142, 2024.
https://doi.org/10.1007/978-3-031-65126-7_13

4. Sebasco, N.P., Sevil, H.E.: Graph-based image segmentation for road extraction from post-disaster aerial footage. Drones **6**(11), 315 (2022). https://doi.org/10.3390/drones6110315
5. Vigneshwaran, S.A., Panneer, S.: Situational Analysis of Road Traffic Accidents-Acase of Madurai District rural areas (2020)
6. Zhang, X., Ma, W., Li, C., et al.: Fully convolutional network-based ensemble method for road extraction from aerial images. IEEE Geosci. Remote Sens. Lett. **PP**(99), 1–5 (2019). https://doi.org/10.1109/LGRS.2019.2953523
7. Cheng, G., Wang, Y., Xu, S., et al.: Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. IEEE Trans. Geosci. Remote Sens. **55**(6), 3322–3337 (2017). https://doi.org/10.1109/TGRS.2017.2669341
8. Alamri, A.M.: RoadVec-Net: a new approach for simultaneous road network segmentation and vectorization from aerial and google earth imagery in a complex urban set-up. GISci. Remote Sens. (2021). https://doi.org/10.1080/15481603.2021.1972713
9. Soni, P.K., Rajpal, N., Mehta, R.: Road network extraction using multi-layered filtering and tensor voting from aerial images. Egypt. J. Remote Sens. Space Sci. **24**(2), 211–219 (2021). https://doi.org/10.1016/j.ejrs.2021.01.004
10. Nguyen, T.L., Han, D.: Detection of road surface changes from multi-temporal unmanned aerial vehicle images using a convolutional Siamese network. Sustainability **12**(6), 2482 (2020). https://doi.org/10.3390/su12062482
11. Wei, Y., Wang, Z., Xu, M.: Road structure refined CNN for road extraction in aerial image. IEEE Geosci. Remote Sens. Lett. **14**(5), 709–713. 1027. https://doi.org/10.1109/LGRS.2017.2672734
12. Wang, S., Mu, X., Yang, D., et al.: Road extraction from remote sensing images using the inner convolution integrated encoder-decoder network and directional conditional random fields. Remote Sens. (2021).https://doi.org/10.3390/rs13030465
13. Alshehhi, R., Marpu, P.R., Woon, W.L., et al.: Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. Isprs J. Photogramm. Remote Sens. **130**(aug.), 139–149 (2017). https://doi.org/10.1016/j.isprsjprs.2017.05.002
14. Ganapathy, P., Skipper, J.A.: A novel ROC approach for performance evaluation of target detection algorithms. In: Conference on Automatic Target Recognition XVII. Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 207 Russ Engineering Center, 3640 Colonel Glenn Hwy, Dayton, OH 45435 (2007)
15. Alshaikhli, T., Liu, W., Maruyama, Y.: Simultaneous extraction of road and centerline from aerial images using a deep convolutional neural network. Int. J. Geo-Inf. (3) (2021). https://doi.org/10.3390/IJGI10030147
16. Pereg, D., Cohen, I., Vassiliou, A.A.: Sparse seismic deconvolution via recurrent neural network. J. Appl. Geophys. **175**, 103979 (2020). https://doi.org/10.1016/j.jappgeo.2020.103979



(a)Completeness (b)Correctness (c)F1

Fig. 4. Curve comparison for different methods.

5 Conclusion

Target detection in aerial images has been widely applied in many fields, including agriculture, forestry, electric power, land resources, urban planning, etc. In the acquisition process of aviation data, aircraft or UAV are constrained by the external environment, stability, wind resistance ability and clarity are limited, jitter phenomenon often occurs, camera Angle changes, etc. These uncertain factors will directly lead to the difficulty of road extraction.

In this paper, a semi-automatic framework combining DCNN and SAE is studied to extract road information from aerial images. SAE model is used to learn the correlation between complex data, and the brief expression is found from the feature perspective. The decoder network samples the feature map extracted from the encoder network back to the input image of the same size, and finally the correct classification output is obtained by softmax classifier. Experimental results show that the proposed algorithm reduces the complexity of the model and improves the speed of calculation.

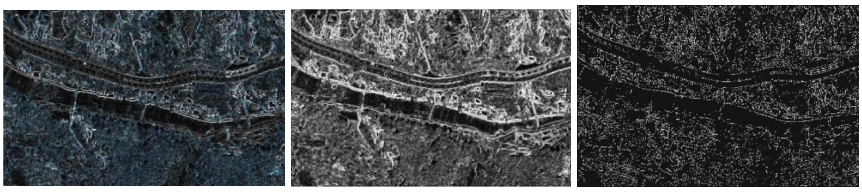
Acknowledgment. This work was supported in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01-03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474, and by Science and Technology Program in Xi'an city under Grant 21XJZZ0055.

References

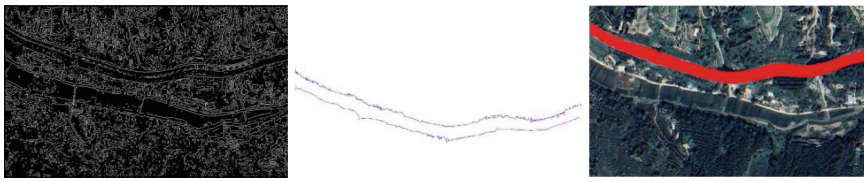
1. Eerapu, K.K., Lal, S., Narasimhadhan, A.V.: O-Seg-Net: robust encoder and decoder architecture for objects segmentation from aerial imagery data. IEEE Trans. Emerg. Top. Comput. Intell. **PP**(99), 1–12 (2021). https://doi.org/10.1109/TETCI.2020.3045485
2. Abdollahi, A., Pradhan, B.: Road extraction from open-source remote sensing dataset based on the modified deep convolutional autoencoders model. In: 43rd COSPAR Scientific Assembly Sydney, Australia, 28 January–04 February 2021. 2021
3. Tabibi, Z., Schwebel, D.C., Zolfaghari, H.: Road-crossing behavior in complex traffic situations: a comparison of children with and without ADHD. Child Psychiatry Hum. Dev. 1–8 (2021). https://doi.org/10.1007/s10578-021-01200-y



(a)Input (b)binary (c)Otsu



(d)Prewitt (e)Sobel (f)Valley



(g)Canny (h)Our method (i)Ground truth

Fig. 3. Comparison of different methods

Table 1. Objective Comparison

Methods	Completeness			Correctness			F1		
	aver	max	min	aver	max	min	aver	max	min
Ref. [4]	0.587	0.735	0.418	0.534	0.851	0.376	0.553	0.755	0.396
Ref. [6]	0.596	0.807	0.329	0.544	0.758	0.327	0.560	0.749	0.408
Ref. [8]	0.619	0.871	0.448	0.549	0.786	0.308	0.574	0.773	0.387
Our method	0.906	0.926	0.864	0.901	0.929	0.859	0.903	0.926	0.861

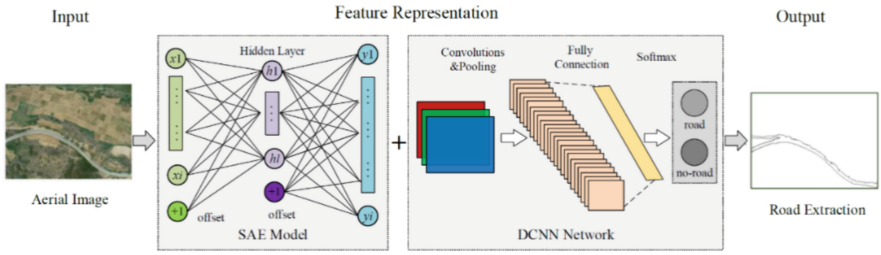


Fig. 2. Framework

4 Experimental Result and Analysis

4.1 Dataset Description

The experimental results of the above network framework are as follows. The dataset consists of two categories (urban roads and rural roads) with 900 images per category, where 400 images for training and 50 images for each group. For each image, the classification of ground truth is annotated by manual with the advice of experts carefully.

Through a large number of experiments, different initial values are selected, and the parameters with the highest performance in the cluster are selected to complete the network design. The evaluation system including Completeness, Correctness and F1 is used to test the road extraction performance. The formula can be denoted as in Eq. (4):

$$Com = \frac{TP}{TP + FN} \; Cor = \frac{TP}{TP + FP} \; F_1 = 2\frac{Com \times Cor}{Com{+}Cor} \tag{4}$$

where, *Com* means the completeness of matched with *GT* (ground truth) calculated by *TP* (truth positive) and *FN* (false negative), and *Cor* is correctness of matched with ground truth by *TP* and *FP* (false positive). F_1 is an overall that combines *Com* and *Cor*.

4.2 Result and Analysis

The proposed method is compared and to test robustness and flexibility of the related methods. The showing example from the testing dataset are shown in Fig. 3. In our testing images, the images were numbered as follows.

Figure 3 shows the different ways to achieve the road extraction, Table 1 gives the objective comparison of the results using Completeness, Correctness and F_1. Figure 4 is the line chart, which could exhibit the objective comparison more intuitively. In terms of Completeness, Correctness, and F1-score, the proposed method gives the best result in general.

The test can verify that the proposed method has some advantages compared to some existing methods in this field, and the results by our method is very close to ground truth, which are higher than the other methods. But for the Correctness, it is Performance is a bit poor which is caused by the almost indistinguishable gray level between the roads and the background in the bottom of the image.

3.2 Deep CNN

Methods based on deep learning have aroused widespread concerns, which establish a high-level semantic mapping relation by extracting the features. As a kind of feed-forward deep learning network, the Deep CNN is suitable for image feature extraction and recognition [15]. Usually, CNN architecture includes convolutional, mapping, pooling, fully connected, and output layers, that can be formed by stacking multiple underlying network structures.

First, feature extraction is performed in convolutional layer, and the formula can be denoted as in Eq. (3):

$$y_i = b_i + \sum\nolimits_i k_{ij} \ddot{A} x_i \tag{3}$$

where, y_i means the output image, x_i means the input image, $\otimes$ denotes convolution operator and k_{ij} is kernel function, finally, b_i is deviation value.

Second, the mapping layer employs a nonlinear activation function to obtain the feature map from the convolutional layer. The commonly used activation function is ReLU, sigmoid, tanh and softplus. Usually, the ReLU (Rectified Linear Units) function is employed as the activation function because the output will be zero, which could reduce the network and smooth the over-fitting problem.

Then, the pooling layer could avoid over-fitting phenomenon and maintain spatial invariance. And the full connection layer connects to all the previous layers including convolution layer or another full connection layer.

To train the network as a better performance, some operators, such as the local response normalization and dropout regularization method, are added to optimize results and speed up the training process. It randomly reduces the output of some neurons and reduces the neurons in the network that are no longer involved in the computation.

Finally, the classifier layer with full link is used to output in probabilistic form for each category. The most used loss function output in is the softmax function.

3.3 Framework

Therefore, a semi-supervised based deep learning method was proposed. The specific steps are as follows: Feature extraction part adopts the SAE model to study and find out the relationship between the optimal, get a concise expression, DCNN decoder of network from the encoder on the extraction of feature mapping samples back to the same size of the input image, and finally, at the end of the DCNN network using softmax classifier for the probability of road pixels in the final output.

Figure 2 shows the framework of our proposed method. The features learned by SAE, are applied to the convolution of a large number of training sets and test sets. The proposed DCNN network layers include one input, five conventional, two pooling, and one output. A max-pooling operation is performed between layer 1 and Layer 2. The average pooling layer follows the convolution of the five layers.

3.1 SAE Model

The performance of image classification is largely depended on the pros and cons of extracted features. SAE model is more suitable for unsupervised learning, which does not need a large number of tags during training massive aerial images. It can avoid the annotation of massive remote sensing images, and greatly improve the automation of the method. The unnecessary of annotation work can greatly improve the automation and efficiency of the algorithm [16].

The classic structure of SAE usually includes an input layer, in Fig. 1, a hidden layer, and an output layer, where +1 is the offset term.

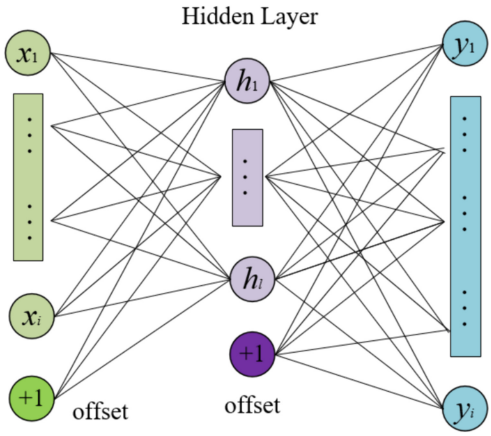


Fig. 1. SAE model

The loss function for neural network can be denoted as in Eq. (1):

$$J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}(x^i)-y^i\right\|^2\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^l)^2 \tag{1}$$

where, m is the amount of input samples, $(W,\ b)$ is the network parameter, n_l stands for the layers amount, s_l denotes the node amount in L layer, λ means the regularization and $h_{W,b}(x^i)$ means the output sample.

The SAE algorithm constrains the output of the hidden layer, so that the average could be high as 0. The loss function of SAE algorithm can be denoted as in Eq. (2):

$$J_{sparse}(W,b)=J(W,b)+\beta\sum_{j=1}^{s2}KL(\rho\|\hat{\rho})\tag{2}$$

where, ρ stands for the sparse parameter, $KL(\rho||\hat{\rho})$ measures the distributions.

Therefore, due to the aerial images have more complex backgrounds and targets. In this paper, a semi-automatic method combined Deep CNN with SAE (Sparse Auto-Encoder) is proposed to detect the road information from aerial image. First, the SAE model is carried out to learn the relationships and features of complex data and extract concise expressions from them automatically. Second, the encoder network completes the operation, and after the deep pooling and deconvolution operations, the intermediate features are extracted by the decoder network and sampled back to the input image of the same size on the map. Both convolution and pooling are implemented to reduce model complexity and boost distance calculation. Third, the final output is obtained by using the classifier, which is the probability distribution in the image representing the likelihood that the pixels in each region belong to the road and the non-road.

2 Related Work

In recent years, many methods have studied on road extraction from aerial image. Pradhan [8] proposed an automatic road extraction method by the neural network, which was superior to many methods of previous studies due to their ability to incorporate both multi-source information. Soni [9] presented a neural network to extract roads by a variety of texture parameters, and followed by the road vectorization stage. Experiments were carried out on different IKONOS and Quick Bird sample images to prove the road extraction capability of the proposed method.

Moreover, Nguyen [10] proposed a road extraction scheme based on feature learning, using convolutional neural network to capture the local structure of the road network. Due to the powerful learning ability of CNN, the road extraction method that we proposed can obtain high quality results. Wei [11] introduced a concise CNN for road extraction in aerial image. The paper proposed a new loss function which integrates the road geometry information into the cross-entropy loss. Experimental results showed that the model could perform well in accuracy, recall, F-score and accuracy.

Also, Wang [12] adopted a single patch architecture to extract roads from high-resolution images. Alshehhi [13] proposed a CNN network with integrated structure based on Alex-Net and VGG-net. Due to the large network structure, Alex-Net paid attention to the information of the large area. VGG networks focused on local details because of their small size. In this work, the training, verification and testing of the current popular deep learning models under different parameters have a good foundation for the identify and extraction of large geological and scientific data such as roads and buildings [14]. The accuracy of the road extraction is significantly improved.

3 Methodology

In our work, a semi-supervised based deep learning method was proposed, which combines Deep CNN (Convolutional Neural Network) with SAE (Sparse Auto-Encoder) to detect the road information from aerial image. In this part, the detail description of the concrete algorithms applied in our network is shown at first, and the overall framework and the algorithm execution process are elaborated on the follow.



Deep Neural Network Based on Sparse Auto-Encoder for Road Extraction

Sheng Liu, Shuxiao Chang, Ting Cao(✉), and Xinyue Li

Department of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi, China
caoting@xaut.edu.cn

Abstract. Road extraction from aerial image has realistic significance for GIS data updating. In view of the complexity challenging for acquiring road information, this paper proposes supervised model that combines Convolutional Neural Network (CNN) with Sparse Auto-Encoder (SAE) to cope with the road extraction task. First, the road features are extracted from the amount of non-annotated data using SAE model that aim to train the road features using CNN principle with implementing convolution and pooling to reduce model complexity. Second, the encoder network completes the operation, and after the deep pooling and deconvolution operations, the intermediate features are extracted by the decoder network and sampled back to the input image of the same size on the map. Third, the soft-max classifier categorizes images into roads and non-roads. Finally, the experiments verify that the proposed method outperforms the traditional methods and could achieve the satisfy result.

Keywords: Road extraction · aerial image · Deep learning · Convolutional Neural Network · Sparse Auto-encoder

1 Introduction

Road extraction from aerial images has vital usage in many applications including geographic information system, intelligent transportation system, environmental security and protection [1]. Various road extraction approaches can achieve road extraction successfully when the road exhibit obvious contrast respect with the non-road areas [2]. However, when the road with complex situation, such as road vehicles, buildings, tree occlusion cases, road extraction often appears discontinuous or gaps [3]. It is still challenging to deal with shadow or occlusion, geospatial information (urban, suburban or rural), and image scales, and obtain full and smooth road network automatically [4].

With the rapid development of deep learning in recent years [5], road extraction can be regarded as a classification task to distinguish aerial image into the road areas and the background areas [6, 7]. The state-of-the-art Convolutional Neural Network (CNN) is viewed as a successful deep learning model. CNN has advantages in hierarchical learning that makes it more efficient in feature extraction and image classification.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 120–127, 2024.
https://doi.org/10.1007/978-3-031-65126-7_12

12. Wang, H.: Research on underwater image enhancement based on improved MSRCR algorithm. **10**(06), 74–78+85 2020
13. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. **1**(4), 321–331 (1988)
14. He, K., et al.: Guided image filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. ECCV 2010. LNCS, vol. 6311, pp. 1–14. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15549-9_1
15. Cui, Q.N., Tian, X.P., Wu, C.M.: Improved algorithm of haze removal based on guided filtering and dark channel prior. **45**(05), 85–290 (2018)
16. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 79–116 (1998)
17. Haralick, R.M., Linda, G.S.: Computer and Robot Vision, vol. 1. Addison-Wesley Longman Publishing Co., Inc., Boston (1992)
18. Marr, D., Hildreth, E.: Theory of edge detection. Proc. R. Soc. Lond. Seri. B. Biol. Sci. **207**(1167), 187–217 (1980)
19. LeCun, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
20. Malik, J., Perona, P.: Preattentive texture discrimination with early vision mechanisms. J. Opt. Soc. Am. A **7**(5), 923–932 (1990)
21. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Process. Lett. **9**(3), 81–84 (2002)

Click to select an image and select the one you want to operate on. Click on 'Identify' to remove fog and recognize traffic signs on the fog map. The image after defogging and recognition will be generated in the right display box.

5 Conclusion

This project mainly focuses on the recognition of traffic sign images in haze weather. Combining the basic theory of digital image processing with traditional Retinex vision theory, research and improvement on image defogging are carried out. A Retinex defogging algorithm with guidance filtering influence factors is designed, and the recognition algorithm for traffic signs is analyzed and implemented. Effective information extraction is carried out on the image. The classification of traffic images is carried out on CNN pre trained networks, and training and learning are conducted using the BelgiumTS traffic sign dataset, Finally, our Convolutional neural network can get 93.89% recognition accuracy, but in practical application, this number still needs to be improved.

Acknowledgment. This work was supported in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01-03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, 2022JQ-264, and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474, and by Science and Technology Program in Xi'an city under Grant 21XJZZ0055.

References

1. Improved Dark Channel Based License Plate Recognition Method and System Implementation in Foggy Environment by Wu Tianyuan, Chongqing University of Posts and Telecommunications, 2019
2. Feng, S.G.: Research on traffic sign image recognition algorithm based on haze weather (2019)
3. FusedGAN: Moving foggy image restoration beyond the limits by Ren et al. (2021)
4. Recurrent Squeeze-and-Excitation Context Aggregation Net for Single Image Dehazing by Liu et al. (2020)
5. Xue, B., Li, W., et al.: Review on feature extraction of traffic sign recognition **40**(06), 1024–1031 (2019)
6. Huang, F.: Parallelization implementation of the multi - scale Retinex image enhancement algorithm based on a many integrated core platform. Concurr. Comput. Pract. Exp. **32**(22) (2020)
7. Jobson, D.J., et al.: A multiscale Retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans. Image Process. **6**(7), 965–976 (1997)
8. McCann, M.T., et al.: The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. IEEE Trans. Inf. Theory **56**(1) (2010)
9. Gao, F., et al.: A novel multi-scale Retinex algorithm for image enhancement. In: Proceedings of the 6th International Conference on Image and Graphics (2011)
10. Methods and Applications of Image Retinex Problem by Yang Xue, Lanzhou University, 2020
11. Xiaofang, W., Dengjie, F., et al.: SRCR image defog algorithm based on multi-scale detail optimization. **37**(09), 92–97 (2020)

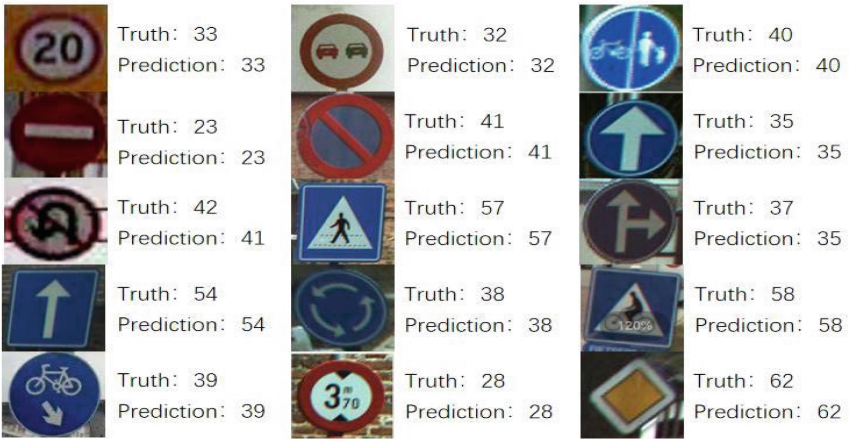


Fig. 8. Partial recognition results of improved defogging algorithm

4.3 Image Defogging and Traffic Sign Recognition System Based on Improved Retinex

The design of the improved Retinex based image defogging and traffic sign recognition system in this article is mainly divided into three parts: haze image selection, image display after defogging, and traffic sign box selection image display.

The Home screen of image processing consists of module selection and system menu, including four controls: button, panel, coordinate axis and text box. Each button has a corresponding callback function to switch the main interface to each module sub interface. The Home screen of GUI image processing system is shown in Fig. 9.

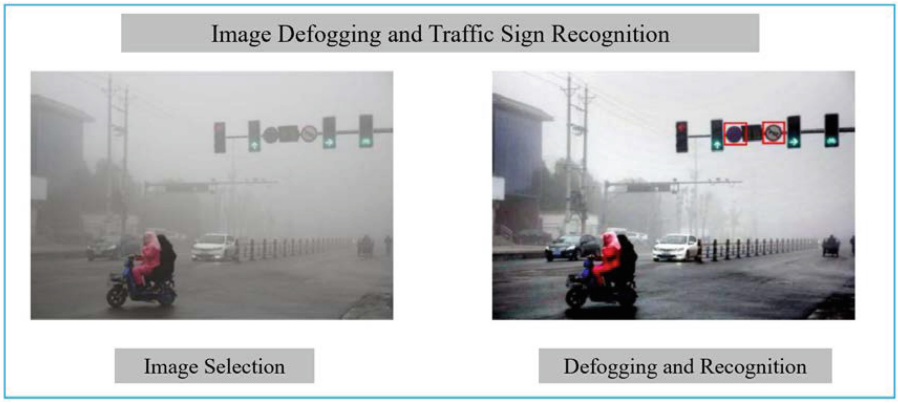


Fig. 9. Image Defogging and Traffic Sign Recognition System

Table 3. Experimental Results of the Second Group

algorithm	picture	Edge intensity factor	PSNR
SSR	pic (d)	41.0715	16.2509
MSR	pic (f)	42.0565	16.2701
Guided filtering	pic (h)	42.2539	16.7990
Weighted guided filtering algorithm	pic (j)	43.0994	16.8973

set graph prediction. In the figure, Input represents the true group identifier of the group of images to be predicted, while Prediction represents the predicted group identifier.

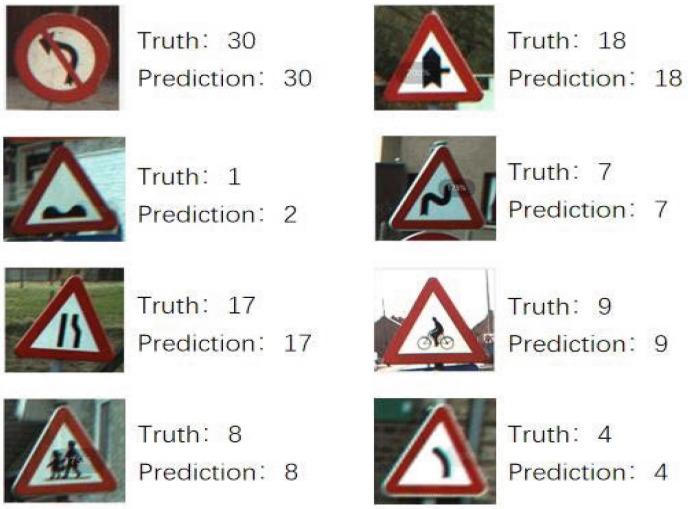


Fig. 7. Prediction Results of Test Set Part

It can be seen that there was an error in the recognition of the third image, but overall, the recognition rate can be maintained at a high level. After predicting all 2520 images in the test set, the accuracy of the prediction result is 92.73%.

Figures 8 show the prediction results of some traffic sign images processed by the improved Retinex defogging algorithm in this recognition algorithm, with two images showing recognition errors.

Figures 8 show the predicted results of some traffic sign images in the first set of experimental results of the defogging algorithm mentioned above. In this step, a total of 43 images from the first and second groups of the defogging algorithm results were used as predictive materials, with a total of 39 images predicting accurately, achieving an accuracy rate of 90.69%.



Fig. 5. Guiding Filter Processing Results



Fig. 6. Weighted Guided Filtering Processing Results

Table 2. Experimental Results of the First Group

algorithm	picture	Edge intensity factor	PSNR
SSR	pic (c)	82.8520	13.0193
MSR	pic (e)	83.5945	13.1975
Guided filtering	pic (g)	85.5671	13.9176
Weighted guided filtering algorithm	pic (i)	84.5239	13.8082

image noise. It can be seen that the experimental results of the algorithm combined with guided filtering contain more detailed information than the traditional Retinex algorithm. Therefore, it can be concluded that the improved Retinex algorithm can achieve edge information preservation to a certain extent.

4.2 Presentation and Analysis of Experimental Results on Traffic Sign Recognition

This part uses the Convolutional neural network pre training model designed in Sect. 3 to predict the test set, as shown in Fig. 7, the terminal output results of some kinds of test



Fig. 2. Original image



Fig. 3. SSR processing results



Fig. 4. MSR processing results

Therefore, two parameters, edge intensity factor and peak signal-to-noise ratio (PSNR), are selected as the comparison criteria, the images were divided into two groups for processing in the experiment. The average values of the experimental results of the two groups of images are shown in Tables 2 and 3 [20, 21]:

Among them, the edge intensity factor reflects the amount of edge information contained in the image. The larger the edge intensity factor, the clearer the image edges and the more edge information it contains; PSNR represents the ratio of signal to noise and is often used to evaluate noise and signal strength. A larger PSNR indicates less

Table 1. Network Architecture Diagram

Hierarchical network	feature maps	Convolutional kernel/pooling size	step
Conv2D	64 × 5	5 × 5	5
MaxPool2d	32 × 5	2 × 2	2
Conv2D	32 × 5 × 5	5 × 5	5
MaxPool2d	16 × 5 × 5	2 × 2	2
Fully connected layer	120	–	None
Fully connected layer	84	–	None
Fully connected layer	62	–	None

sets of images, including a total of 2520 images. The design and implementation of the training model are coded in the Python environment and PyTorch dependency package, which includes two convolutional layers, two maximum pooling layers, and three fully connected layers. A CNN network is implemented to achieve traffic image classification.

The structure of the designed network training model is shown in Table 1.

As shown in the above figure, after each convolutional layer is extracted, a pooling layer is added to reduce the dimensionality of feature information, thereby reducing computational complexity and accelerating network training speed.

After the training is completed, the model parameters are used to predict the test set, and the prediction accuracy for the test set can reach 92.73%. But for practical application requirements, this accuracy is not high. The CNN network model designed in this section is only a simple pre trained network, and its recognition performance still has room for improvement.

4 Experimental Results

4.1 Display and Analysis of Experimental Results of Defogging Algorithm

Here we use traditional single scale Retinex algorithm, multi-scale Retinex algorithm, and improved Retinex algorithm for experiments. In the traditional Retinex algorithm experiment, different Gaussian scales c are continuously adjusted to achieve better processing results. The final scale selection is: the scale in the single scale SSR algorithm is set to 15% of the image size; The multi-scale MSR algorithm has a mesoscale setting of 5% of the image size for small scales, 15% for medium scales, and 40% for large scales.

As shown in Fig. 2, the experimental example is shown in the original image. 4–2 shows the processing results of the single scale SSR algorithm, 4–3 shows the processing results of the multi-scale MSR algorithm, 4–4 shows the guided filtering processing results, and 4–5 shows the weighted guided filtering processing results (Figs. 3, 4, 5, 6):

It can be seen that both the traditional Retinex algorithm and the improved guided filtering algorithm can achieve good defogging results, achieving image enhancement results. However, the processing quality of edge information in each group of experimental results is difficult to compare with the naked eye.

1. Read in the guidance image I and the pending image P;
2. Calculate the mean and variance of I, the mean of the image P to be processed, and the product IP of I and P;
3. Calculate the linear correlation coefficient based on this $a_k = (IP - I_{mean}P_{mean})/(I_{Var} + \varepsilon)$; $b_k = P_{mean} - aI_{mean}$;
4. Calculate the mean of a_k and b_k;
5. Export filtering results: $q = a_{mean} * I + b_{mean}$;

This method uses guided filtering instead of Gaussian filtering to estimate illuminance images, ultimately resulting in an improved Retinex algorithm.

3.3 LOG Filtering Method

The edge detection algorithm of images requires both noise suppression and accurate positioning of edge information, and the LOG filtering method is an effective edge detection method. The LOG filter operator, also known as the Laplacian of Gaussian operator, and its corresponding operator, also known as the LOG operator, is the optimal filter for detecting image edge information based on image signal-to-noise ratio. This method comprehensively considers noise suppression and edge detection [16–18].

Perform the LOG operator on the test image to extract edge information, and the effect is shown in Fig. 1.



Fig. 1. Edge information extraction using LOG operator

Due to the extraction results being not suitable for observation, the pixel values of the resulting image were increased by a bias of 20 for observation. It can be seen that the LOG operator can effectively extract edge information from images.

3.4 Traffic Sign Image Recognition Based on CNN Network

This part will use Convolutional neural network (CNN) based on deep learning to implement a traffic sign image classification and recognition algorithm [19]. The training data is based on the BelgiumTS traffic sign dataset, which includes both training and testing sets. The training set contains 62 sets of images, each containing a certain number of logo images for training, with a total of 4575 images; The test set is also divided into 62

3.2 Improved Retinex Algorithm

The traditional Retinex algorithm uses Gaussian filtering for implementation, which has the effect of smoothing the image after processing. During the enhancement process of the image, there will be a loss of detail information, resulting in blurred information in the logo. So when calculating the illumination information of an image, we use guided filtering to estimate the illumination information of the image. Guided filtering is an edge preserving filter, and here we use guided filtering for illumination estimation [13, 14].

$$q_i = \sum_j W_{ij}(I) p_j \tag{1}$$

where p is the input image to be processed; I is the guiding image; q is the filtered output; W_{ij} is the filtering kernel, equivalent to $F(x, y)$, W_{ij} in the traditional Retinex algorithm is a function of guide image I. In actual calculations, we generally consider the output image q as the linear calculation result of guide image I. Assuming The output and input of the $W_{ij}(I)$ function are in a two-dimensional window Satisfy linear relationship within W_k:

$$q_i = a_k I_i + b_k, \forall i \in w_k \tag{2}$$

Among them, a_k and b_k is the constant term coefficient that needs to be calculated by us, and it is also the coefficient when the window center is located at k; w_k is the window; i and k are pixel indices.

As a local linear model, guided filtering is defined as the following Loss function in order to find linear correlation and minimize the difference between the output value of the fitting function and the true value p:

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left((a_k I_i + b_k - p_i)^2 + \varepsilon a_k^2 \right) \tag{3}$$

ω_k is right for a_k Correction compensation when the value is too large; The parameters about ε are used to adjust the blurriness of the image and the detection accuracy of edge information; εa_k^2 is used to suppress a_k value is too large. In terms of results, if the guide map does not contain edge information, the corresponding output mean filtering fuzzy result; If the guide map contains more edge information, the edge information will be reflected in the output image to achieve the preservation of edge information [15]. When calculating the coefficients of each window, a single pixel is usually described by multiple calculated linear functions. When calculating the output value of a single point, we take the mean of all calculated coefficients, and the final output result is as follows:

$$q_i = \frac{1}{|\omega_k|} \sum_{i \in \omega_k} (a_k I_i + b_k) = \overline{a}_i I_i + \overline{b}_i \tag{4}$$

Calculate the value of the linear coefficient from this. The algorithm flow of the guided filter is as follows:

Among the existing defogging algorithms, the retinex method has good performance and adaptability in image defogging. However, traditional retinex methods have drawbacks such as high computational complexity, difficulty in parameter selection, and limited effectiveness [2]. To address the problems of traditional retinex, we propose an improved retinex image clarity algorithm for image defogging and recognition of traffic signs in the image.

2 Related Work

Wang introduced traffic sign images into Convolutional neural network as training data to realize the classification function of traffic sign images. This method uses Convolutional neural network to study and classify the features of traffic signs, and can accurately recognize and classify different types of traffic signs. Li et al. proposed a method called FusedGAN to overcome the limitations of traditional defogging algorithms. This method combines the Generative adversarial network (GAN) and traditional image defogging technology, and restores images with complex haze by introducing multi-scale and multi-channel information fusion. FusedGAN can better remove the haze effect in the image and improve the image clarity and contrast [3]. Liu et al. proposed a single image defogging method based on the Recurrent Squeeze and Extraction Context Aggregation Network (R-SECA-Net). This method improves the quality and detail retention ability of image defogging by introducing attention mechanism and context aggregation. R-SECA-Net can adaptively adjust defogging processing, effectively reducing the problem of detail loss caused by haze [4]. Huang et al. proposed a traffic sign recognition algorithm based on CNN networks [5, 6]. The algorithm uses Convolutional neural network to extract and classify the features of traffic sign images, which can achieve high accuracy of traffic sign recognition.

3 Methodology

3.1 Traditional Retinex Algorithm

Among traditional Retinex image enhancement algorithms, the most common ones are single scale SSR algorithm, multi-scale MSR algorithm, etc., followed by iterative McCann algorithm and multi-scale Retinex algorithm with color restoration (MSRCR) [7–9]. The SSR algorithm needs to maintain a balance between contrast and image features, but the images to be processed vary in terms of shooting environment and imaging results. Therefore, the MSR algorithm is proposed based on the single scale algorithm. In order to compensate for the color deviation caused by interference such as haze and noise, a color restoration factor parameter C is added to the multi-scale MSR algorithm to adjust for the color deviation problem caused by the enhancement of local area contrast in the image. This corresponding algorithm is called the multi-scale Retinex algorithm with color restoration (MSRCR) [10–12].

Research on Traffic Sign Image Recognition Algorithms Under Complex Weather Conditions

Sheng Liu, Liming Qi, and Ting Cao(✉)

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
caoting@xaut.edu.cn

Abstract. In the transportation system, the influence of haze is more significant, such as license plate recognition, real-time monitoring, etc. The visibility of both people and equipment is greatly affected in foggy weather, leading to the emergence of foggy image processing. We analyzed the recognition requirements of traffic signs in foggy weather and conducted research on algorithms for removing fog from foggy images and extracting image edges. This topic mainly improved on the traditional Retinex algorithm, recognizing the loss of detail information in images under Gaussian filtering conditions. We applied guided filtering to the estimation of illumination images to achieve the preservation of image edge information. In terms of image recognition, the currently best performing LOG operator and Canny edge extraction algorithm were applied to achieve the extraction of detail information. Then, based on the background knowledge of Convolutional neural network, a small Convolutional neural network model is designed for training to realize the recognition and classification of traffic sign images. The experimental results show that the method proposed in this paper can achieve good functions in fog removal and traffic sign recognition.

Keywords: Haze · Edge detection · Retinex · Guided filtering · LOG operator · Convolutional neural network

1 Introduction

Traffic signs are the most important source for drivers to obtain road information during driving. As an important auxiliary facility in the road traffic system, traffic signs play an irreplaceable role. Haze inevitably reduces atmospheric visibility, and the accuracy and timeliness of driver information acquisition will be greatly negatively affected. Countless traffic accidents occur every year as a result. In addition, in foggy weather, the implementation of technologies such as intelligent monitoring, intelligent recognition, automatic navigation, and target tracking in outdoor environments has a significant negative impact. Therefore, in order to minimize the impact of haze weather on images, studying the implementation of traffic sign recognition algorithms under haze weather has extremely important theoretical value and practical significance [1].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 109–119, 2024.
https://doi.org/10.1007/978-3-031-65126-7_11

4. Zhi-Liang, D., Yi-Qun, P., Jiantong, X., et al.: Application of reinforcement learning algorithm in operation optimization of air conditioning system . Build. Energy Conserv. **7**, 7 (2020)
5. Wang, X., Wu, J., Liu, C., Yang, H., Du, Y., Niu, W.: Fault time series prediction based on LSTM recurrent neural network. J. Beijing Univ. Aeronaut. Astronsut. **44**(4), 13 (2018)
6. Zhao, S., Dong, X.: Research on speech recognition based on improved LSTM deep neural networks. J. Zhengzhou Univ. Eng. Ed. **39**(5), 5 (2018). (in Chinese)
7. Ren Zhihui, X., Haoyu, F.S., et al.: Chinese lexical segmentation for sequence annotation based on LSTM network. Appl. Res. Comput. **34**(5), 5 (2017)
8. Rao, Q.: Research on Dimension Emotion Recognition based on Context. Jiangsu University (2017)
9. Lei, X.: Research on Short-time Vehicle Flow Prediction Model based on Integrated LSTM. Chongqing University of Posts and Telecommunications (2019)
10. Zhang, X.W., Li, Y.Y., Huang, S., et al.: Population situation prediction method of COVID-19 based on LSTM. CN111798991A (2020)

data vehicles in the training set. Although the relative error of a few predicted values is slightly larger, most of the results can meet the requirements.

This paper also introduces the mean absolute percentage error (MAPE), which indicates the accuracy of the model, and from the numerical values in the MAPE, how accurate the model is (Table 5).

Table 5. Lists the data.

Map	MAPE value
Working day	0.083704
Holidays and festivals	0.106437
Rainy day	0.126403
Traffic control day	0.107493

The average MAPE value of the calculated scene is 0.106, indicating that the average error is about 10.6% and the prediction accuracy is 89.4%, indicating that the model has a high precision forecast of short-term traffic flow.

5 Conclusion

As an important part of ITS, traffic flow forecasting system based on LSTM provides great help to people's travel and traffic management department's work. The system uses python language, uses Tensortflow + keras architecture to establish LSTM prediction model, and processes the acquired data set, which is divided into training set and test set. The prediction results of the training set are compared with the test set, and a good prediction accuracy is obtained. The model can be used to predict the future traffic flow, which provides help for people's travel and the work of relevant departments.

Acknowledgment. This work was supported in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01-03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474, and by Science and Technology Program in Xi'an city under Grant 21XJZZ0055.

References

1. Wei, C.: Research on Taxi Shelter Design based on Service Design Concept. Xi 'an Polytechnic University (2017)
2. Liu, C.: Research on Urban Macro Travel Speed Prediction based on Classical Model and LSTM Model. Beijing Jiaotong University (2019)
3. Bonnin, R.: TensorFlow Machine Learning Project. Yao Pengpeng, trans. Beijing: People's Posts and Telecommunications Press (2017)

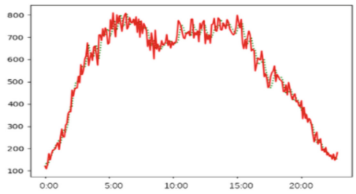


Fig. 2. Forecast of weekday traffic flow

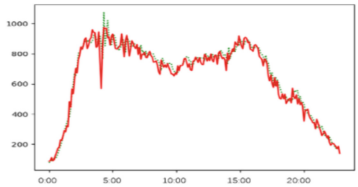


Fig. 3. Forecast of holiday traffic flow

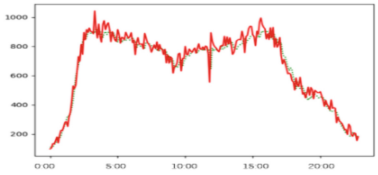


Fig. 4. Forecast of vehicle flow in rainy days

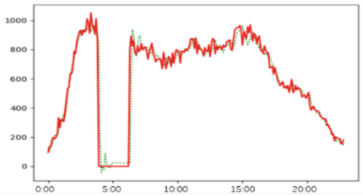


Fig. 5. Flow prediction under traffic control

which are brought into the model for training, and the prediction results are obtained. Use the plotting function to represent the prediction results. Among them, the real value is set as the red solid line, and the prediction result is the green dotted line. It can be seen from the figure that the short-term vehicle flow prediction model based on LSTM has a better vehicle flow prediction result for working days.

In general, the relative error of the sequence value of traffic flow on the predicted date can be obtained from the figure. To a certain extent, the main reason for the error in the analysis is that the model only considers the time characteristics of vehicle flow while other influencing factors are not taken into account, and there are not enough

3.2 Neural Network Training Steps for Long and Short Time Memory

Based on the above error propagation formula, the training steps of LSTM neural network are as follows:

Firstly, the output value of each neuron is calculated by the forward calculation formula, and the output value of the whole network is calculated according to the output value of each neuron;

The total output is compared with the actual value, the total error value is obtained, and the deviation of each neuron is calculated by the back propagation algorithm; The gradient descent algorithm is used to calculate and update each weight gradient according to the corresponding error value; After the new network weight is obtained, the forward calculation formula is continued to calculate the output of each neuron according to the new input data, and the final output of the network is obtained and then compared with the actual value.

When a certain error accuracy is reached, the network parameter is saved. At this time, when the network training is completed, if the error accuracy is not reached, the iteration is carried out continuously. Until the network output reaches a certain error accuracy.

4 Experimental Results and Analysis

4.1 Software Running Environment and Hardware Configuration

In terms of hardware environment, the Windows OS version is win11, the operating system is 64-bit, the processor model is Intel i5 8300 h, the display adapter is NVIDIA GTX1050ti, and the memory is 16 GB. In terms of software environment, the programming language is python, the python version is 3.6.8, and pycharm is selected as the integrated development environment.

4.2 Operation Results and Analysis

The built training set data was fed into the prediction model, and four different datasets were built using real-time data collected from individual detectors in the highway system across California's metropolitan areas between April 8 and May 10, 2018, classified by four different conditions: weekdays, holidays, rainy days, and traffic control days. Each part of the data is 4-day traffic flow data, and the sample set is built with the number of vehicles passing through high-speed cameras every 5 min. The vehicle flow prediction model of a short-duration memory network is constructed and the trained model is saved. The model input is obtained by the sequential sampling method described above, and then the vehicle flow of the same day is predicted on a rolling basis.

Here, plt.plot() function in Matplotlib is used to plot the prediction results and the real values, and 4 graphs of the traffic flow prediction results under different scenarios are obtained. Figures 2, 3, 4 and 5.

As shown in the figure above, the traffic flow of the following day is predicted based on the traffic flow data of the previous three days, and the total number of samples is 1152 each time. The 864 samples of the first three quarters are used as training sets,

The LSTM neural network uses two gates to control the cell state c, one is the amnesia gate, whose function is to judge how many cell states exist from the last time c_{t-1} to the time c_t, and the other is the input gate. Its function is to determine how much timely input x_t to the cell state c_t. Another gate is the output gate [9], whose function is to determine how much cell state c_t is output to the LSTM's current output value (Fig. 1).

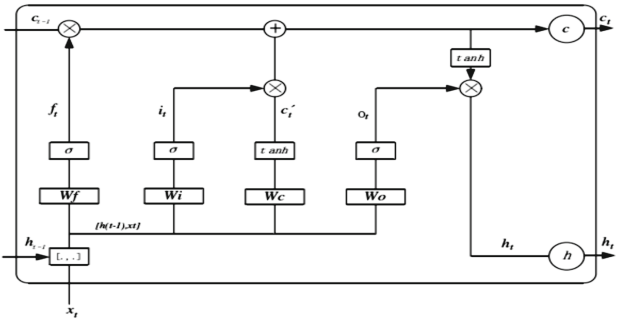


Fig. 1. Internal calculation flow diagram of a short-duration memory neural network unit

In the LSTM network, forward propagation and computation of information are also carried out through neuron transmission, and the network forward computation can be expressed by six formulas [10]. The first is the forgetting door and the calculation of the forgetting door is:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{3}$$

The input gate is calculated as follows:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{4}$$

The unit state describing the current input is calculated based on the output of the previous moment and the input of the current moment, and its calculation expression is as follows:

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{5}$$

$$\widetilde{c_t} = tanh(W_c \times [h_{t-1}, x_t] + b_c) \tag{6}$$

The number of gates that control the long-term memory acting on the current instantaneous output is output gate o_t, and its calculation formula is (7).

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{7}$$

The output result of the cell is determined by the output gate o_t and the cell state c_t, expressed as follows:

$$h_t = o_t * tanh(c_t) \tag{8}$$

The above formula 8 is the forward calculation expression of the whole long and short time memory network.

$$X_{scaled} = X_{std} * (max - min) + min \tag{2}$$

2.2 Model Output

In this paper, Tensorflow + keras architecture is used to construct the prediction model of long and short term memory network and train the neural network model with data set. As an open library [3], Tensorflow is a symbolic mathematical system based on data stream programming, which has been widely used in the research of machine learning, deep neural networks and other fields. Tensorflow is composed of multi-level institutions, which can use GPU and TPU for numerical calculation, and supports C and Python, which is completed in python language in Pycharm. In the process of using Tensorflow to build a neural network model, the environment should first be built. The version selected here is Tensorflow-GPU-2.6.0, and suitable CUDA should be installed. Secondly, various necessary packages should be imported into Pycharm. Such as keras, matplotlib, MKL sklearn etc., after setting various parameters, in this paper, the model set to 2 layer LSTM model. The process of constructing LSTM neural network using Tensorflow library is as follows: First, build a Sequential model with Sequential() to add layers. When building the neural network model, it is necessary to set the parameters of the model in the program, which has been introduced before. The number of hidden neurons in the first layer LSTM network is 50 and the output dimension is 50. Return_sequences is set to True and only the output of the last state is returned. The output dimension of the second layer is 100, the output dimension of the Dense layer is 1, and the activation function is liner.

After the traffic flow prediction model is built, it is necessary to train the neural network model with the sample set. In this paper, the loss function loss is defined as the mean square error function. RMSprop is used in the optimizer and batch gradient descent algorithm is adopted. According to the above, the training batch size is 8, which means that 8 data are extracted from the training set at a time for model training. During the experiment, the data of the training set was continuously converted into the prediction model, and the prediction results were compared with the real data of the verification set [4]. After the deviation was obtained, the backpropagation algorithm was used to optimize the training network, and the model was continuously trained. In this paper, 300 iterations of training were set as the end condition.

3 Related Technology

3.1 The Internal Structure of Long - Term Memory Network and the Calculation Method of Data

Long term memory neural network has completely overcome the shortcomings of ordinary RNNS [5], and is the most widely used RNN at present, which is widely used in many fields such as speech and picture recognition [6], natural language processing [7], emotion recognition [8] and so on. In the LSTM neural network, in addition to the short-term input signal sensitive state h, the cell state c is added and used to store the long-term state.

data of the first three days are constructed and trained as a sample set and retained as a vehicle flow prediction set of the following day to test the accuracy of the model.

In this paper, vehicle data at n time intervals before the road surface is used to estimate vehicle data at a subsequent time unit. Therefore, the sample set is constructed by inputting data at n time from past time into the network, where X represents the input network arrangement, the data at n + 1 time becomes the network output, and Y is taken as the network output arrangement. Neural networks generally associate input and output with information. After training, the network will form another input network and obtain a new output by direct association with specific data. Therefore, network input X and network output Y jointly construct the sample set of the experiment, and the data set is obtained in the form of rolling forward sampling, extracting n + 1 pieces of information at a time. Previously, n pieces of information were input X, and rolling prediction was also made in the prediction to build a larger number of samples. Therefore, in order to process the experimental data sequence into a short-duration memory network which can adapt to the data arrangement, it is necessary to use functions to complete the above requirements. There are a total of 3 * 288 = 864 experimental data in the three days from 4–8 to 4–10. When m + 1 experimental data is set at one time, the sample number constructed when m + 1 experimental data is used as the input of the experiment and m + 1 experimental data is used as the output of the experiment is (864-m) item. A data type in the constructed training sample set is shown in Table 4 (m = 11).

Table 4. Data after sequence transformation

X
……
130,155,125,124,111,117,91,119,79,112,81
155,125,124,111,117,91,119,79,112,81,88
125,124,111,117,91,119,79,112,81,88,71
124,111,117,91,119,79,112,81,88,71,78
111,117,91,119,79,112,81,88,71,78,118
117,91,119,79,112,81,88,71,78,118,82
91,119,79,112,81,88,71,78,118,82,71
119,79,112,81,88,71,78,118,82,71,83
……

Since the sigmoid function is used in the hidden layer of the experimental network, in order to achieve the speed of network convergence and prevent the problem of neuron saturation, it is generally required to normalize the training information. Here, MinMaxScaler() function in numpy library is selected to normalize the data.

$$X_{std} = \frac{X - X.\min(axis = 0)}{X.\max(axis = 0) - X.\min(axis = 0)} \tag{1}$$

The collected data includes six data types: Local Date, TIME, ID, CXDM, HPZL, and SYXZ. Table 2 describes the meanings of each data type.

Table 2. Original fields

fieldName	Local Date	TIME	ID	CXDM	HPZL	SYXZ
Meaning	Date	Record timestamp	Vehicle number	Vehicle code	Type of license plate	Nature of vehicle use

Based on the research content, part of the data is selected for extraction, and the data used are two types: Local TIME (date) and TIME (record time).

Based on the past traffic flow of a specific section of the road, the vehicle data of the next moment can be predicted, so that it is possible to train the neural network by constructing a data set based on the past traffic flow value. Before constructing the data set, the traffic flow value of the intersection should be extracted successively in time period. Sort the traffic according to the time sequence, and then use the resample() function in Pandas to summarize the time data in a period of 5 min when the traffic passes through the intersection. The summary results are shown in Table 3.

Table 3. Extracted data

Time	Volume
……	……
2018/4/8 0:05	130
2018/4/8 0:10	155
2018/4/8 0:15	125
2018/4/8 0:20	124
2018/4/8 0:25	111
2018/4/8 0:30	117
2018/4/8 0:35	91
2018/4/8 0:40	119
……	……

After pre-processing and time series value extraction, the traffic data passed by the intersection has been summarized into a five-minute time span of traffic flow data and stored in the document. All data starts at 2018-04-03 00:00:00 and ends at 2018-05-31 23:55:00. The data span is 5 min.

In this paper, the vehicle flow data of 30 days from midnight of April 3 to evening of May 10 are divided into four conditions according to the actual situation: working day, holiday, rainy day and traffic control. According to the situation, the vehicle flow

For the problem of traffic flow prediction, domestic and foreign scholars have conducted a lot of research, and built three research models, which are support vector machine model (SVM), deep learning model and neural network model (NN). Because of its own characteristics, SVM has been widely used in short-term traffic flow prediction, solving the problem of too small data and other problems. With the continuous development of deep learning, deep learning has also been studied in traffic flow prediction to a certain extent. The research on deep learning in traffic flow prediction is just at the beginning stage. According to the existing results, deep learning has high accuracy in predicting the results of specific data, but it also has certain limitations, and there are certain problems in the aspects of large calculation amount and long consumption time. Over the years, scholars have developed a variety of neural network models. Compared to previous machine learning, neural networks have more hidden units, using methods to abstract objects at a deeper level and extract hidden features that the data cannot see. Neural network also has the characteristics of strong adaptability.

On the whole, the development direction of short-term traffic prediction has gradually transited from linear model to nonlinear model, and from non-intelligent direction to intelligent prediction direction [2]. Based on the analysis of estimation accuracy, training model difficulty and reliability, the neural network model is usually better than the traditional model in the case of the same data, and is more suitable for short-term traffic flow prediction.

This paper will use the previous traffic data and build a model to predict the short-term traffic flow and analyze the results, so as to provide a helping hand to promote the smooth layout of ITS, reduce the current increasingly congested traffic situation, and ensure the safe and convenient travel of the people and the improvement of road travel experience.

2 Short Time Traffic Flow Forecasting Model

2.1 Model Input

The information used here is based on data from the Traffic Flow Prediction Project database collected in real time from individual detectors in the highway system across California's metropolitan areas. The time span is 1 month, 38 days from April 3 to May 10, 2018 (Table 1).

Table 1. Original data examples

Local Date	TIME	ID	CXDM	HPZL	SYXZ
2018/4/8	00:03:40	13631	K33	02	A
2018/4/28	14:38:29	41256	K31	02	A
2018/5/9	23:19:03	67293	K33	03	D
……	……	……	……	……	……

Design and Implementation of Traffic Flow Prediction Model Based on Short and Long Time Memory Network

Sheng Liu, Xinyue Li, Ting Cao(✉), and Shuxiao Chang

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
caoting@xaut.edu.cn

Abstract. Due to the randomness, fuzziness, time variability and uncertainty of traffic flow, it is difficult for traditional forecasting models based on time series or artificial neural networks to accurately reflect the actual traffic situation, etc. This paper takes the demand for short-term traffic flow forecasting of urban rail transit as the research object, and analyzes the implementation methods suitable for short-term traffic flow forecasting. LSTM neural network was used to construct the model for simulation experiment analysis. The results of data analysis show that the LSTM neural network model obtains the minimum average absolute percentage error MAPE value of 10.6% and the highest average accuracy of 89.4%, which has a good prediction effect and can improve the prediction work of short-term traffic flow.

Keywords: Neural network · Integrated learning · LSTM · Short time traffic flow forecast

1 Introduction

In recent years, with the rapid development of our economy, the number of motor vehicles and non-motor vehicles in urban areas is also increasing. The overall number of motor vehicles has been increasing, and the problem of the sharp increase in urban traffic pressure has also followed. In order to solve many problems such as the worsening of urban traffic congestion, in the context of the continuous development of modern social science and technology, the Internet of Things and other emerging technologies have made people put forward more advanced ways to improve traffic conditions, and gradually formed the concept of intelligent transportation system (ITS).

ITS is an efficient management system, which relies on road engineering to reduce traffic congestion and natural pollution and ensure smooth traffic operation. Traffic flow prediction is a very key part of ITS [1], which can quickly help the system to achieve timely, dynamic, accurate and reliable quantitative prediction of vehicle data and future road traffic flow conditions.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 99–108, 2024.
https://doi.org/10.1007/978-3-031-65126-7_10

Transportation Networks

6. Deckersbach, T., Wilhelm, S., Keuthen, N.J., Baer, L., Jenike, M.A.: Cognitive-behavior therapy for self-injurious skin picking: a case series. Behav. Modif. **26**(3), 361–377 (2002)
7. Odlaug, B.L., Grant, J.E.: Clinical characteristics and medical complications of pathologic skin picking. Gen. Hosp. Psychiatry **30**(1), 61–66 (2008)
8. Tucker, B.T., Woods, D.W., Flessner, C.A., Franklin, S.A., Franklin, M.E.: The skin picking impact project: phenomenology, interference, and treatment utilization of pathological skin picking in a population-based sample. J. Anxiety Disord. **25**(1), 88–95 (2011)
9. Prochwicz, K., Antosz-Rekucka, R., Kałużna-Wielobób, A., Sznajder, D., Kłosowska, J.: Negative affectivity moderates the relationship between attentional control and focused skin picking. Int. J. Environ. Res. Public Health **19**(11), 6636 (2022)
10. Brämer, G.R.: International statistical classification of diseases and related health problems. Tenth revision. World Health Stat. Q. **41**(1), 32–36 (1988)
11. Guha, M.: Diagnostic and statistical manual of mental disorders: DSM-5. Ref. Rev. **28**(3), 36–37 (2014)
12. Cohen, L.J., et al.: Clinical profile, comorbidity, and treatment history in 123 hair pullers: a survey study. J. Clin. Psychiatry **56**(7), 319–326 (1995)
13. Woods, D.W., et al.: The trichotillomania impact project (tip): exploring phenomenology, functional impairment, and treatment utilization. J. Clin. Psychiatry **67**(12), 1877 (2006)
14. Higashi, S., Goto, D., Okada, S., Shiozawa, N., Makikawa, M.: Development of wearable EMG measurement system on forearm for wrist gestures discrimination. In: 2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech), pp. 250–251. IEEE (2019)
15. Turlapaty, A.C., Gokaraju, B.: Feature analysis for classification of physical actions using surface EMG data. IEEE Sens. J. **19**(24), 12196–12204 (2019)
16. Fajardo, J.M., Gomez, O., Prieto, F.: EMG hand gesture classification using handcrafted and deep features. Biomed. Signal Process. Control **63**, 102210 (2021)
17. Campanini, I., Disselhorst-Klug, C., Rymer, W.Z., Merletti, R.: Surface EMG in clinical assessment and neurorehabilitation: barriers limiting its use. Front. Neurol. **11**, 556522 (2020)

OCD presents a complex and heterogeneous mental health challenge, requiring in-depth investigation to comprehend its underlying mechanisms fully. This paper ventures into the unexplored territory of OCD and its potential link to human skin textures, with a specific focus on Excoriation Disorder, commonly known as chronic skin-picking. By conducting a comprehensive analysis of existing literature, this study seeks to elucidate the cognitive aspects, memory impairments, and potential neurobiological factors contributing to this unique association.

5 Conclusion

In summary, Obsessive-Compulsive Disorder (OCD) is a complex mental health condition challenging our understanding. This study explores the potential link between OCD and human skin textures, focusing on Excoriation Disorder (chronic skin-picking) through comprehensive literature analysis. Examining cognitive, memory, and neurobiological factors, it also considers human-computer interaction (HCI) in analysis and treatment, emphasizing skin texture aspects. Two avenues for understanding OCD through skin texture emerge: repetitive movements due to memory issues, resulting in enlarged objects with skin texture imprints, and identifying OCD patterns through distinctive skin marks from compulsive skin peeling. Inspired by Exposure and Response Prevention therapy, magnifying skin texture details simulates ERP, countering perfectionism. This innovative approach provides insights into OCD complexities, underlining skin texture's role in understanding and treating the disorder. By integrating cognitive and neurobiological aspects, this study advances our understanding of the intricate OCD-skin texture relationship, aiding OCD research and interventions for a comprehensive perspective on this condition.

Acknowledgment. This work is supported by the Shenzhen Science and Technology Innovation Commission (Stabilisation Support Programme).

References

1. Shusta, S.R.: Successful treatment of refractory obsessive-compulsive disorder. Am. J. Psychother. **53**(3), 377–391 (1999)
2. Savage, C.R., Baer, L., Keuthen, N.J., Brown, H.D., Rauch, S.L., Jenike, M.A.: Organizational strategies mediate nonverbal memory impairment in obsessive-compulsive disorder. Biol. Psychiat. **45**(7), 905–916 (1999)
3. Benzina, N., Mallet, L., Burguière, E., N'diaye, K., Pelissolo, A.: Cognitive dysfunction in obsessive-compulsive disorder. Curr. Psychiatry Rep. **18**, 1–11 (2016)
4. Law, C., Boisseau, C.L.: Exposure and response prevention in the treatment of obsessive-compulsive disorder: current perspectives. Psychol. Res. Behav. Manage. **12**, 1167–1174 (2019)
5. Grant, J.E., Chamberlain, S.R.: Trichotillomania and skin-picking disorder: an update. Focus **19**(4), 405–412 (2021)

precision score, recall score, f1 score, and accuracy. Compared results are shown in Fig. 3.

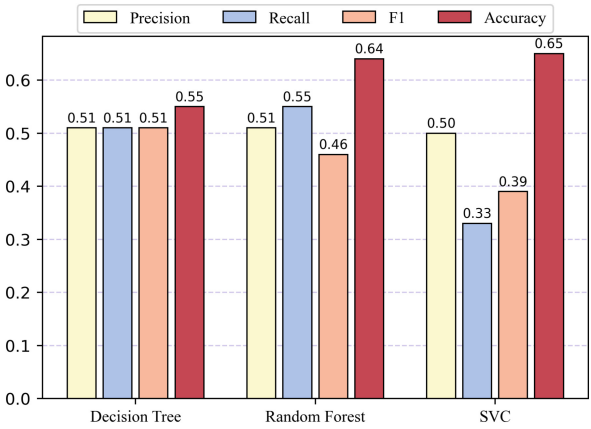


Fig. 3. Compared results of EMG classification on the WESAD dataset.

We can see that the SVC achieves the best accuracy which is 0.65 and the Decision Tree model achieves the best F1 score which is 0.51. We can also find that although the accuracy of SVC is high, its F1 is not high, because the number of samples in the two categories of the data set is unbalanced, and SVC is more likely to classify the samples to the one with a larger number.

4 Discussion

The findings carry important implications for OCD research and intervention. Understanding the connection between OCD and skin texture enhances comprehension of this intricate disorder. The immersive ERP approach offers a promising therapeutic avenue for directly confronting obsessions and compulsions.

Despite the study's strengths, certain limitations must be acknowledged. Sample size and specific OCD subtypes may influence the generalizability of the findings. Additionally, the accuracy and limitations of wearable devices can impact data quality.

In conclusion, this academic discourse explores the intricate relationship between OCD and human skin textures. Utilizing wearable devices and an immersive ERP approach, valuable insights into obsessive-compulsive traits through skin texture analysis are gained. The combination of EMG signals and skin features enables a holistic exploration of OCD's underlying mechanisms. This research contributes to the field by advancing our understanding of OCD and proposing potential skin texture-based interventions in OCD assessment and treatment.

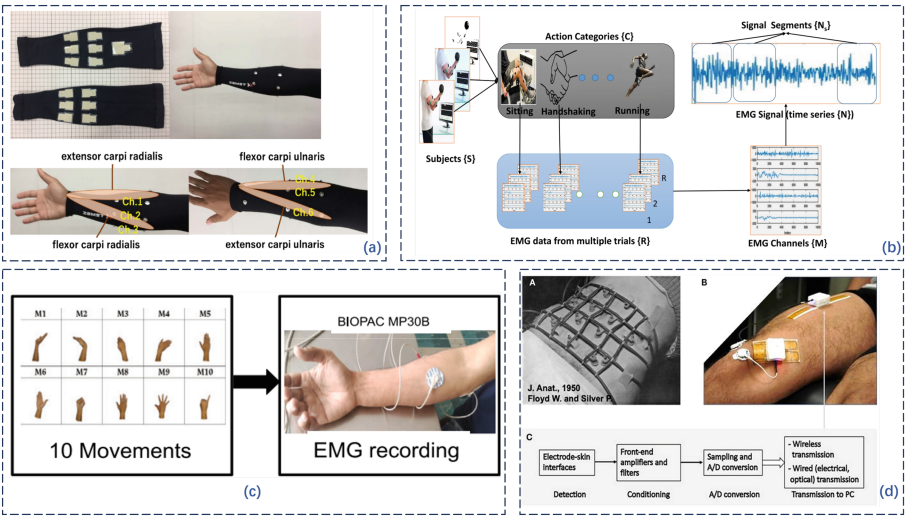


Fig. 2. Existing schematic diagram of AIoT Framework for EMG-based OCD detection.

In summary, the proposed framework aims to collect the EMG signal, then capture the correlation between OCD and EMG, and finally complete OCD detection. We also review some existing schematic diagrams of the AIoT Framework for EMG-based OCD detection as Fig. 2.

Figure 2 (a) shows a wearable EMG measurement system on the forearm for wrist gesture discriminations, in which twelve electrodes are utilized to measure EMG from six locations [14]. Figure 2 (b) shows surface EMG Data collection for the classification of physical actions, which observations consist of C categories with R trials [15]. Figure 2 (c) shows arm EMG signal collection for hand gesture classification, which contains 10 gesture classes, and each gesture is repeated 100 times by the participants during collection [16]. Figure 2 (d) illustrates EMG detection with wearable sensors for abdominal muscle monitorings in 1950 as shown in (A) and the modern system for knee EMG detection in (B), where (C) illustrates the system [17].

3.4 EMG Classification Case

To better understand the EMG signal, we provide a classification case with the Wearable Stress and Affect Detection (Wesad) dataset. This dataset provides EMG signals with four different emotion states: neutral, stressed, amused, and meditated. We first divided these states into neutral (neutral and neutral) as label 0 and active (stressed and amused) as label 1. Then we downsample the signal to 256 Hz and use a sliding window approach to obtain 10-second segments. Finally, we use Decision Tree, Random Forest, and Support Vector Classifier to make a classification. To evaluate the performance of models, we compare the

paving the way for the development of targeted interventions for affected individuals.

3.3 Proposed AIoT Framework for EMG-Based OCD Detection

We propose an AIoT for EMG-based OCD detection, which contains three main modules: signal collection, signal processing, and OCD detection. The proposed framework is shown in Fig. 1.

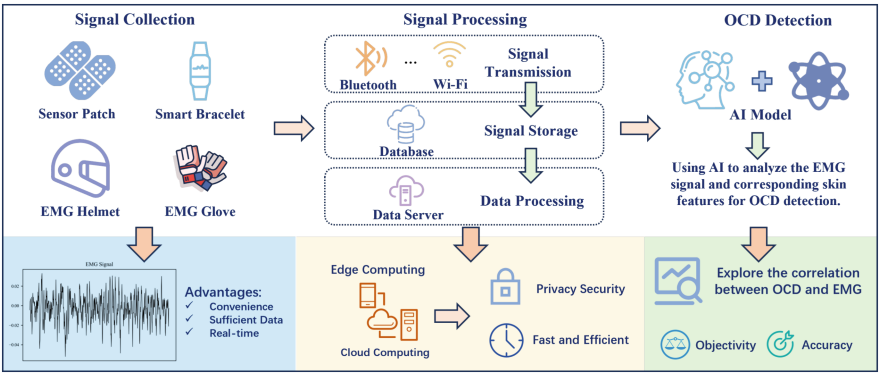


Fig. 1. AIoT Framework for EMG-based OCD detection. The proposed AIoT Framework for EMG-based OCD detection can be divided into three main parts: signal collection, signal processing, and OCD detection.

First, wearable devices are widely used for EMG signal collection, such as skin sensor patches, smart bracelets, EMG helmets, and EMG gloves. These devices can collect sufficient EMG data in a convenient way, and provide a way to collect data in real time. These wearable devices can also be personalized according to the needs of the user and the collected signal to meet the requirements of different users. Then, IoT techniques are used for EMG signal processing. Specifically, EMG signals are transferred with Bluetooth or Wi-Fi from wearable devices to the database for signal storage, and data servers are used for data processing with edge computing and cloud computing. In this way, it can provide fast and efficient signal storage and processing with privacy security. In addition, the database can store long-term signal data to achieve continuous recording, which can achieve better correlation mining. Finally, AI models such as CNNs and RNNs are used for analysis. Using AI models to analyze the EMG signal and corresponding skin features can help explore the correlation between OCD and EMG in an objective and accurate way. In addition, other AI technologies, such as big data, can also provide effective assistance to explore the correlation between OCD and EMG. With the correlation explored, a classification model can be constructed for OCD detection.

utilization of wearable technology enhances understanding of skin-picking disorder's complexities and its potential connections to OCD, potentially leading to more effective therapeutic approaches for managing these conditions.

3.2 Method for Trichotillomania

The research methodology revolves around data analysis using wearable devices, specifically focusing on EMG signals and corresponding skin features. This approach aims to explore the correlation between OCD and EMG data while integrating physiological and skin texture information to gain deeper insights into the association between OCD symptoms and changes in skin texture.

The study intends to utilize the EMG signals to investigate individuals with Trichotillomania, analyzing the muscle activity involved in hair-pulling. By capturing EMG signals and corresponding skin features, the research seeks to understand the patterns and frequency of hair-pulling episodes, as well as the physical consequences, such as hair loss and potential skin injuries. The analysis of this data may reveal connections between EMG signals, skin texture changes, and the severity of Trichotillomania symptoms, offering a comprehensive exploration of the disorder's impact on the skin.

Additionally, the wearable devices' data analysis will be extended to explore hair-pulling patterns across different body areas, such as the scalp, eyebrows, eyelashes, and pubic area, as mentioned in the previous. This investigation aims to identify any differences in the intensity or frequency of hair pulling in various regions, contributing to a more nuanced understanding of Trichotillomania and informing targeted interventions.

Furthermore, the duration of hair-pulling episodes will be assessed through the EMG signals, enabling researchers to understand the persistence and intensity of these behaviors. The study seeks to correlate the duration of hair-pulling episodes with the severity of Trichotillomania symptoms and potential skin damage, facilitating the development of interventions to interrupt hair-pulling behavior and promote healthier coping strategies.

The research methodology also aims to explore correlations between OCD symptoms and EMG data related to skin-picking or hair-pulling behaviors. By capturing EMG signals during episodes of skin picking and analyzing corresponding skin texture features, the study intends to investigate potential associations between the intensity of skin-picking behaviors, the severity of OCD symptoms, and observable skin changes. This integrated approach offers valuable insights into the relationship between OCD and Body-Focused Repetitive Behaviors (BFRBs) like skin picking, contributing to a deeper understanding of their underlying mechanisms and potential links between the disorders.

The research methodology, centered on wearable devices and data analysis of EMG signals and skin features, provides a unique opportunity to comprehensively explore the association between OCD symptoms and changes in skin texture in BFRBs like Trichotillomania and skin picking. The integration of physiological and skin texture information enhances our understanding of these disorders and may lead to advancements in both OCD and BFRB research,

3 Proposed Method

The research methodology is centered on data analysis using wearable devices. Electromyography (EMG) signals and corresponding skin features are captured to explore the correlation between OCD and EMG data. Integrating physiological and skin texture information aims to provide deeper insights into the association between OCD symptoms and skin texture changes.

3.1 Method for Skin-Picking Disorder

The research methodology revolves around data analysis using wearable devices, with a primary focus on capturing EMG signals and corresponding skin features. The objective is to explore the correlation between OCD and EMG data while integrating physiological and skin texture information to gain deeper insights into the association between OCD symptoms and skin texture changes. Through the application of wearable devices, the study aims to investigate individuals with skin-picking disorder, analyzing the EMG signals during episodes of skin-picking and correlating them with the corresponding skin texture changes. This analysis seeks to uncover patterns in EMG activity associated with skin-picking behaviors, contributing to the identification of potential markers of the disorder.

Moreover, the research methodology seeks to identify specific characteristics of OCD through the analysis of skin texture changes in conjunction with EMG data. By integrating physiological and skin texture information, the study endeavors to explore whether certain patterns in skin texture changes are associated with OCD symptoms, providing a deeper understanding of the underlying mechanisms.

Additionally, the research methodology aims to assess the psychological consequences of skin scratching by analyzing EMG data and corresponding skin features. The study intends to explore the correlation between the severity of skin-picking behaviors and the level of distress experienced by individuals, shedding light on the psychological impact of skin-picking and its potential role in the maintenance of the disorder.

The research methodology can contribute to the classification of skin picking as a Body-Focused Repetitive Behavior (BFRB) by capturing EMG signals during recurrent and habitual actions directed at the body. By establishing a link between the repetitive behaviors observed in skin picking and the broader category of BFRBs, the study provides a foundation for understanding the relationship between different BFRBs and contributes to a comprehensive understanding of these conditions.

In conclusion, the research methodology focuses on data analysis using wearable devices, particularly EMG signals and skin features, to comprehensively explore the association between OCD symptoms and skin texture changes in skin picking disorder. By integrating physiological and skin texture information, the study aims to advance knowledge in both OCD and BFRB research, informing the development of targeted interventions for affected individuals. The

able to control himself to think about this phenomenon over and over again; some obsessive-compulsive disorder sufferers cannot accept the imperfections on their skin, such as scars or pimple marks and thus repeatedly think about them and touch them, which not only affects their daily life, but also causes more damages to their skin itself in severe cases. Individuals with OCD demonstrate slowed performance on neuropsychological tests due to excessive focus on test accuracy and interference from intrusive obsessive thoughts. This impairment may be associated with dysfunction in the ventral prefrontal cortex system [3]. Individuals with OCD may have concerns about incorrectly filling their exam numbers during exams, leading to repetitive checking behaviors that impact their cognitive reasoning and problem-solving skills, potentially resulting in an inability to finish the exams within the time constraints.

ERP is a form of Cognitive-Behavioral Therapy (CBT) designed to assist patients in confronting and exposing themselves to fears or discomfort related to their obsessive thoughts, with the goal of preventing the performance of compulsive behaviors to alleviate the distress. The ultimate goal of ERP (Exposure and Response Prevention) is to challenge patients' conditioned fear and response to stimuli, allowing them to experience the outcome and gain an understanding that the feared stimuli are, in fact, safe [4]. Amplifying the skin texture under certain circumstances exposes the patient to an immersive imperfect skin texture, which reduces the excessive focus on perfection and imperfection and acts as a palliative.

Trichotillomania (hair-pulling disorder) and skin-picking (excoriation) disorder are neuropsychiatric disorders that usually occur in conjunction with OCD, but are not recognized by professionals [5]. Skin-picking disorder is a psychiatric disorder characterized by repeated scratching or picking at the skin, resulting in skin injuries such as minor ulcers, hyperpigmentation, shallow scars, and even, less commonly, more severe skin disfigurement and skin infections [6–8]. Some of the characteristics of OCD can be explored by analyzing the textures of the area of skin injuries with the electromyographic information generated during the behaviors. In addition to medical consequences, psychological sequelae of skin scratching have been identified, including clinically significant distress and different areas of dysfunction [9]. Skin picking is categorized as a body-focused repetitive behavior (BFRB) characterized by recurrent and habitual actions directed at the body [10].

Trichotillomania is characterized by a repetitive act of pulling one's own hair, leading to hair loss and potential dysfunction [11]. This condition primarily involves pulling from the scalp, eyebrows, and eyelashes, although any body part with hair, such as the pubic area, can also be affected [12,13]. It is not uncommon for individuals with Trichotillomania to engage in hair pulling from multiple areas, and the episodes of pulling can vary in duration, ranging from a few minutes to several hours. Magnify and observe the multiple skin areas involved in the act of hair plucking, looking for specific skin features.

enabling the acquisition of a comprehensive dataset. Based on this data, the study draws conclusions regarding the unique manifestations of OCD symptoms related to human skin textures, providing novel insights into the intricate connections between these elements.

The innovation point of this study lies in its interdisciplinary approach, integrating cognitive and neurobiological aspects to analyze OCD through the lens of human skin textures. This holistic exploration of OCD-related behaviors and skin texture traces offers a novel perspective on the complexities of the disorder, potentially paving the way for advancements in research and treatment strategies. By unveiling potential links between OCD and skin textures, this study aims to contribute valuable insights to the broader understanding of the disorder, ultimately aiming to improve the lives of individuals affected by OCD.

2 Background

Obsessive-compulsive disorder (OCD) is a multidimensional heterogeneous disorder characterized by impairments in volitional processes, including attention, association, thinking, and behavioral autonomy. The core of this disorder lies in the disharmony within the self [1]. This phenomenon manifests as a conflict between rational thoughts and the intrusive, distressing obsessions characteristic of OCD. Patients may express frustration, guilt, and shame, recognizing the irrationality of their obsessions and compulsions but finding it exceedingly difficult to resist or control them. This internal turmoil can significantly impact their self-esteem and emotional well-being, contributing to a cycle of distress and compulsion. Addressing this disharmony becomes a central therapeutic goal, as it is vital for helping patients develop effective coping strategies and building resilience in managing OCD symptoms. Some individuals with OCD may present with comorbid dissociative identity disorder (DID), characterized by difficulties in memory judgment related to both actual and imagined execution. Patients exhibit a lack of confidence in their memories, resulting in compulsive symptoms such as repetitive checking [1]. Repeated checking behaviors include skin-to-object contact, such as touching the door handle when repeatedly checking whether the door is closed, rummaging through a bag when checking whether the contents are adequately stored, repeatedly touching the gas, water, and electricity switches with the fingers when checking whether the switches are turned off, and so on.

Due to the presence of memory impairments in OCD, patients tend to focus more on event details, affecting their memory function. Some scholars propose that episodic memory deficits are secondary to executive function impairment, attributed to deficits in memory encoding [2]. As a result of paying attention to detailed content, it is easy for patients to see only specific parts of things or events and become too obsessed with the content of those parts, thus becoming unable to control their thinking as well as their behavior, making it challenging to take a macroscopic view of things. For example, a perfectionist will not be able to tolerate the marks or imperfections on an object, and thus will not be

Keywords: Obsessive-compulsive disorder (OCD) · exposure and response prevention (ERP) therapy · excoriation disorder · chronic skin-picking · human skin textures · cognitive

1 Introduction

Obsessive-Compulsive Disorder (OCD) presents a complex and heterogeneous mental health challenge, requiring in-depth investigation to comprehend its underlying mechanisms fully. This paper ventures into the unexplored territory of OCD and its potential link to human skin textures, with a specific focus on Excoriation Disorder, commonly known as chronic skin-picking. By conducting a comprehensive analysis of existing literature, this study seeks to elucidate the cognitive aspects, memory impairments, and potential neurobiological factors contributing to this unique association. The theoretical framework draws upon Exposure and Response Prevention (ERP) therapy, which encourages patients to confront their fears and obsessions. An innovative approach utilizes skin texture immersion as a simulated ERP, exposing individuals to imperfections and targeting perfectionistic tendencies associated with OCD.

The exploration of OCD demands an understanding of its clinical manifestations, particularly compulsive repetitive movements that often involve the skin. Within the realm of OCD, memory disorders lead individuals to hyperfocus on event details, resulting in behaviors such as constantly enlarging objects, which, in turn, can influence memory function. A noteworthy aspect is the lasting traces of skin texture left on objects during repeated exposures, potentially offering insights into the connection between OCD behaviors and the skin itself. This background serves as a crucial foundation for comprehending the complexities of OCD and its manifestations related to human skin textures.

The motivation driving this study arises from the limitations encountered in existing approaches to understanding OCD fully. Traditional methodologies have faced challenges in exploring the intricate nuances of the disorder, necessitating innovative avenues for investigation. Focusing on human skin textures, particularly in the context of Excoriation Disorder, this paper aims to offer a fresh perspective on OCD analysis. The juxtaposition of this novel approach with conventional methods illuminates the potential benefits of using human skin textures to enhance our comprehension of OCD and its related behaviors. Additionally, this study addresses the shortcomings of previous approaches and underscores the necessity for innovative methodologies to unlock the intricate relationship between OCD and human skin textures.

Central to this thesis are the main ideas and methods, prominently featuring skin-based motivation. The study proposes the use of magnified skin texture details to simulate ERP therapy for OCD. This simulation endeavors to expose patients to imperfections, fostering a gradual reduction in perfectionistic tendencies. Amplifying skin texture to create an immersive environment for ERP therapy serves as a medium for personal interaction and bridges the gap between the individual and their surroundings. Data collection for this research encompasses diverse approaches, such as questionnaires and real-world observations,

Understanding Obsessive-Compulsive Disorder Through Human Skin Textures

Yazhen Zhu[1,2,3], Jian Chen[2,3,4], Yuwei Sun[5], and Wei Wang[2,3,6(✉)]

[1] Royal College of Art, London SW7 2EU, UK
[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
chenj589@mail2.sysu.edu.cn, ehomewang@ieee.org
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen, MSU-BITUniversity, Shenzhen 518172, Guangdong, China
[4] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518000, China
[5] Columbia University, New York, NY 10027, USA
ys3371@tc.columbia.edu
[6] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. Obsessive-Compulsive Disorder (OCD) is a complex and heterogeneous mental health condition that challenges our understanding of its underlying mechanisms. This paper explores the potential connection between OCD and human skin textures, particularly Excoriation Disorder (chronic skin-picking), through a comprehensive analysis of existing literature. Investigating cognitive aspects, memory impairments, and potential neurobiological factors contributing to this association, the study also examines the role of human-computer interaction (HCI) in data analysis and treatment approaches, with a focus on skin texture-related aspects. Additionally, the thesis delves into two entry points for understanding OCD through human skin texture. OCD's clinical manifestations involve compulsive repetitive movements, where memory disorders lead individuals to hyperfocus on event details, causing behaviors of constantly enlarging objects, leaving traces of skin texture on them. Drawing inspiration from Exposure and Response Prevention (ERP) therapy, the paper proposes magnifying skin texture details to simulate ERP, exposing patients to imperfections and reducing perfectionistic tendencies. Secondly, related OCD symptoms, like compulsive skin peeling, leave specific skin marks, providing potential clues for identifying OCD characteristics and patterns. This innovative approach offers valuable insights into the complexities of OCD, highlighting the significance of human skin texture in understanding and treating the disorder. By integrating cognitive and neurobiological aspects, this study provides a comprehensive perspective on the intriguing relationship between OCD and human skin textures, contributing to advancements in OCD research and intervention.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 85–95, 2024.
https://doi.org/10.1007/978-3-031-65126-7_9

11. Zakaria, C., Yilmaz, G., Mammen, P.M., Chee, M., Shenoy, P., Balan, R.: Sleepmore: inferring sleep duration at scale via multi-device wifi sensing. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **6**(4), 1–32 (2023)
12. Yu, B., et al.: Wifi-sleep: sleep stage monitoring using commodity wi-fi devices. IEEE Internet Things J. **8**(18), 13900–13913 (2021)
13. Yang, X., Yu, X., Xie, L., Xue, H., Zhou, M., Jiang, Q.: Sleep apnea monitoring system based on commodity wifi devices. Comput. Mater. Cont **2**(69), 2793–2806 (2021)
14. Liu, W., Chang, S., Liu, Y., Zhang, H.: Wi-PSG: detecting rhythmic movement disorder using cots wifi. IEEE Internet Things J. **8**(6), 4681–4696 (2020)
15. Ridolfi, M., Kaya, A., Berkvens, R., Weyn, M., Joseph, W., Poorter, E.D.: Self-calibration and collaborative localization for UWB positioning systems: a survey and future research directions. ACM Comput. Surv. (CSUR) **54**(4), 1–27 (2021)
16. Li, S., Wang, Z., Zhang, F., Jin, B.: Fine-grained respiration monitoring during overnight sleep using IR-UWB radar. In: Hara, T., Yamaguchi, H. (eds.) International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, pp. 84–101. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-94822-1_5
17. Han, W., Dai, S., Yuce, M.R.: Real-time contactless respiration monitoring from a radar sensor using image processing method. IEEE Sens. J. **22**(19), 19020–19029 (2022)
18. Kwon, H.B., et al.: Attention-based LSTM for non-contact sleep stage classification using IR-UWB radar. IEEE J. Biomed. Health Inform. **25**(10), 3844–3853 (2021)
19. Atlas, D., Srivastava, R., Sekhon, R.S.: Doppler radar characteristics of precipitation at vertical incidence. Rev. Geophys. **11**(1), 1–35 (1973)
20. Baboli, M., Singh, A., Soll, B., Boric-Lubecke, O., Lubecke, V.M.: Wireless sleep apnea detection using continuous wave quadrature doppler radar. IEEE Sens. J. **20**(1), 538–545 (2019)
21. Islam, S.M.M., Lubecke, V.M.: Sleep posture recognition with a dual-frequency microwave doppler radar and machine learning classifiers. IEEE Sensors Lett. **6**(3), 1–4 (2022)
22. Rahman, T., et al.: DoppleSleep: a contactless unobtrusive sleep sensing system using short-range doppler radar. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 39–50 (2015)
23. Chen, L., Xiong, J., Chen, X., Lee, S.I., Zhang, D., Yan, T., Fang, D.: Lungtrack: towards contactless and zero dead-zone respiration monitoring with commodity RFIDS. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(3), 1–22 (2019)
24. Liu, C., Xiong, J., Cai, L., Feng, L., Chen, X., Fang, D.: Beyond respiration: contactless sleep sound-activity recognition using RF signals. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(3), 1–22 (2019)
25. Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S., Bianchi, M.T.: Learning sleep stages from radio signals: a conditional adversarial architecture. In: International Conference on Machine Learning, pp. 4100–4109. PMLR (2017)

In summary, each technology has its own merits and considerations. Contactless sensing also leaves much to be desired, such as greater noise immunity to the varying light conditions of different indoor environments. At the same time, because contactless sensing can capture more information, it faces more serious privacy issues. The choice depends on specific requirements, budget constraints, and the desired level of monitoring accuracy. Besides, more research can focus on how to combine these two methods for better performance and less cost.

5 Conclusion

In this work, we review the existing sleep monitoring methods based on Wifi sensors and wireless sensors. Then we make a comparative analysis between these two methods for a better illustration of wireless sensors used in the field of sleep monitoring. Through the summary of the existing methods, we can better find the direction for the follow-up research. However, in addition to wifi sensors and radar, acoustic and optical sensors are also beginning to be used in this field. Therefore, it is our future work to further summarize and analyze the advantages and disadvantages of these methods.

References

1. Perez-Pozuelo, I., et al.: The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ Digital Med. **3**(1), 42 (2020)
2. Bertisch, S.M., et al.: Insomnia with objective short sleep duration and risk of incident cardiovascular disease and all-cause mortality: sleep heart health study. Sleep 41(6), zsy047 (2018)
3. Zhou, Q., Zhang, M., Hu, D.: Dose-response association between sleep duration and obesity risk: a systematic review and meta-analysis of prospective cohort studies. Sleep Breathing **23**, 1035–1045 (2019)
4. Palagini, L., Hertenstein, E., Riemann, D., Nissen, C.: Sleep, insomnia and mental health. J. Sleep Res. **31**(4), e13628 (2022)
5. Rundo, J.V., Downey, R., III.: Polysomnography. Handb. Clin. Neurol. **160**, 381–392 (2019)
6. Engstrøm, M., Rugland, E., Heier, M.S.: Polysomnography (PSG) for studying sleep disorders. Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke **133**(1), 58–62 (2013)
7. Rottenberg, F., Nguyen, T.-H., Dricot, J.-M., Horlin, F., Louveaux, J.: CSI-based versus RSS-based secret-key generation under correlated eavesdropping. IEEE Trans. Commun. **69**(3), 1868–1881 (2020)
8. Chen, Z., Zhang, L., Jiang, C., Cao, Z., Cui, W.: Wifi CSI based passive human activity recognition using attention based BLSTM. IEEE Trans. Mob. Comput. **18**(11), 2714–2724 (2018)
9. Gui, L., Ma, C., Sheng, B., Guo, Z., Cai, J., Xiao, F.: In-home monitoring sleep turnover activities and breath rate via wifi signals. IEEE Syst. J. **17**, 2355–2365 (2022)
10. Liu, J., Chen, Y., Wang, Y., Chen, X., Cheng, J., Yang, J.: Monitoring vital signs and postures during sleep using wifi signals. IEEE Internet Things J. **5**(3), 2071–2084 (2018)

Doppler radar is widely used in the field of sleep detection due to its excellent ability to measure target displacement remotely. Doppler radar can capture the information of chest displacement due to respiration or heartbeat through the transmitted microwave signals and analyze it through the Doppler effect [19]. A contactless system named PRMS using quadrature microwave doppler radar to monitor sleep apnea events in real time. The system contains a real-time actigraphy and sleep apnea detection algorithm [20]. A novel sleep posture recognition technique is proposed, which employs classifiers that are amenable to optimization through Bayesian hyperparameter tuning. These classifiers operate on data from a dual-frequency monostatic continuous-wave radar system [21]. DopplerSleep, a contact sleep sensing system, uses a single Doppler sensor to track sleep quality. DopplerSleep can monitor both body movements and tiny chest and heart movements, and the system has been experimentally validated to perform well on sleep stage classification tasks [22].

RF signals are widely used for contactless motion and vital signs monitoring in the field of sleep monitoring. Radio Frequency Identification (RFID) is a contactless communication technology that enables two-way data exchange for identification and data transfer using RF signals with flexibility and low cost. A respiration monitoring system with RFID sensors called LungTrack is proposed to achieve dual objective monitoring with an accuracy of above 93% for two targets at a distance of 10 cm at least [23]. TagSleep is a sleep posture recognition system using the concept of two-layer sensing with RFID sensors [24]. A model combining a convolutional network and recurrent neural network is trained on the RF-measured sleep dataset with an adversarial training regime [25].

4 Comparative Analysis

WiFi sensors and other wireless sensors, as non-interference devices, offer both advantages and disadvantages in sleep monitoring. Figure 1 shows a comparison between these two methods. WiFi sensors typically utilize wireless signals and receivers to track variables such as breathing, body movement, and sleeping positions. These sensors analyze movement patterns and breathing rates by observing changes in WiFi signals. They are cost-effective and easy to deploy, but privacy concerns may arise.

On the other hand, radar technology emits high-frequency pulse signals and measures the time it takes for the signals to bounce back. This enables accurate positioning and tracking of objects, including monitoring human movements and breathing patterns during sleep. Radar provides precise distance and position measurements, boasting high accuracy and reliability. However, radar requires specialized hardware and incurs higher costs. Both UWB and doppler radars described previously are capable of real-time sleep monitoring with a high degree of accuracy, but there is the problem of higher equipment costs and more demanding deployment conditions during equipment placement.

While RFID technology offers advantages like low power consumption and affordability, it may have limitations when it comes to more detailed sleep analysis and breathing monitoring.

wifi sensors are used for obstructive sleep apnea (OSA) detection and rhythmic movement disorder (RMD) detection. An intelligent apnea monitoring system can utilizing linear fitting and wavelet transform eliminate the phase error of CSI. The system uses commodity wifi, which is better able to eliminate interference from changes in sleeping posture [13]. A sleep monitoring system named Wi-PSG is proposed to utilize CSI from Wifi infrastructures for RMD-related movement detection, which can achieve an accuracy of above 92% for different RMD movement classifications [14].

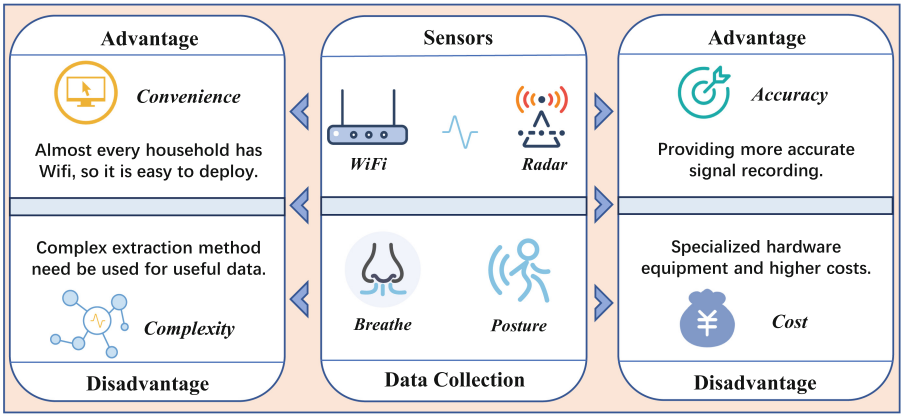


Fig. 1. Comparison between WiFi sensors and radars.

3 Sleep Detection Based on Wireless Sensor

Wireless radars are the most widely used sensors in sleep detection based on wireless sensors. Systems with wireless sensors are usually used for vital signs detection during sleep and sleep quality detection. The main sensors used in these systems are Ultrawideband (UWB) radar, Doppler radar, and Radio Frequency (RF) sensors.

UWB radar is commonly utilized for precise localization, employing low energy levels for short-range and high-bandwidth communications across the radio spectrum [15]. The required sleep information can be extracted by the UWB radar sensor penetrating the clothes and quilt. A fine-grained prototype for overnight respiration monitoring is proposed by exploiting the complementarity between the amplitude and phase of the radar signal [16]. Four respiration patterns are recognized during overnight sleep in this method. Another image processing method converts the raw signals collected by the UWB radar into a 2-D heatmap image and then an image-processing algorithm is used to capture respiratory information for respiratory motion measure [17]. An attention-based LSTM model is proposed to use the vital signs detected remotely by an impulse-radio UWB radar for sleep stage classification [18].

[6], However, the recording of PSG always needs expensive equipment and keep lots of contact with the subjects' body which bring discomfort. These drawbacks make it unsuitable for daily life sleep monitoring.

With the development of information techniques, more and more wireless sensors are used for sleep monitoring. There are already a lot of wearable devices used for sleep monitoring, but they also face resistance because of the discomfort brought to subjects and instability during sleep. Contactless sensors can effectively address the problem that invasive sensors bring natural sleep difficulties. There are various contactless sensors used in sleep monitoring now. The main of them are wireless sensors. Wifi sensor is also a kind of wireless sensor but since it has received more attention than other wireless sensors, it is put in a separate category.

Since wireless sensors are now widely used in sleep monitoring and have shown great potential, it is meaningful to review sleep monitoring research based on wifi sensors and wireless sensors. This can help develop contactless devices to achieve stable, safe, and non-contact sleep detection. In this work, we will first review the main sleep detection methods based on wifi sensors and wireless sensors respectively, and then a comparative analysis is made to summarize the difference between wifi sensors and wireless used in sleep monitoring. Finally, we provide a conclusion of our work.

2 Sleep Detection Based on Wifi Sensor

Wifi-based sleep monitoring activities are generally carried out through high precision indoor positioning, and the commonly used methods include Received Signal Strength (RSS) and Received Signal Strength (CSI) [7]. With the development of the technology, the CSI technique has demonstrated greater stability and accuracy and has become the more mainstream method nowadays. While using wifi sensors for sleep monitoring, CSI can be used to capture the effect of sleep activity contained by the Wifi signals [8].

Existing methods that use Wifi sensors to monitor sleep quality include heart rate monitoring and respiration monitoring [9]. A method is proposed to track the breathing rate and heart rate during sleep with Wifi [10]. They exploit to utilize the fine-grained channel information of existing Wifi networks to extract the minute movements that come with breathing and heartbeats. Wifi network activity is also used in a sleep-tracking approach called SleepMore which utilizes machine learning methods [11]. SleepMore constructs a semi-personalized random forest model to make a classification of the network activity behavior and the results are divided into sleep and awake states in minute dimensions. The experimental results show that SleepMore achieves an indistinguishable result with the Oura ring baseline within a 5% uncertainty rate.

Wifi sensors are also used for sleep stage classification and sleep-related disorders detection. An advanced signal processing and fusion method is proposed to extract accurate respiration and body movement for four-stage sleep classification, which achieves an accuracy of 81.1% [12]. In disorders monitoring,

Review of Sleep Monitoring Research Based on Wireless Sensor

Yuzhu Hu[1,2,3], Jian Chen[1,2,3], Shen Zhao[1(✉)], Kexin Tan[2], Kuai Yu[2], and Wei Wang[2,3,4]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518000, China
{huyzh27,chenj589}@mail2.sysu.edu.cn, z-s-06@163.com
[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{1120200259,1120200296}@smbu.edu.cn, ehomewang@ieee.org
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen MSU-BITUniversity, Shenzhen 518172, Guangdong, China
[4] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. Since sleep quality is crucial to human health, sleep monitoring has become a hot spot in the field of smart healthcare. Previous methods depend on polysomnography and wearable devices need immediate contact with the subject, which brings discomfort. Contactless sensors can address this issue. The most common contactless sensors used in sleep monitoring are wireless sensors (including radar and WiFi). To clarify the research in this area, we summarized the existing sleep monitoring methods based on WiFi sensors and wireless radar and made a comparison. The conclusion shows that the two kinds of methods have advantages and disadvantages, so the development of complementary methods is very promising for sleep monitoring.

Keywords: Sleep monitoring · contactless sensors · wireless sensing

1 Introduction

Sleep is one of the most important basic life activities of human beings, and it is also an important basis for maintaining physical and mental health [1]. Chronic poor sleep has also been linked to cardiovascular disease, obesity, and even some mental health problems [2–4]. Therefore, sleep monitoring is important for health status monitoring and is now become a hot topic for research.

Polysomnography (PSG) is the most widely used tool to monitor sleep, and it is regarded as the gold standard to detect sleep-related breathing disorders [5]. PSG can provide comprehensive information on sleep stages on the basis of Electroencephalography (EEG) activity, eye movements, and muscular tension

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 79–84, 2024.
https://doi.org/10.1007/978-3-031-65126-7_8

15. Vos, T., et al.: Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet **396**(10258), 1204–1222 (2020)
16. Walker, E.R., McGee, R.E., Druss, B.G.: Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry **72**(4), 334–341 (2015)

developing countries (Fig. 2); anxiety contributes to the development of depression and must be taken into account as well.

Considering all above we can suggest the authorities to take more measures to ease the burden and stress of the depraved people. As other studies showed [4,6,8], low-income group are at the higher risk of getting depression and having worse health condition in general [3,5], so, some government financial help is better be provided (subsidiaries, money allowance, etc.).

Acknowledgment. This work is supported by the Shenzhen Science and Technology Innovation Commission (Stabilisation Support Programme).

References

1. Chesney, E., Goodwin, G.M., Fazel, S.: Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry **13**(2), 153–160 (2014)
2. Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J., Penninx, B.W.: Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses. Am. J. Psychiatry **171**(4), 453–462 (2014)
3. Diener, E., Biswas-Diener, R.: Will money increase subjective well-being? Soc. Indic. Res. **57**, 119–169 (2002)
4. Dwyer, R.J., Dunn, E.W.: Wealth redistribution promotes happiness. Proc. Natl. Acad. Sci. **119**(46), e2211123119 (2022)
5. Headey, B., Muffels, R., Wooden, M.: Money does not buy happiness: or does it? a reassessment based on the combined effects of wealth, income and consumption. Soc. Indic. Res. **87**, 65–82 (2008)
6. Kahneman, D., Deaton, A.: High income improves evaluation of life but not emotional well-being. Proc. Natl. Acad. Sci. **107**(38), 16489–16493 (2010)
7. Kartaev, P.S.: How to teach econometrics to economists: bachelor level..... 72 macroeconomic policy. Sci. Res. Fac. Econ. Electron. J. **11**(2), 72–90 (2019)
8. Killingsworth, M.A.: Experienced well-being rises with income, even above $75,000 per year. Proc. Natl. Acad. Sci. **118**(4), e2016976118 (2021)
9. Patel, V., et al.: Addressing the burden of mental, neurological, and substance use disorders: key messages from disease control priorities. The Lancet **387**(10028), 1672–1685 (2016)
10. Pearce, M., et al.: Association between physical activity and risk of depression: a systematic review and meta-analysis. JAMA Psychiatry **79**, 550–559 (2022)
11. Reger, M.A., Stanley, I.H., Joiner, T.E.: Suicide mortality and coronavirus disease 2019-a perfect storm? JAMA Psychiatry **77**(11), 1093–1094 (2020)
12. Son, J., Shin, J.: Bimodal effects of sunlight on major depressive disorder. Compr. Psychiatry **108**, 152232 (2021)
13. Stock, J.H., Watson, M.W.: Introduction to Econometrics, vol. 104. Addison Wesley Boston (2003)
14. Viswanathan, M., et al.: Screening for depression and suicide risk in children and adolescents: updated evidence report and systematic review for the us preventive services task force. JAMA **328**(15), 1543–1556 (2022)

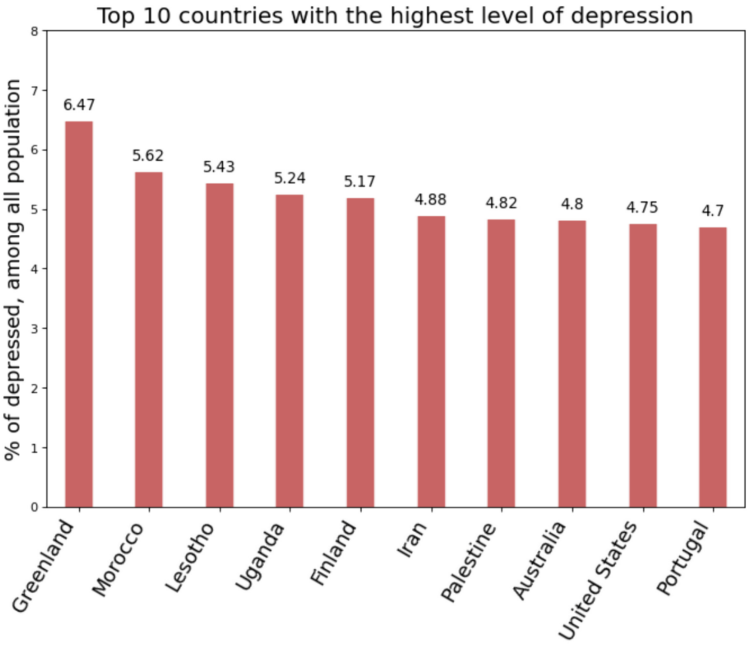


Fig. 2. Top 10 countries with the highest level of depression

As can be noticed, top ten includes mainly developing countries where GDP per capita quite small. The exceptions are Finland, Australia, United States, Portugal and Greenland as a special region of Denmark, where the GDP per capita is medium or higher. In case with Finland and Greenland, such higher level of depression could be explained by two factors: 1) isolated and low populated communities, as can be seen from the depression model, the population size is significant factor; 2) the lack of sunny days, what negatively effects on mood and emotional conditions [12]. As for other countries, the further deep analysis is required.

4 Discussion

This large-scale study based on worldwide panel data about depression showed that people who live in countries with low GDP per capita are more vulnerable to depression. We find that the relationship between depression and GDP per capita is strongly negative, and because of analyses of huge massive of date, the results are universal. At the same time, the connection between depression and anxiety disorders is strongly positive, thus, the following conclusions could be made: in countries with lower GDP per capita, more people tend to suffer from depression. Actually, this fact can be proved even statistically: the majority of the countries in the list of top 10 countries with high level of depression are

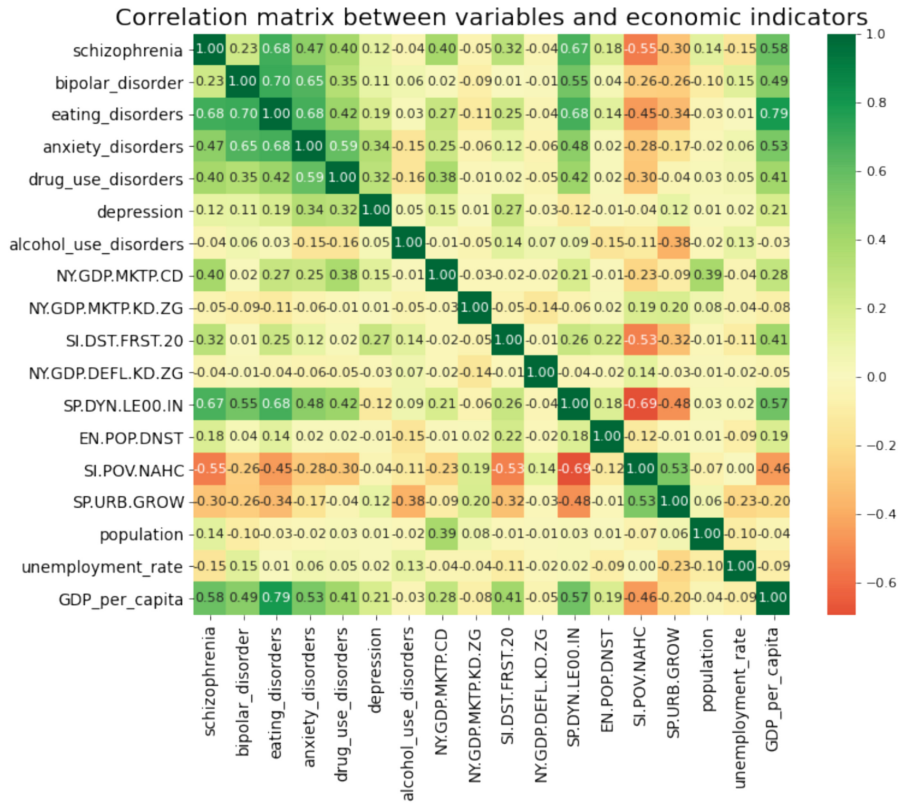


Fig. 1. Correlation matrix between variables and economic indicators

- GDP per capita - all things being equal, with an increase in GDP by one dollar, the number of people suffering from depression decreases by 2,74*e-6%
- Income share held by lowest 20% - all things being equal, with an increase in Income share held by lowest 20% by one dollar, the number of people suffering from depression decreases by 0,024%
- Life expectancy at birth, total - all things being equal, with an increase in life expectancy at birth, by one year, the number of people suffering from depression decreases by 0,013%
- Population - all things being equal, with an increase in population by one people, the number of people suffering from depression decreases by 1,05*e-9%
- Anxiety disorders - all things being equal, with an increase in anxiety disorders by one percent, the number of people suffering from depression increases by 0,35%

Now countries that have the highest rates of depression are shown in Fig. 2.

– SI.POV.NAHC - Poverty headcount ratio at national poverty lines (% of population)
– SP.URB.GROW - Urban population growth (annual %)
– Unemployment - Unemployment rate, (% of work force)
– GDP_PER_CAPITA - GDP per capita, (current US$)

2.3 Panel Study

In order to avoid omitted variable bias, we took regressors from different spheres [7,13]: pure economic, social and urban. We also have added the variable of control - Anxiety, as, by all means, anxiety disorders influence on the development of depression and other mental disorders. We conducted all measures using special econometric program Gretl.

3 Results

We calculated the correlation between all mental disorders and economic indicators as Fig. 1.

At the same time, we got the following depression model as Tabel 1.

Table 1. Depression Model

	Coefficient	St. error	t-statistics	p-value	
const	325,388	0,707587	4,599	¡0,0001	***
anxiety disorders	0,354180	0,145535	2,434	0,0168	**
NYGDPMKTPCD	0,000000	0,000000	4,352	¡0,0001	***
NYGDPMKTPKDZG	0,00112238	0,00100377	1,118	0,2663	
SIDSTFRST20	−0,0239324	0,00750013	−3,191	0,0019	***
NYGDPDEFLKDZG	−0,000230614	0,000412541	−0,5590	0,5775	
SPDYNLE00IN	−0,0134092	0,00500963	−2,677	0,0088	***
ENPOPDNST	0,000163380	0,000295864	0,5522	0,5821	
SIPOVNAHC	−0,00107081	0,00145502	−0,7359	0,4636	
SPURBGROW	−0,00724004	0,00900659	−0,8039	0,4235	
population	−1,04580e−09	1.73E−05	−6,051	< 0,0001	***
unemployment_rate	−0,00358409	0,00199508	−1,796	0,0756	*
GDP_PER_CAPITA	−2,74364e−06	9.21E−02	−2,978	0,0037	***

The LSDV R-square for this model is 0,9956, '*' means that variable is significant on 10%, '**' - 5%, and '***' - 1%. Therefore, we could interpret four variables of interest (on 5%):

World Health Association, around 280 million of people worldwide are suffering from depression, moreover, the World Health Organization assumes that 5% of men and 9% of female experience depressive disorders in their lifetime [10, 15]. Depression can lead to the development of other illnesses what effect on premature mortality [1,2,16] and even increase the suicide rates [9,11,14], that is why it is crucial for authorities to be aware of development of such illnesses. The innovation of this work is that it includes factors and figures from different spheres and examine their impact on the development of depression and other mental disorders. This allows us to broaden our thinking and to make more clear judgments [7,13]. Particularly, in addition to social-economic indicators, we also added urban population growth in our list of economic indicators, what allows to see the big picture. This article is aimed to determine how the main economic indicators are connected with mental disorders. After establishing the relationships, it will be possible to judge whether the country at the risk of mass depression. We believe that with the help of our research local authorities will be able to identify the upcoming health threats more effectively, and, what is the key point, much earlier, thus, many human lives would be improved or even saved.

2 Methods

This is a panel study which includes data from 196 countries throughout 27 years. In our research we mainly used econometrics and ordinary least square (OLS) analysis to make proper models. All implemented models have passed the Ramsey Test, the check for heteroscedasticity and multicollinearity, thus, all described models are trustful. Besides, in case with the depression analysis, the Fixed Effects model was used due to take into account each country peculiarity [7,13].

2.1 Dependent Variables

In addition to Depression, we also considered the following types of mental diseases: Schizophrenia, Bipolar disorder, Eating disorders, Anxiety disorders, Drug use disorders, Alcohol use disorders. All variables are examined as % of all population.

2.2 Economic Indicators

For each variable we make an econometric model with the following regressors:

- NY.GDP.MKTP.CD - GDP (current US$)
- NY.GDP.MKTP.KD.ZG - GDP growth (annual %)
- SI.DST.FRST.20 - Income share held by lowest 20%
- NY.GDP.DEFL.KD.ZG - Inflation, GDP deflator (annual %)
- SP.DYN.LE00.IN - Life expectancy at birth, total (years)
- EN.POP.DNST - Population density (people per sq. km of land area)

Identification of Economic Factors for Mass Depression Based on Panel Study and Machine Learning

Iaroslava Pravolamskaya[1,2,3,4], Jian Chen[3,4,5], and Wei Wang[3,4,6(✉)]

[1] Faculty of Economics, Shenzhen MSU-BIT University, Shenzhen 518172, China
[2] Faculty of Economics, Lomonosov Moscow State University, Moscow 119991, Russia
[3] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
[4] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen, MSU-BITUniversity, Shenzhen 518172, Guangdong, China
[5] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518000, China
chenj589@mail2.sysu.edu.cn
[6] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
ehomewang@ieee.org

Abstract. Panel study and machine learning are important tools for analyzing various aspects of the economy. They allow researchers to study the dynamics of changes in different economic indicators, such as GDP, inflation, unemployment, etc. In addition, these tools can be used to determine causal relationships between social, economic and psychological factors what can allow us to predict the development of the economy and changes in people's life in the future. However, previous works in this sphere studied the connections between income and happiness, not taking into account the relationships between economic indicators and mental disorders. This article is aimed to analyze the relationship between economic factors and the level of mass depression based on a panel study and machine learning methods. Experimental results based on panel study and machine learning demonstrate effectiveness of our proposed econometric model.

Keywords: Panel Study Machine Learning Depression Identification Economic Factors Econometrics Models

1 Introduction

The increasing number of people suffering from depression and other mental diseases is one of the most challenging issues in the 21 century. According to

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 72–78, 2024.
https://doi.org/10.1007/978-3-031-65126-7_7

14. Koyama, K., Hoshikawa, H., Kojima, G.: Eddy current nondestructive testing for carbon fiber-reinforced composites. J. Press. Vessel. Technol. **135**(4), 041501 (2013)
15. Kostopoulos, V., Vavouliotis, A., Karapappas, P., Tsotra, P., Paipetis, A.: Damage monitoring of carbon fiber reinforced laminates using resistance measurements. Improving sensitivity using carbon nanotube doped epoxy matrix system. J. Intell. Mater. Syst. Struct. **20**(9), 1025–1034 (2009). https://doi.org/10.1177/1045389X08099993

conducted to get the correct data and then compared to draw conclusions. Damage to the carbon/glass blend and fracture of the carbon fibers was observed by using Three Point Bending method. The pictures show that where pressure is applied the upper layers are damaged by shear stresses leading to kinking and the lower layers are damaged mainly in the form of delamination leading to failure.

In conclusion, this study has been designed, experimented and concluded that it is feasible to monitor the electrical conductivity of this hybrid carbon/glass fiber blend and that this composite fiber can also be seen as a self-sensor.

Acknowledgments. This project is supported by the funding of Guangdong Province Key Laboratory of Intelligent Detection in Complex Environment of Aerospace, Land and Sea. (2022KSYS016).

References

1. Maleque, M.A., Salit, M.S.: Materials Selection and Design. SM, Springer, Singapore (2013). https://doi.org/10.1007/978-981-4560-38-2
2. Jalalvand, M., Czél, G., Wisnom, M.R.: Damage analysis of pseudo-ductile thin-ply UD hybrid composites – A new analytical method. Compos. A Appl. Sci. Manuf. **69**, 83–93 (2015). https://doi.org/10.1016/j.compositesa.2014.11.006
3. Sauer, M.: Composites Market Report 2019—The Global CF-und CC-Market 2019: Market Developments, Trends, Outlook and Challenges. Composites United eV, Berlin, Deutschland (2019)
4. Czél, G., Wisnom, M.R.: Demonstration of pseudo-ductility in high performance glass/epoxy composites by hybridization with thin-ply carbon prepreg. Compos. A Appl. Sci. Manuf. **52**, 23–30 (2013)
5. David-West, O., et al.: A review of structural health monitoring techniques as applied to composite structures. Struct. Durability Health Monit. **11**(2), 91–147 (2017)
6. Rev, T., et al.: A simple and robust approach for visual overload indication-UD thin-ply hybrid composite sensors. Compos. A Appl. Sci. Manuf. **121**, 376–385 (2019)
7. Chapuis, B.: Introduction to structural health monitoring. In: Chapuis, B., Sjerve, E. (eds.) Sensors, Algorithms and Applications for Structural Health Monitoring. IC, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-69233-3_1
8. Vavouliotis, A., Paipetis, A., Kostopoulos, V.: On the fatigue life prediction of CFRP laminates using the electrical resistance change method. Compos. Sci. Technol. **1**(5), 630–642 (2011)
9. Smith, R.: Composite defects and their detection. Mater. Sci. Eng. **3**(1), 103–143 (2009)
10. Gregor Trtnik, M.G.: Recent advances of ultrasonic testing of cement based materials at early ages. Ultrasonics **54**, 66–75 (2013)
11. Song, S., Jing, J., Cheng, W.: Online monitoring system for macro-fatigue characteristics of glass fiber composite materials based on machine vision. IEEE Trans. Instrum. Meas. **71**, 1–12 (2022)
12. Manjunatha, P.A.: Vision-based and data-driven analytical and experimental studies into condition assessment and change detection of evolving civil, mechanical and aerospace infrastructures. Doctoral dissertation, University of Southern California (2022)
13. Bayraktar, E., Antolovich, S.D., Bathias, C.: New developments in non-destructive controls of the composite materials and applications in manufacturing engineering. J. Mater. Process. Technol. **206**(1–3), 30–44 (2008). https://doi.org/10.1016/j.jmatprotec.2007.12.001

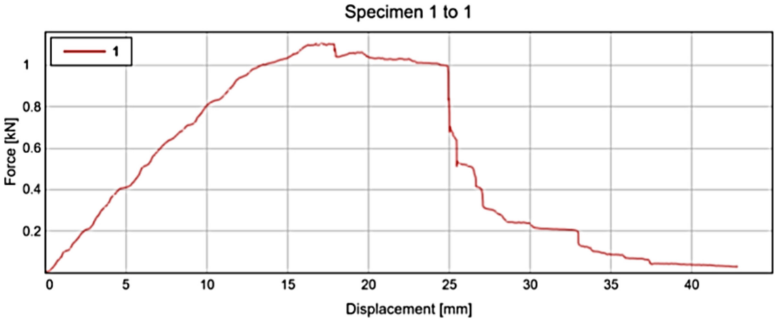


Fig. 11. 10 mm experimental group displacement and Force curve

to the phenomenon of pseudo-stretchability. Analysis of the sample damage showed that shear damage dominated at the upper end of the sample, while delamination dominated from the middle to the lower plies, shown as Fig. 12. However, the main change in resistance in this test was due to the fracture of the thin carbon fibers, which was mainly due to tensile stresses, while the delamination of the lower plies was mainly due to shear stresses, so the design of this test was reasonable.

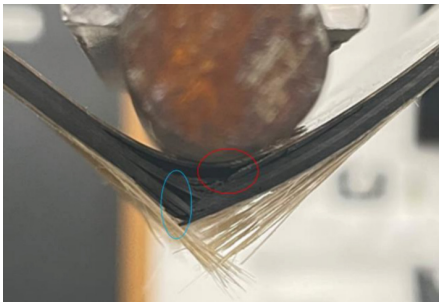


Fig. 12. Injury patterns under three -point bending

6 Conclusion

In this project, a hybrid thin-layer carbon/glass fiber self-sensing method is proposed. It is innovative in that it changes the traditional case of applying the carbon fibers directly to the object to be sensed. Also, by using an S -shape instead of the traditional direct strip, it allows for greater coverage and a larger area to be monitored than just partial detection, while its more holistic nature makes it more effective for monitoring a whole plane rather than monitoring a broken location, and also has a greater improvement in monitoring the effects of certain unseen damage. Regarding the experiment, this experiment uses the controlled variable method to create differences for different variables. By designing groups of different widths as well as different styles, several experiments were

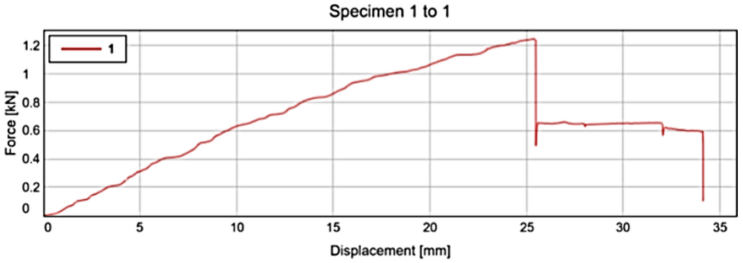


Fig. 9. 5 mm experimental group displacement and Force curve

in the 25–30 mm zone, the resistance begins to rise, indicating that the carbon fiber body is further destroyed in this zone, when in the 30–35 zone, the carbon fiber is completely destroyed and the resistance rises to 4000 Ω (Fig. 8). When the carbon fibers are completely destroyed, the loading force is removed and the fibers spring back, at this time some of the fibers reduce in resistance because the stress is reunited (Fig. 9).

10 mm Bending Test

The 10 mm three-point bending test is also primarily a comparison with 5 mm, observing the change in resistance of two different widths of carbon fiber to determine which is more appropriate.

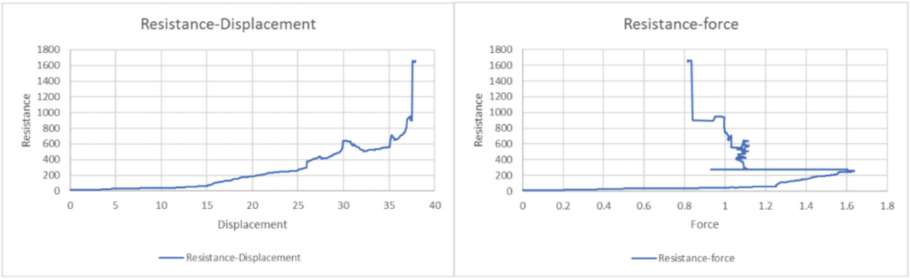


Fig. 10. 10 mm experimental group resistance displacement curve(left) and 10 mm experimental group resistance Force curve(right)

By comparing the two sets of plots, it can be found that the 5 mm images of resistance and Force are relatively similar to the 10 mm images on the three -point bending method, but on the resistance displacement curve, it is obvious that the rising trend of the 10 mm curve is smoother, so after the comparison, it is more recommended to use a 10 mm wide thin layerof carbon fiber as a self-sensor (Figs. 10 and 11).

Damage Mode Analysis

In this section, the damage pattern of the experimental product and the image in the above figure will be analyzed in detail, as the damage to the sample occurred gradually over the course of the test and this fiber hybridization slowed down the catastrophic rate and so led

Table 1. Mechanical properties of curing

Properties	Numerical value	Unit
Tg Onset(DMA)	**140**	°C
Tensile Strength	**645**	MPa
Compressive Strength	**515**	MPa
Flexural Strength	**882**	MPa
Flexural Modulus	**60.1**	GPa
Interlaminar Shear Strength	**69.8**	MPa
Tg Peak(DMA)	**148**	°C

of plain carbon fiber strips alone. As the fiber orientation was also considered in this carbon fiber experiment to affect the magnitude of the current, a unidirectional thin layer of carbon fiber was used in this case so that the consistency of the current could be maintained throughout.

In the three-point bending test, since the three -point bending test causes large shear stresses, data were collected from the start to sample failure and finally the changes in resistance and the reasons for these changes were analyzed in conjunction with the changes in the curves.

5 mm Bending Test

In this section the experimental data on the 5 mm three-point bending method is described. Unlike the above, as this design is a hybrid design, the standard T700 carbon fiber bending performance criteria above can only be used as a reference value, so according to the experimental process, the bending performance is significantly lower compared to T700, only around 800 Mpa, so it is speculated that it is possible that the mixture of glass fiber and thin-layered carbon fiber has affected the bending performance.

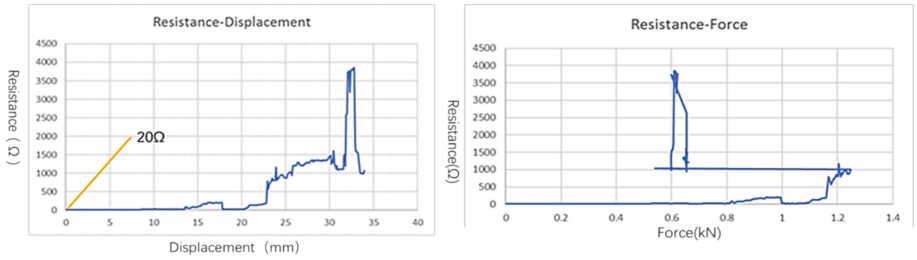


Fig. 8. 5 mm experimental group resistance displacement curve(top) and 5 mm experimental group resistance Force curve(bottom)

According to the data we can see that there is a relatively obvious increase in resistance after the indentation test, as can be seen from the graph, at 25 mm of the experiment is the maximum stress, when the carbon fiber begins to destroy, it can be concluded that

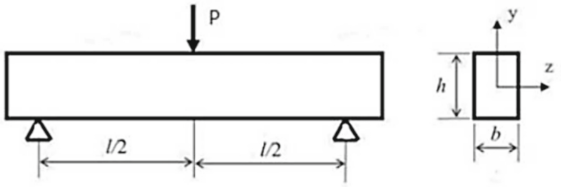


Fig. 7. Three-point bending method model

In the three-point bending test, when viewed from the front, it can simply be seen as a simply supported beam subjected to a concentrated pressure. Three-point bending should theoretically result in a linear distribution of positive stresses along the beam in the cross-sectional area when loaded.

$$\sigma = \frac{M}{I_z} y \tag{1}$$

where σ is the stress, M is the moment, Iz is the moment of inertia of the cross-section to the z-axis and y is the distance in the cross-section to the y-axis. The maximum positive stress at the danger point of the beam is:

$$\sigma_{Max} = \frac{M_{Max}}{I_z} y_{Max} \tag{2}$$

For rectangular section specimens:

$$M = \frac{P \times l}{4} \qquad I_z = \frac{bh^3}{12} \tag{3}$$

Substituting Eqs. (3) into (2) yields the new equation

$$\sigma_{bb} = \frac{3P \times l}{2bh^2} \tag{4}$$

where P is the load and L is the span, b is for width, h is for thickness.

In the case of a sample based on this equation, the maximum shear stress is calculated a

$$s\sigma_{bb} = \frac{3P \times l}{2bh^2} = \frac{3 \times 1.5\,KN \times 150\,mm}{2 \times 50\,mm \times 9\,mm^2} = 800\,Mpa \tag{5}$$

According to the Table 1, its standard value is 880 Mpa, However, as this design contains other fibers of different thicknesses or patterns, this data can only be used as a reference value for the main body of the sample, so in principle the maximum acceptable shear stress for this design should be lower than this value.

5 Results

This experiment focused on the fabrication process of the self-sensor, which was designed using an innovative mixture of carbon fiber and thin layers of E glass fiber, and investigated the advantages and differences between this combination and the use

widely used in destructive testing due to its simple construction and the fact that it does not require much manipulation. For this test, the sample is placed on a jig and a multimeter is connected to the two sections of copper to read the resistance data. The movement speed of 7 mm/min is entered into the control of the hydraulic press and the test is started (Fig. 4).



Fig. 4. Three -point bending test.

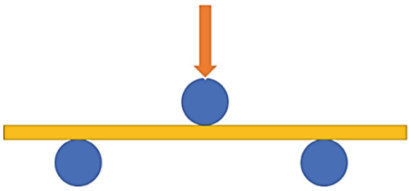


Fig. 5. Injury patterns under three-point bending

The three-point bending process is achieved mainly by applying pressure to the tip, during which the sample undergoes a process of gradual destruction. The following diagram shows the principle of three -point bending (Fig. 5).

During the three-point bending experiment, the sample started to break gradually when it was loaded to a high enough stress. As this sample was a mixed sample, the surface glass fiber started to break when it was loaded to 0.7 KN, the glass fiber broke completely when it was loaded to 1.2 KN, then the load was reduced to 0.6 KN and then the carbon fiber started to break gradually.

The images show that the entire damage process is produced gradually, and based on the experimental images it can be seen that the samples start with damage and end up with damage (Fig. 6).

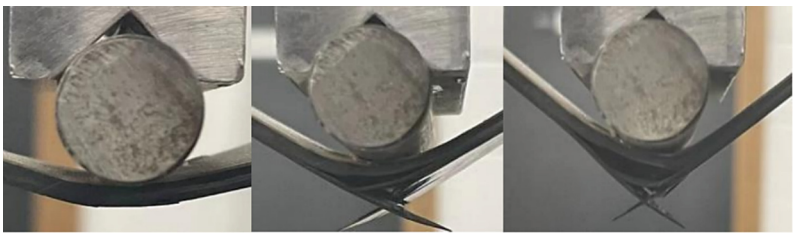


Fig. 6. The process of three-point bending test

Theory of Three-Point Bending Test

When bending deformation occurs in the three-point bending method, the fibers near the bottom elongate and those near the top shorten. According to the planar hypothesis, the fiber state changes gradually from stretching to compression along the height of the cross section from the bottom to the top, then there must be a layer in between where the length of the fiber remains constant, this layer is called the neutral layer (Fig. 7).

known as multi -directional (MD) fibers (Fig. 2). This multi -axial material has better tensile and compressive resistance than uniaxial material, but because it is manufactured at an angle, it is less malleable.

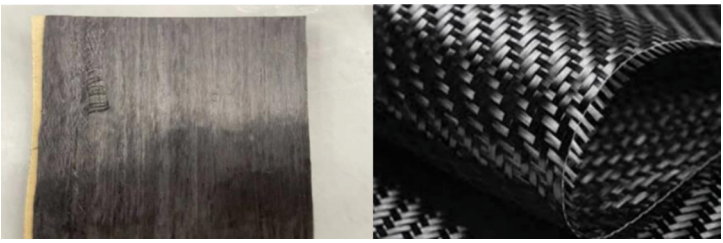


Fig. 2. UD carbon fiber(A), MD carbon fiber(B)

Experiment Test Group

The carbon fibers in the experimental group will be linked to each other, showing S-shaped connections, which then means that the data from the experimental group will affect each other. This control group can be used to see if the resistance will be affected by the occurrence of fiber breaks. Having established that the resistance will change due to fiber breakage, then this experimental group has the advantage that only two electrodes are needed to complete the experiment due to the large area it covers. The main reason for using two different widths of samples was to see the rate of change in resistance by comparing the two sizes of 5 mm and 10 mm. In the graph below, sample 1 is the control group of 10 mm, sample 2 is the control group of 5 mm, sample 3 is the experimental group of 5 mm and sample 4 is the experimental group of 10 mm, shown as Fig. 3.



Fig. 3. Kinds of simple.

Three-Point Bending Test

After the indentation test, the material is tested using the three-point bending method, which is one of the simplest and most effective methods of testing laminates and is

4 Research Methodology

It describes the methods and techniques used in the selection of materials, the preparation process, and the design of the testing process for this project. In particular, first, it will be described in detail what materials are used to prepare the samples, as well as the preparation process and methods. Then, it will be described the test process of the experiment and other equipment used in the experimental process, last, it is going to be described the detail of the whole experiment and the theoretical data will be given, including the theoretical currents and the theoretical stresses generated by the experiment.

Material Selection

In addition, for sample preparation we used T700/XC130 unidirectional prepreg carbon fiber as the sample body and S -Glass/913 and M46JB unidirectional prepreg thin carbon fiber as the sensor. The base material was made from T700 carbon fiber manufactured by Toray of Japan. When selecting the substrate material, it was considered that the main carbon fibers are mainly T300 and T700, both of which contain a large amount of carbon, but the overall performance of T700 is significantly better than that of T300. In the selection of the sensing layer, we chose to use a thin carbon fiber sandwiched between the two glass fibers. As the thin carbon fiber chosen, M46JB, has a similar tensile strength to T700, but obviously the compressive strength of M46JB is weaker than that of T700.T In this experiment, glass fibers were chosen to wrap the thin carbon fiber because, as seen in Meisam's model, there are three different damage modes for composites made from high and low strain materials, so in order for the sensing layer of carbon fibers to break before the glass fibers in the isolation layer, a thin layer of carbon fibers with a lower degree of strain than the glass fibers must be used as the induction material (Fig. 1). (Fotouhi, Jalalvand et al., 2017)

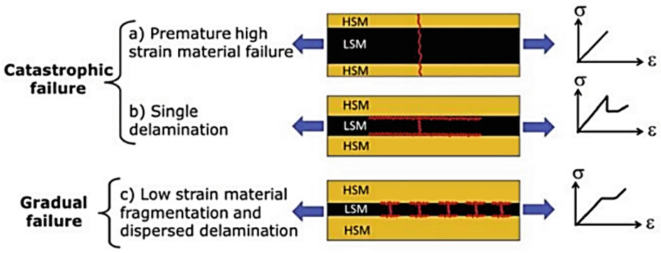


Fig. 1. Possible failure modes in a three layers UD hybrid made from HSM and LSM (red lines show fracture) (a) single crack through the whole specimen, (b) single crack in the LSM followed by instantaneous delamination, and (c) multiple fracture and localised stable pull-out of the LSM

Typically, the basic constituent material of a carbon fiber reinforced material is usually a combination of multiple or unidirectional fiber orientations, rearranged to provide different mechanical properties. A single unidirectional (UD) fiber arrangement, where all the fibers in the resin are aligned in one direction with no voids or breaks. Another type of arrangement is where the fibers are aligned at 0 and 90 degrees, this is

In fields such as aviation and construction, it is important to ensure safety margins, as sudden damage can potentially lead to injury or death as well as huge financial losses. In these important areas, sudden failures as well as small residual load capacities are not allowed. This is why higher safety margins and more conservative structural designs are the dominant design approach in current designs, while another problem with engineered materials is that they break down without prior detectable damage and warning [2].

2 Literature Review

The composite material is made by two or more components. Composite material can be made by using fiber that are cured within a resin. Using a combination of different materials, taking advantage of their strengths and reducing the impact of their weaknesses is an important idea in designing composite materials. The most common types are combining carbon fibers and glass fibers with a thermosetting resin to create either a CFRPs or GFRPs. The use of multiple fiber-reinforced polymer laminations to form a new composite material with enhanced mechanical properties, such as compressive and tensile resistance, and the consideration of how to efficiently monitor the new composite material, based on previous research by scientists, has become an important key, and finally some of the advantages and disadvantages of previous monitoring methods in relation to the composite material in this experiment are presented.

3 Aims and Objective

How to monitor the health of composite materials is an important current research issue. It can effectively contribute to the development of several areas where composites are used, such as aerospace engineering and the automotive industry. However, in the traditional composite fiber industry the damage itself is unpredictable and sudden as it can be caused by different kind of stresses and pseudo-plastic deformations. There are many different methods of detection, but most of them do not reflect the changes in the material in real time and are also too expensive. This project therefore proposes a method to monitor changes in material resistance based on the fact that the resistance of thin carbon fibers changes gradually during damage. Research done by Meisam has shown that damage to fibers varies linearly (Rev, Jalalvand et al. 2019) and therefore the health of the material can be monitored by detecting changes in resistance based on this theory. The aim of this project is to design another sensor based on resistance variation, to experiment in order to check if the sensor is usable, and to design and test the sensor in order to achieve the best possible results, observing through experimentation and results whether it is sensitive and efficient.

Research Objective 1. Research test samples and target sensors based on existing literature and conjecture.

Research Objective 2. Analysis of the change in resistance of the sample and its force and displacement profiles.

Research Objective 3. Using a hybrid S-shape to detect damage on a flat plane and resistance changes monitored using a carbon glass hybrid.

Structural Health Monitoring of Carbon Fiber Composite Lamination Using Electrical Resistance

Guiping Lu[1(✉)], Xiaofeng Zhang[1], Shan Lu[2], Binghua Su[1], Kejun Wang[1], and Jiaran Liang[1]

[1] Beijing Institute of Technology, Zhuhai, Zhuhai, China
344088386@qq.com
[2] BMW Brilliance Automotive Ltd, Shenyang, China

Abstract. It focuses on a composite material made of glass and carbon fibers in this paper. The composite can be actively monitored and controlled by the self-sensing of the carbon fibers. However, due to the high stiffness and brittleness of the composite material, damage often occurs instantaneously. It is difficult to monitor damage patterns and control damage through factors such as fiber type variables and displacement relationships. This is why monitoring the health of composite fibers is an important direction, which has major implications for the aerospace, industrial and automotive sectors. In this project, the main focus is to monitor the electrical conductivity of carbon fibers online by breaking thin layers and observing the changes in their conductivity, and to understand changes in condition through changes in current. In addition, the composite design of this project can be applied to the monitoring of large planar materials, as well as to applications in important areas such as aerospace. In making further comparisons, it can be seen that the 5 mm thin layer of carbon fiber is more sensitive in the process of self-sensing, while the change in resistance is more noticeable when damage is received in the period.

Keywords: Structural Health Monitoring · Carbon Fiber Composite Lamination · Electrical Resistance · Three-point bending test

1 Introduction

There are more and more high-tech products made of composite materials, such as aerospace or automotive, and even the latest batteries. They are becoming increasingly popular due to their outstanding properties, such as their high strength, low weight and fatigue resistance, Composites are combinations of two or more materials with different physical behavior and chemical states. In particular in this test, the materials used are fiber reinforced polymers. As the properties of composites are usually more variable, engineers consider their design structure and components to minimize failure during the design of composites [1].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 61–71, 2024.
https://doi.org/10.1007/978-3-031-65126-7_6

5. Carrard, V., et al.: Medical student mental health. https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health
6. Davis, C., Martin, G., Kosky, R., O'Hanlon, A.: Early intervention in the mental health of young people: a literature review. In: ERIC (2000)
7. Ediz, B., Ozcakir, A., Bilgel, N.: Depression and anxiety among medical students: examining scores of the beck depression and anxiety inventory and the depression anxiety and stress scale with student characteristics. Cogent Psychol. **4**(1), 1283829 (2017)
8. Eva, E.O., et al.: Prevalence of stress among medical students: a comparative study between public and private medical schools in bangladesh. BMC. Res. Notes **8**(1), 1–7 (2015)
9. Ge, F., Zhang, D., Wu, L., Mu, H.: Predicting psychological state among chinese undergraduate students in the covid-19 epidemic: a longitudinal study using a machine learning. Neuropsychiatric Disease Treat. **16**, 2111–2118 (2020)
10. Ghrouz, A.K., Noohu, M.M., Dilshad Manzar, M., Warren Spence, D., BaHammam, A.S., Pandi-Perumal, S.R.: Physical activity and sleep quality in relation to mental health among college students. Sleep Breathing **23**, 627–634 (2019)
11. Henry, S.K., Grant, M.M., Cropsey, K.L.: Determining the optimal clinical cutoff on the CES-d for depression in a community corrections sample. J. Affect. Disord. **234**, 270–275 (2018)
12. Jungbluth, C., MacFarlane, I.M., Veach, P.M., LeRoy, B.S.: Why is everyone so anxious?: an exploration of stress and anxiety in genetic counseling graduate students. J. Genet. Couns. **20**(3), 270–286 (2011)
13. Mao, Y., Zhang, N., Liu, J., Zhu, B., He, R., Wang, X.: A systematic review of depression and anxiety in medical students in china. BMC Med. Educ. **19**(1), 1–13 (2019)
14. McGorry, P.D., Killackey, E.J.: Early intervention in psychosis: a new evidence based paradigm. Epidemiology Psychiatric Sci. **11**(4), 237–247 (2002)
15. Moutinho, I.L.D., et al.: Depression, stress and anxiety in medical students: a cross-sectional comparison between students from different semesters. Revista da Associação Médica Brasileira **63**, 21–28 (2017)
16. Womble, M., Jennings, S., Schatz, P., Elbin, R.: A-173 clinical cutoffs on the state-trait anxiety inventory for concussion. Arch. Clin. Neuropsychol. **36**(6), 1228–1228 (2021)

Regarding depression: Concerning study duration, age, academic efficacy scores from the MBI questionnaire, and QCAE affective empathy scores, the corresponding p-values are 0.851, 0.626, 0.578, and 0.405, all significantly greater than 0.05. From a statistical standpoint, this indicates that these features do not manifest significant differences at the given level. In other words, we lack sufficient evidence to support significant relationships or disparities between these features and anxiety.

However, in the context of gender, history of psychological counseling, native language, health status, and MBI Cynicism scores, the corresponding p-values are all below 0.05. This implies that these features may possess some degree of correlation, association, or influence with anxiety. In terms of statistical analysis, these divergences suggest that these features might hold a certain impact or role in relation to anxiety emotions.

5 Conclusion

In this study, we investigated the statistical relationships between various factors and the occurrence of psychological disorders, revealing patterns of variation in the proportions of individuals affected by psychological disorders and the severity of these disorders across different populations. We identified several factors closely associated with psychological disorders, with gender, native language, and health status potentially exhibiting more significant correlations with anxiety and depression.

Nevertheless, our study does have certain limitations. The size of the dataset is relatively small, and the number of features is limited, which could potentially impact the accuracy of our conclusions. To arrive at more universally applicable conclusions, we require a more comprehensive dataset of medical student information and a larger sample size.

Acknowledgment. This work is supported by the Shenzhen Science and Technology Innovation Commission (Stabilisation Support Programme).

References

1. Ahad, A., Chahar, P., Haque, E., Bey, A., Jain, M., Raja, W.: Factors affecting the prevalence of stress, anxiety, and depression in undergraduate indian dental students. J. Educ. Health Promot. **10**, 266 (2021)
2. Al-Dabal, B.K., Koura, M.R., Rasheed, P., Al-Sowielem, L., Makki, S.M.: A comparative study of perceived stress among female medical and non-medical university students in dammam, saudi arabia. Sultan Qaboos Univ. Med. J. **10**(2), 231 (2010)
3. Behere, S.P., Yadav, R., Behere, P.B.: A comparative study of stress among students of medicine, engineering, and nursing. Indian J. Psychol. Med. **33**(2), 145–148 (2011)
4. Carrard, Valerie, C., et al.: The relationship between medical students' empathy, mental health, and burnout: a cross-sectional study. Med. Teacher **44**(12), 1392–1399 (2022)

The presence of a job or a partner appears to have limited influence on anxiety or depression.

4.2 Correlation Analysis

Through t-tests and chi-squared tests, we will determine features that exhibit robust correlations with anxiety and depression, as well as those with weaker correlations.

Table 2. Demographic characteristics of college students - Anxiety and Depression

Variable	Anxiety p-Value	Depression p-Value
Gender	0.000***	0.000***
Job	0.439	0.018**
Part	0.242	0.012**
psyt	0.000***	0.000***
year	0.083*	0.084*
age	0.76	**0.626**
glang	0.000***	0.000***
stud_h	**0.987**	**0.851**
health	0.001***	0.000***
qcae_cog	0.216	0.15
qcae_aff	0.074*	**0.405**
mbi_ex	**0.587**	0.053*
mbi_cy	0.005***	0.000***
mbi_ea	**0.529**	**0.578**

Note: ***, **, * represent significance levels of 1%, 5%, and 10%, respectively.

Based on the results from Table 2, the following insights can be derived:

For anxiety disorder: In the examination of study duration, the significance p-value is 0.987; concerning the emotional exhaustion and academic efficacy scores from the MBI questionnaire, the respective p-values are 0.587 and 0.529; regarding the presence of a job, the p-value is 0.439. These outcomes indicate that statistically, the aforementioned features do not exhibit significant differences at the given level. In other words, we lack sufficient evidence to support a significant association or disparity between these features and anxiety.

However, when considering features such as gender, history of psychological counseling, native language, and health status, all respective p-values are below 0.05. This suggests the potential existence of some degree of correlation, association, or influence between these features and anxiety. This statistical divergence implies that these features might have a certain impact or role in relation to anxiety emotions.

Statistical Analysis of Other Variables' Relationship with Psychological Disorders. The statistical graphs illustrating the proportions of anxiety and depression, as well as the average scores of the affected population, varying with different features, are presented in Fig. 2.

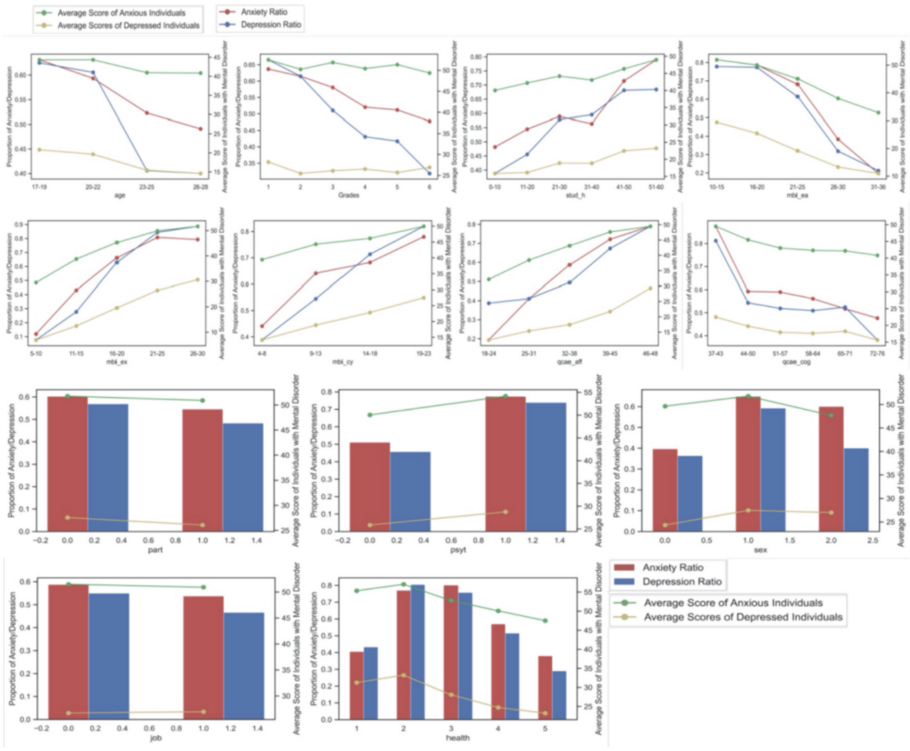


Fig. 2. Anxiety or depression statistical graphs

Features that exhibit a negative correlation with the proportion and severity of individuals with anxiety and depression include: age, academic year, academic efficacy, cognitive empathy, and health status. We observed that as age increases or academic year advances, the proportion of individuals with anxiety or depression decreases. Notably, the proportion of individuals with depression significantly drops after the age of 23. This may be attributed to medical students gradually adapting to the pace of learning, acquiring effective study methods, and consequently reducing the occurrence of anxiety and depression.

Features that show a positive correlation with the proportion and average scores of individuals with anxiety and depression include study duration, emotional exhaustion, cynicism, and affective empathy. We found that individuals with longer study durations exhibit a higher prevalence of psychological disorders, coupled with increased severity.

4 Result and Discussion

4.1 Statistical Analysis

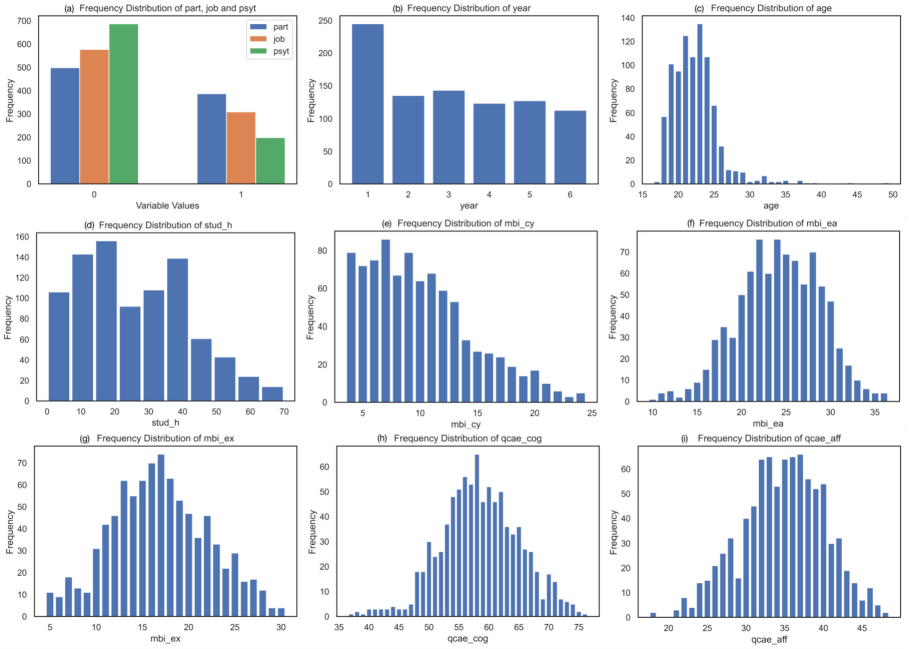


Fig. 1. Frequency Distribution of Each Feature. In Figure (a), the horizontal axis scale of 0 and 1 represents no partner (no job, no psychological treatment) and have a partner (job, psychological treatment) categories, respectively.

Univariate Statistical Analysis. The statistical graphs for each feature are depicted in the Fig. 1.

From Fig. 1(a), it is evident that students without partners outnumber those with partners, and similarly, students without jobs exceed those with jobs. Most students have not undergone psychological therapy over the past year.

Figure 1(b) illustrates that the first-year student count significantly surpasses other academic years, while second to sixth-year students exhibit a more even distribution.

Figure 1(c) indicates that the age distribution of medical students in the sample is concentrated between 18 to 25 years.

Figure 1(d), the highest number of individuals falls within the 10 to 20 h per week study time range. Most individuals do not exceed 40 h of study time per week.

The distributions of other features approximate a normal distribution.

Statistical Description. Creating statistical graphs for individual features provides a more intuitive display of the distribution of each feature's quantity, aiding in gaining a deeper understanding of the overall feature distribution within the sample population.

We have chosen two indicators, the proportion of individuals with psychological disorders and the average scores of the affected population, to depict the quantity and severity of psychological disorders. By visualizing the trends of these two indicators in relation to other features, we can gain a clearer insight into the influence of these features on psychological disorders.

Statistical Inference. This study primarily employs two hypothesis testing methods: the t-test and the chi-squared test, to conduct an analysis of dissimilarities among various features.

The independent samples t-test is utilized to compare differences between categorical and quantitative samples (samples A and B). The main steps are as follows:

1. Hypothesis formulation: The null hypothesis assumes no significant difference between samples A and B, while the alternative hypothesis assumes the presence of a difference.
2. Assumption of sampling distribution: Independent samples A and B are assumed to be approximately normally distributed, satisfying the conditions for t-distribution.
3. Calculation of t-value:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}}$$

4. Calculation of confidence interval for means: Using the computed t-value, along with sample sizes and confidence level, the confidence interval for means is calculated, allowing for statistical inference regarding mean differences.

The Pearson chi-squared test is employed for analyzing differences between two categorical sample variables. The statistical measure used is

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i},$$

which assesses the disparity between theoretical frequencies and observed values.

commonly utilized, each specifically designed for screening anxiety and depression symptoms, respectively.

3 Methods

3.1 Dataset Introduction

The dataset [4,5] for this study was released in 2020 and encompasses information from 886 medical students. The features comprise individual demographic details ('age', 'year', 'sex', 'glang', 'part', and 'job'), educational aspects ('stud_h', 'mbi_cy', and 'mbi_ea'), psychological conditions ('qcae_cog', 'qcae_aff', and 'mbi_ex'), and physical well-being ('health'). Two labels describing the psychological disorder status are 'stai_t' and 'cesd', which are derived from the STAI-T and CESD questionnaires respectively. These questionnaires are widely employed for screening anxiety and depression patients. The introduction of each feature is shown in the Table 1.

Table 1. Study variables

variable name	description
age	age at questionnaire
year	curriculum year
sex	gender
glang	mother tongue
part	having a partner
job	have a paid job
stud_h	how many hours per week spend on study
health	How satisfied are you with your health
psyt	consulted a psychotherapist or a psychiatrist for health
qcae_cog	QCAE Cognitive empathy score
qcae_aff	QCAE Affective empathy score
mbi_ex	MBI Emotional Exhaustion
mbi_cy	MBI Cynicism
mbi_ea	MBI Academic Efficacy

3.2 Statistical Analysis

This study primarily engages in statistical description and hypothesis testing of the dataset, aiming to identify the relationships between psychological disorders (anxiety or depression) and various features.

1 Introduction

The psychological well-being of college students has garnered extensive attention. The university phase, which signifies the transition between academic and social realms [10], marks the initial steps of students venturing into the societal arena. However, due to factors such as uncertainty about the future, substantial academic pressure, challenges in interpersonal relationships, and insufficient self-confidence, college students are susceptible to experiencing psychological health issues such as anxiety and depression [12].

Among various academic disciplines, medical students particularly warrant significant concern as they encounter heightened psychological health challenges [2,3,8]. Their prolonged academic duration, substantial academic pressures, and the weight of future employment prospects create a formidable environment. Moreover, the daily exposure to patients' ailments and suffering brings about negative emotions, thereby increasing the likelihood of psychological health problems.

Despite the plethora of research focusing on factors contributing to psychological disorders, there remains a relative scarcity of investigations concentrating on medical students. Consequently, this study primarily revolves around medical students as a specific sample group, delving into their prevalence and severity of psychological disorders. Concurrently, we aim to discern potential factors contributing to the onset of psychological ailments and analyze the varying degrees of correlation between these factors and psychological disorders.

2 Related Work

Many researchers have initiated investigations into the psychological well-being of medical students. Medical students exhibit higher levels of depression, anxiety, and stress symptoms [7,15]. Such psychological disorders as anxiety and depression can potentially have adverse effects on medical students' personal and professional lives, leading to issues like insomnia and even triggering thoughts of suicide [9].

Mao *et al.* [13] found that the occurrence of depression and anxiety among medical students is influenced by a variety of factors, including individual characteristics, socioeconomic status, and environmental factors such as gender, academic year, family structure, family income, parental educational background, and social support. Additionally, Ahad *et al.* [1] revealed that age, gender, employment status, and accommodation situation are significant factors affecting stress levels among medical students. Notably, female students tend to experience higher stress levels, and those engaged in clinical internships face greater stress compared to pre-internship periods. It's noteworthy that the findings by Moutinho *et al.* [15] emphasize significant variations in the psychological well-being of medical students across different semesters.

Early detection and treatment of mental disorders are crucial for achieving favorable recovery outcomes and reducing the risk of relapse [6,14]. Typically, questionnaire surveys are employed for the early screening of anxiety and depression patients. Among these, the STAI-T [16] and CESD [11] questionnaires are



Analysis of Factors Related to Anxiety and Depression in Medical Students

Zheng Jinfang[1,2,3], Pan Jiachen[1,2,3], Zhang Peiyi[1,2,3], Xiao Yi[2,3], and Wang Wei[2,3,4(✉)]

[1] Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
1120200266@smbu.edu.cn

[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
xiaoyi@smbu.edu.cn, ehomewang@ieee.org

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

[4] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. The psychological well-being of university students, particularly those pursuing medical education, has garnered widespread attention. These students hold the potential to shape the future of societal progress, with medical students shouldering a crucial responsibility for the development of overall community health. However, many medical students are susceptible to psychological disorders such as anxiety and depression due to high levels of stress. While numerous studies have investigated factors contributing to the prevalence of psychological ailments in the general population, there has been a limited focus on analyzing this phenomenon specifically among medical students. This study utilizes a sample of 886 medical students, gathering information regarding their personal backgrounds, academic pursuits, psychological states, and physical health conditions. The aim is to discern which subgroups have a higher prevalence of anxiety or depression. Employing statistical analysis, the relationships between various factors and the occurrence of psychological disorders are examined. Through differential analysis, factors with a stronger correlation to psychological disorders are identified. Notably, factors like study duration and emotional fatigue exhibit a positive association with anxiety and depression, while factors such as academic year and academic efficacy demonstrate a negative correlation. Furthermore, gender and health status exhibit robust correlations with the manifestation of anxiety and depression.

Keywords: Anxiety · depression · medical students · correlative factors

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 52–60, 2024.
https://doi.org/10.1007/978-3-031-65126-7_5

34. Williams, J.B., First, M.: Diagnostic and statistical manual of mental disorders. In: Encyclopedia of Social Work (2013)
35. Wongkoblap, A., Vadillo, M.A., Curcin, V., et al.: Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study. JMIR Mental Health **8**(8), e19824 (2021)
36. Wu, J., Zhou, Z., Wang, Y., Li, Y., Xu, X., Uchida, Y.: Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction. In: 2019 International Conference on Multimodal Interaction, pp. 582–588 (2019)
37. Yoon, J., Kang, C., Kim, S., Han, J.: D-vlog: Multimodal vlog dataset for depression detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 12226–12234 (2022)
38. Zhang, H., et al.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18802–18812 (2022)
39. Zheng, W., Yan, L., Wang, F.Y.: Two birds with one stone: knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. IEEE Trans. Affect. Comput. **14**(4), 2595–2613 (2023)
40. Zhou, L., Liu, Z., Yuan, X., Shangguan, Z., Li, Y., Hu, B.: Caiinet: neural network based on contextual attention and information interaction mechanism for depression detection. Digit. Sig. Process. **137**, 103986 (2023)
41. Zhu, Y., Shang, Y., Shao, Z., Guo, G.: Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Trans. Affect. Comput. **9**(4), 578–584 (2017)
42. Zogan, H., Razzak, I., Wang, X., Jameel, S., Xu, G.: Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web **25**(1), 281–304 (2022)

14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Lin, C., et al.: Sensemood: depression detection on social media. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 407–411 (2020)
17. Mann, P., Matsushima, E.H., Paes, A.: Detecting depression from social media data as a multiple-instance learning task. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE (2022)
18. de Melo, W.C., Granger, E., Hadid, A.: A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Trans. Affect. Comput. **13**(3), 1581–1592 (2020)
19. de Melo, W.C., Granger, E., Lopez, M.B.: MDN: a deep maximization-differentiation network for spatio-temporal depression detection. IEEE Trans. Affect. Comput. 14(1), 578–590 (2021)
20. Meng, Y., Bridge, J., Addison, C., Wang, M., Merritt, C., Franks, S., Mackey, M., Messenger, S., Sun, R., Fitzmaurice, T., et al.: Bilateral adaptive graph convolutional network on CT based covid-19 diagnosis with uncertainty-aware consensus-assisted multiple instance learning. Med. Image Anal. **84**, 102722 (2023)
21. Mitra, V., et al.: The SRI avec-2014 evaluation system. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 93–101 (2014)
22. Organization, W.H., et al.: Depression and other common mental disorders: global health estimates. World Health Organization, Technical Report (2017)
23. Safa, R., Bayat, P., Moghtader, L.: Automatic detection of depression symptoms in twitter using multimodal analysis. J. Supercomput. **78**(4), 4709–4744 (2022)
24. Saldanha, O.L., et al.: Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. NPJ Precision Oncol. **7**(1), 35 (2023)
25. Salekin, A., Eberle, J.W., Glenn, J.J., Teachman, B.A., Stankovic, J.A.: A weakly supervised learning framework for detecting social anxiety and depression. Proc. ACM Interact. Mobile Wearable Ubiquitous Technol. **2**(2), 1–26 (2018)
26. Shangguan, Z., Liu, Z., Li, G., Chen, Q., Ding, Z., Hu, B.: Dual-stream multiple instance learning for depression detection with facial expression videos. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 554–563 (2022)
27. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: a survey. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8135–8153 (2022)
28. Sotelo, J.L., Nemeroff, C.B.: Depression as a systemic disease. Personalized Med. Psychiatry **1**, 11–25 (2017)
29. Vahia, V.N.: Diagnostic and statistical manual of mental disorders 5: a quick glance. Indian J. Psychiatry **55**(3), 220 (2013)
30. Valstar, M., et al.: Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2014)
31. Valstar, M., et al.: Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2013)
32. Verhoeven, J.E., Révész, D., Epel, E.S., Lin, J., Wolkowitz, O.M., Penninx, B.W.: Major depressive disorder and accelerated cellular aging: results from a large psychiatric cohort study. Mol. Psychiatry **19**(8), 895–901 (2014)
33. Wang, T., Li, C., Wu, C., Zhao, C., Sun, J., Peng, H., Hu, X., Hu, B.: A gait assessment framework for depression detection using kinect sensors. IEEE Sens. J. **21**(3), 3260–3270 (2020)

detection task, and the best performance results illustrate the effectiveness and superiority of our proposed method. We hope that our work can add more effective contributions to the field of weakly supervised depression detection. In future work, we hope to add more modal social media such as text for depression detection.

Acknowledgment. This work was supported by the National Nature Science Foundationof China (62102266, 62231020, 62272317), Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University, Public Technology Platform of Shenzhen City (GGFW2018021118145859), Shenzhen Science and Technology Innovation Commission (R2020A045), Natural Science Foundation of Guangdong Province-Outstanding Youth-Program (2019B151502018), Pearl River Talent Recruitment Program ofGuangdong Province (2019ZT08X603, 2019JC01X235), National Key R&DProgram of China (2020YFA0908700), and the Natural Science and Engineering Research Council of Canada (corresponding author: Xiping Hu,huxp@bit.edu.cn).

References

1. Al Jazaery, M., Guo, G.: Video-based depression level analysis by encoding deep spatiotemporal features. IEEE Trans. Affect. Comput. **12**(1), 262–268 (2018)
2. Alghowinem, S., Goecke, R., Wagner, M., Parker, G., Breakspear, M.: Eye movement analysis for depression detection. In: 2013 IEEE International Conference on Image Processing, pp. 4220–4224. IEEE (2013)
3. Bourke, C., Douglas, K., Porter, R.: Processing of facial emotion expression in major depression: a review. Aust. NZ. J. Psychiatry **44**(8), 681–696 (2010)
4. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Commun. **71**, 10–49 (2015)
5. Cummins, N., Sethu, V., Epps, J., Schnieder, S., Krajewski, J.: Analysis of acoustic space variability in speech affected by depression. Speech Commun. **75**, 27–49 (2015)
6. Feng, J., Zhou, Z.H.: Deep miml network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
7. Ge, Y., Zhou, Q., Wang, X., Shen, C., Wang, Z., Li, H.: Point-teaching: weakly semi-supervised object detection with point annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 667–675 (2023)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
9. Gui, T., et al.: Cooperative multimodal approach to depression detection in twitter. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 110–117 (2019)
10. Hashimoto, N., et al.: Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3852–3861 (2020)
11. Hendra, C., Pratanwanich, P.N., Wan, Y.K., Goh, W.S., Thiery, A., Göke, J.: Detection of m6a from direct RNA sequencing using a multiple instance learning framework. Nat. Methods **19**(12), 1590–1598 (2022)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433 (2016)

As show in Fig. 1, the max pooling operation performs the worst, but is comparable to machine learning-based methods. The mean pooling operation outperforms the max pooling operation and achieves comparable results to multiple deep network based methods. In contrast, our proposed attention-based pooling operation achieves the best result, which shows that the attention mechanism effectively improves the performance of the MIL framework.

4.5 Effect of Instance Temporal Size

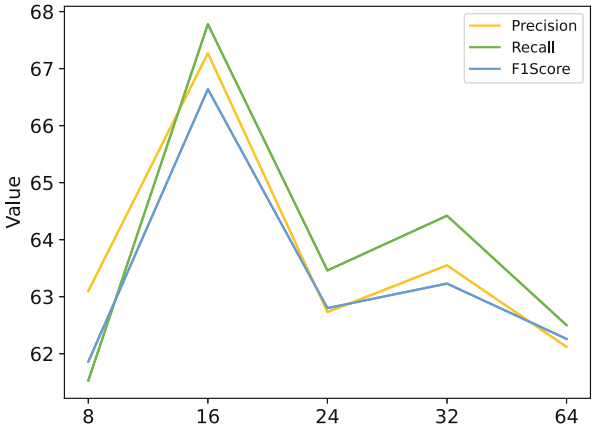


Fig. 2. Evaluation of varying instance temporal size on D-Vlog dataset

To assess the effect of instance time segments size on the method, we construct time segments k of different sizes for the study. As shown in Fig. 2, the best results are achieved in all metrics when k = 16. Moreover, it is worth noting that the results of the model do not exhibit linear change when the value of k increases or decreases, demonstrating that the smaller or larger time segments are not appropriate in depression detection. Technically, the size of the time segment determines the length of the time information contained in the instance. When the size of time segment decreases, continuous time series entering the time window is too short to perform sufficient feature aggregation through the multiple instance pooling layer. When the size of time segment increases, the redundancy of too much temporal information may affect the training of the model.

5 Conclusion

In this study, we perform an attention-based multiple instance learning method to detect depression using social media. We conduct sufficient experiments on the D-Vlog dataset and report the state-of-the-art model performance compared to us. The experimental results show the advantages and potential of multiple instance learning in depression

Table 1. Evaluation of the proposed methods on D-Vlog dataset

Method	Precision(%)	Recall(%)	F1Score(%)
LR	54.86	54.72	54.78
SVM	53.10	55.19	52.97
RF	57.69	58.49	57.84
KNN-Fusion	57.86	59.43	54.25
BiLSTM	60.81	61.79	59.70
TFN	61.39	62.26	61.00
Transformer_Concat	62.51	63.21	61.10
Transformer_Add	59.11	60.38	58.11
Transformer_Multiply	63.48	64.15	63.09
Depression_Detector	65.40	65.57	63.50
Temporal Convolutional	65.40	64.70	65.00
CAIINET	66.57	66.98	66.56
Ours	**67.27**	**67.77**	**66.64**

4.4 Comparison with the MIL Methods

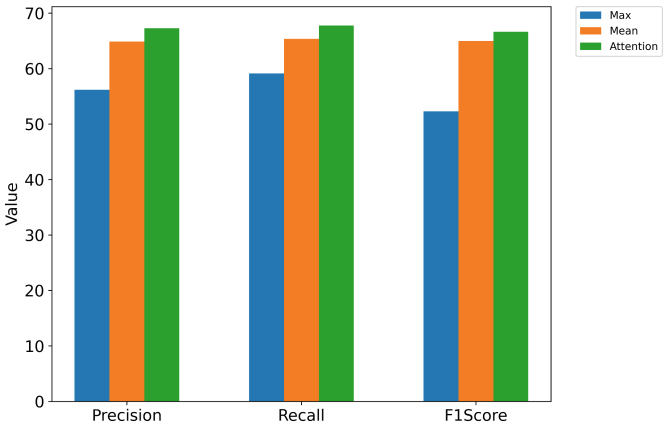


Fig. 1. Evaluation of the MIL methods on D-Vlog dataset

In order to explore the potential of MIL in depression detection and compare with the attention-based MIL method used in this paper, we present the methods based on max pooling [6] and mean pooling [6] as comparison experiments. Notably, the max pooling and mean pooling operation select the feature with the highest and average feature among the instance to obtain the bag-level feature.

4 Experiments

4.1 Experimental Dataset

In this work, we use the D-Vlog dataset [37] collected from YouTube, which contains 961 vlogs videos from 816 subjects composed of 322 males and 639 females. The dataset has a total of 505 depressed subjects and 406 healthy controls, and the average length of the vlog is 596 s. According to the ratio of 7:1:2, the dataset is divided into training set, validation set, and test set, respectively. The preliminary label assignment of the dataset comes from the title keywords of the vlog. Usually, vlogs containing keywords such as "depression daily vlog", "depression journey vlog" and "depression vlog" are labeled as depressed vlog. In addition, vlogs containing keywords such as "daily vlog" and "haul vlog" are labeled as non-depressed vlogs. Then, two tasks are used to ensure the plausibility of labels. First, videos that do not conform to the "vlog" format (e.g., videos without appearance) are removed. Second, specific annotators are assigned to judge whether the subjects have depression by watching the vlog videos with automatic text generation. For privacy protection considerations, D-Vlog only provides the features of the extracted voice and facial expression in the video, which are the 15-dimensional extended Geneva Minimalistic Acoustic Parameter Set and the 68-dimensional facial landmarks, respectively.

4.2 Experimental Setup

In this paper, the size of the time segment that constitutes the instance is 16, and the total length of the bag is limited to 596. All models are trained for 30 epochs, using Adam [15] as the optimizer with learning rate, weight decay and eps are set to 1e-4, 5e-4, and 1e-8, respectively and the batch size is set to 16. We report weighted average precision, recall, and F1 score to evaluate model performance. In order to prevent overfitting, the model uses an early stopping mechanism during training. All experiments are implemented in pytorch, running on a server with NVIDIA 1660 s and 16 GB RAM.

4.3 Comparison with the Previous State-of-the-Art Models

We compare with current state-of-the-art methods to evaluate the effectiveness of our method, and the results are shown in Table 1. Specifically, the recent methods for comparison include: 10 methods using in the D-Vlog dataset [37] as the baseline, the Knowledge-Embedded Temporal Convolutional Transformer method proposed by Zheng et al. [39] and the CAIINET method proposed by Zhou et al. [40]. The traditional machine learning methods including LR, SVM and RF don't perform well on the D-vlog dataset, which is due to the lack of nonlinear fitting ability of machine learning. Moreover, corresponding deep learning methods including BISTM, TFN and Depression Detector achieve better performance compared to machine learning methods.

Compared with baseline, our proposed method improves at least 1.87%, 2.2% and 3.14% on the weighted average precision, recall and F1 score metrics. In addition, compared with the recently proposed Knowledge-Embedded Temporal Convolutional Transformer method and the CAIINET method, our proposed method has achieved the highest results in all metrics, indicating the effectiveness of the proposed method.

$$F_{max} = W_2 h_{max} \tag{6}$$

where W_1 and W_2 represent trainable weights respectively. Moreover, in order to integrate the information of the obtained vectors F_{mean} and F_{max}, we concat them and use a one-dimensional convolution operation to obtain the contextual kernel α:

$$\alpha = f_c\left([F_{mean} : F_{max}]\right) \tag{7}$$

where f_c represents the convolution operation with convolution kernel size is 1. Formally, by combining the context kernel α with the final output O of BILSTM, we obtain the instance features with temporal context. This step can be formulated into:

$$z = \sum_{t=1}^{T} a_t O_{w,t} \tag{8}$$

where,

$$a_t = \frac{\exp\left(\alpha O_{w,t}^{\top}\right)}{\sum_{\tau=1}^{T} \exp\left(\alpha O_{w,\tau}^{\top}\right)} \tag{9}$$

Technically, a_t is the attention weight to indicate the effectiveness of the BiLSTM output feature. Also, instance feature with temporal information is obtained by combining the attention weight with the output feature, which helps to articulate the dynamic information of depressed patients.

3.3 AD-MIL

Recently, many studies have attempted to use attention mechanisms to integrate them into the MIL framework [11,14,38]. Notably, Ilse et al. [14] demonstrate that MIL based on attention pooling can achieve better performance compared to conventional multiple instance pooling such as max pooling and mean pooling. Inspired by these, we use attention pooling to aggregate the instance features obtained in the previous section.

Formally, we denote $Z = \{z_1, \ldots, z_M\}$ as a bag of M instance features, and attention-based MIL pooling can be defined as:

$$e = \sum_{m=1}^{M} b_m z_m \tag{10}$$

with,

$$b_m = \frac{\exp\left\{q^{\top} \tanh\left(V z_m^{\top}\right)\right\}}{\sum_{k=1}^{M} \exp\left\{q^{\top} \tanh\left(V z_k^{\top}\right)\right\}} \tag{11}$$

where q and V are trainable parameters and hyperbolic tangent $\tanh(\cdot)$ is the element-wise non-linearity. In addition, b_m represents as an attention weight indicating the contribution of a given instance to the prediction of the whole bag. Therefore, different attention weights can be used as an implicit feature selection to make the final bag features more informative.

m-th instance of i-th bag. Furthermore, each instance s_{im} is assumed to have implicit label $y_m \in \{0,1\}$ to represent negative or positive, which is not given in practice due to labeling difficulties.

Traditional MIL meets the following constraints: A bag is positive if there is at least one positive instance, while a negative bag is only if all the instances making up the bag are negative. Formally, it follows that

$$Y = \begin{cases} 0, & \text{iff } \sum_m y_m = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

However, in the case of our work, there will be cases where both positive and negative instances are included in one bag, so assumption here is not strict. Hence, we propose an attention-based algorithm for depression detection by introducing a looser version of the attention mechanism to assign implicit weights to instances.

3.2 AD-LSTM

The proposed AD-LSTM module first uses Bi-directional LSTM (BiLSTM) [8] to extract the temporal information of the two directions of LSTM [12] as output, and then uses the attention mechanism to integrate the semantic features with temporal information to obtain the feature of the instance.

We develope BiLSTM combining information in both directions of LSTM at the same time to obtain richer semantic information in the instance. Notably, each layer of BiLSTM consists of LSTM in two directions, and the outputs of the layer are as follows:

$$h_{i,t} = l_f\left(O_{i-1,t}\right) \tag{2}$$

$$H_{i,T-t} = l_b\left(O_{i-1,T-t}\right) \tag{3}$$

$$O_{i,t} = \left[h_{i,t}, H_{i,T-p}\right] \tag{4}$$

where T represents the total length of the segment. l_f and l_b represent the forward and backward LSTM models, respectively. $h_{i,t}$ and $H_{i,T-t}$ represent the output of the i-th layer at the time t of the forward LSTM and the output of the i-th layer at the $T-t$ time of the backward LSTM. Then, we add the forward and backward features of the last layer w of BILSTM to get $O = \{O_{w,1}, \ldots, O_{w,T}\}$, and we connect the forward and backward output features of BILSTM at time T to form contextual feature $h = \{h_{1,T}, H_{1,T}, \ldots, h_{w,T}, H_{w,T}\}$. Similar to the feature map in CNN, each $h_{i,T}$ in h represents the feature of BILSTM at the last time. Therefore, in order to obtain rich contextual information, we use the mean pooling operation and max pooling operation, which is commonly used in spatial information processing, to obtain context features h_{mean} and h_{max}. Further, we use the two-layer network model to introduce the non-linear operations of the two pooling features:

$$F_{mean} = W_1 h_{mean} \tag{5}$$

2.2 Weak Supervision and Multiple Instance Learning

Multiple instance learning (MIL) is a form of weakly supervised learning, which is used to deal with model training under insufficient labels. Typically, in multiple instance learning, the model only receives coarse-grained bag-level labels, and the labels of the instances that make up the bag are unknown. According to the different MIL settings, the current MIL algorithm can be divided into instance-level [13] and bag-level [10]. Due to the difficulties and high costs in the actual labeling process, specific annotators can only assign bag-level labels in the context. Hence, MIL has been widely used in many fields including object detection [7], pneumonia detection [20] and tumor detection [24].

Recently, several works of MIL has been applied for depression detection. Concretely, in the use of weakly supervised learning framework, Salekin et al. [25] proposed a MIL method to identify depression from voice speech containing labels of depressed patients without providing specific segments of symptoms. Shangguan et al. [26] proposed a dual-stream MIL deep network to identify depression by using raw facial expressions. In addition, extensive works have used MIL for detecting depression on social media due to its superiority. Wongkoblap et al. [35] proposed two multiple instance learning models to predict depression using textual information from Twitter. Moreover, a MIL method for detecting depression using students' posts from university was presented by Mann et al. [17]. They performed theoretical and experimental analysis by using Transformer and LSTM model on the dataset of university students.

Previous work using MIL to identify depression in social media has mainly focused on text information, and few works have used the information of subjects' facial expressions and voices to identify depression. Since the facial expressions and voice can express the mental state of the subjects and the effectiveness of MIL in the detection of depression, it is very necessary to establish a model for detecting depression using MIL based on these two modalities. Inspired by the work of these pioneers, we aim to expand the scope of the current literature on depression detection through the application of MIL and attention mechanisms.

3 Methods

We propose a weakly supervised learning model for the depression detection task in a single end-to-end deep network. Concretely, our model receives the vocal features and visual features extracted by OpenSmile and Dlib respectively, and then the AD-LSTM module extracts the temporal information within the instances. Finally, the AD-MIL module integrates the information of the instance for identifying depression. In this section we present the formulation of MIL and the details of the proposed MIL model.

3.1 Preliminaries

The MIL algorithm receives N labeled sample pairs $\mathrm{D} = \{(S_1, Y_1), \ldots, (S_N, Y_N)\}$, where S_i (i from 1 to N) is the whole bag and Y_i is $\{0, 1\}$ for binary classification of depression and health. Also, $S_i = \{s_{i1}, s_{i2}, \ldots, s_{iM}\}$ consists of M instances where s_{im} represents the

approaches using weakly supervised learning. In Sect. 3, we introduce the relevant preliminaries and the details of our proposed model. We provide the datasets, experimental settings and results used in the experiments in Sect. 4. Finally, we conclude our work in Sect. 5.

2 Related Work

This section briefly reviews the related methods of depression detection and weakly supervised learning.

2.1 Deep Learning for Depression Detection

Since the emerging applications in affective computing, the deep learning-based methods can use behavior signals for depression detection. The datasets for depression detection tasks can be divided into task-specific collection and non-specific task collection. In a specific task, the process of data collection comes from recording subjects completing a certain task according to the requirements of the examiner, such as answering some specific questions or discussing the certain topics. In a non-specific task, the process of data collection comes from external information, such as, voice, video, and text of individuals on the Internet.

AVEC2013 [31] and AVEC2014 [30] are typical task-specific datasets which focus on video modalities. For example, Zhu et al. [41] proposed a two-stream deep network to detect depression by considering the appearance and movements of subjects. By leveraging the optical flow of dynamic information of facial expressions, they improved the performance of the model. Similarly, Jazaery et al. [1] used a convolutional 3D network (C3D) to capture spatio-temporal information and to learn the features of continuous segments through Recurrent Neural Network (RNN). To reduce the model size for depression detection, Melo et al. [19] proposed the 2D deep network (a.k.a., MDN) to capture the spatio-temporal information in facial videos. By embedding the maximization block and difference block in the 2D deep network, the model captured the subtle changes and sudden transitions between face expression, and achieved comparable performance to 3D deep network.

Since a considerable number of users share recent life emotions and states on the Internet, social media can provide data information under non-specific tasks for depression detection. There are several approaches that use multi-modal data of social media for depression detection. For example, Safa et al. [23] used the biological features, features generated by analyzing user profile pictures, and banner images to detect depression. By using image and text information posted by users on social media, Gui et al. [9] introduced a new collaborative multi-agent reinforcement learning method to predict depression. Zagan et al. [42] presented a novel interpretable depression detection framework, the Hierarchical Attention Network, which used textual, behavioral, temporal, and semantic aspects of social media features for deep learning. Moreover, a deep visual-textual multimodal learning system dubbed SenseMood was proposed to predict the mental state of the users on social networks. Lin *et al.* [16] used CNN and Bert to extract deep representations of pictures and text on social media, which were combined for further depression classification.

of depression mainly relies on the subjective and complex reports of the subjects and the professional judgment of the psychiatrists. For example, the clinician rating scales (e.g., Hamilton rating scale) require rigorous training of raters. The self-rating depression scales (e.g., Self-rating depression scale) rely on accurate description, assessment, and expression of subjects and may change the purpose of their report [29]. Due to the lack of medical resources, the great harm of depression, and the large number of patients, the subjective assessment and diagnosis cannot meet the current demands for depression diagnosis. In this vein, the automatic detection of depression has attracted ever-increasing research attention due to objectivity, fast deployment, and long endurance.

With the advancement of affective computing, previous studies use behavioral signals as objective indicators to conduct research on depression detection, which provides an objective and effective way for auxiliary diagnosis of depression. Many current research outcomes have shown that common behavioral signals can be used as objective indicators for depression detection, such as, eye movement [2], voice [4], gait [33], and facial expression [3].

Different from eye movements and gaits that need to be collected during professional experiments, the leveraged facial expressions and voices in our research obtained through more relaxed methodologies (e.g., social media). In this paper, we use social media data collected from vlog of people documenting their daily lives on the Internet. Compared with data collected from the experimental environments, vlog data has three advantages: 1) easier to obtain; 2) larger quantity; and 3) consistent increase in volume. The three advantages of the vlog data allow to build a more generalized model and explore the ability to articulate the datasets in the wild. In general, a vlog dataset has both facial expressions and voice modalities, and there are dedicated annotators to judge whether the subjects are depressed. However, a complete piece of vlog data has only one binary classification sparse label, because it is impractical for annotators to accurately label the symptoms reflecting depression at a fine-grained level. The traditional depression detection methods [1,5,18,21] assign the same coarse-grained labels to the training instances (video clips or single frames) and may lead to overfitting and corresponding performance loss [27,36].

To address this challenge, we propose a weakly supervised method to identify the binary classification of the subjects (depression or health) using vlog data. Our model takes temporal segments of a certain size as instance input, and uses AD-LSTM to extract temporal contextual information to obtain instance-wise representations. Then, AD-MIL views the vlog video of each subject as a bag containing multiple instances that may be positive or negative. More specifically, the AD-MIL model first uses an attention mechanism to identify the contribution weights of instances to the final classification, and then obtains individual subject representations by combining the weights with instance representations. Technically, using the attention mechanism can effectively alleviate the impact of instances that are not related to the classification label and integrate the information of the whole bag. We conduct a series of experiments on the D-vlog dataset [37], and the experimental results show that our proposed method exceeds the state-of-the-art works, indicating the effectiveness of the proposed method.

The remaining parts of this work are organized as follows. In Sect. 2, we present recent approaches to automatic depression detection using deep learning as well as



Automatic Depression Detection Using Attention-Based Deep Multiple Instance Learning

Zixuan Shangguan[1], Xiaxi Li[2], Yanjie Dong[2], and Xiaoyan Yuan[1(✉)]

[1] Beijing Institute of Technology, Beijing, China
yxy_newly@163.com
[2] Shenzhen MSU-BIT University, Shenzhen, China
{1120200239,ydong}@smbu.edu.cn

Abstract. Depression is a serious mental illness and one of the leading causes of suicide worldwide. However, the social prejudice and the lack of psychiatrists for depression lead to a significant number of depressed patients without accurate diagnosis and subsequent serious consequences. With the rise of social media, previous studies have found that the information of depressed patients on social media can be analyzed to automatically detect depression for auxiliary diagnosis. In the context of weakly supervised learning framework, a multiple instance learning (MIL) method is proposed to identify depression from social media with visual and vocal information. By leveraging the state-of-the-art attention-based deep LSTM (AD-LSTM), the proposed MIL method can handle the problem with sparse labels (i.e., one label for a long-term sequence of visual information). More specifically, the AD-LSTM module is used to process a fixed-length visual and vocal segments to extract temporal representations of instances, and the AD-MIL module is used to aggregate the obtained temporal representations for individual subject predictions. Compared with current benchmarks, our experiments demonstrate that our proposed MIL method can achieve the best weighted average precision, recall and F1 score with the corresponding values as 66.56%, 66.98% and 66.55%, respectively. The numerical results illustrate that the potential and effectiveness of our proposed MIL method in the field of depression detection.

Keywords: Depression Detection · Multiple Instance Learning · Social media

1 Introduction

Major depressive disorder (a.k.a.. depression) is a critical mental illness with serious consequences for individual physical and mental health. More than 300 million people worldwide, which is equivalent to 4.4% of the global population, are currently suffering from varying degrees of depression [22]. Depressed people often exhibit low mood, loss of interest in practice, sleep disturbance, loss of appetite, lack of self-confidence, loss of energy, and inability to concentrate [34]. In addition, depression increases the risk of diabetes, heart disease, Alzheimer's and, in more severe cases, suicide [28,32].

Accurate diagnosis of depression can be effectively controlled and treated through psychological consulting and psychotropic medication. However, the current diagnosis

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 40–51, 2024.
https://doi.org/10.1007/978-3-031-65126-7_4

29. Shen, G., et al.: Depression detection via harvesting social media: a multimodal dictionary learning solution. In: IJCAI, pp. 3838–3844 (2017)
30. Shen, G., e al.: Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 3838-3844. AAAI Press (2017)
31. Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of suicide ideation in social media forums using deep learning. Algorithms **13**(1), 7 (2019)
32. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences (2018)
33. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans. Knowl. Data Eng. **32**(3), 588–601 (2018)
34. Wang, N., etal.: Learning models for suicide prediction from social media posts. arXiv preprint arXiv:2105.03315 (2021)
35. Wang, Y., Wang, Z., Li, C., Zhang, Y., Wang, H.: A multitask deep learning approach for user depression detection on sina weibo. arXiv preprint arXiv:2008.11708 (2020)
36. Yang, T., et al.: Fine-grained depression analysis based on chinese micro-blog reviews. Inf. Process. Manage. **58**(6), 102681 (2021)
37. Yao, X., Yu, G., Tang, J., Zhang, J.: Extracting depressive symptoms and their associations from an online depression community. Comput. Hum. Behav. **120**, 106734 (2021)
38. Zhou, S., Zhao, Y., Bian, J., Haynos, A.F., Zhang, R., et al.: Exploring eating disorder topics on twitter: machine learning approach. JMIR Med. Inform. **8**(10), e18273 (2020)
39. Zogan, H., Razzak, I., Jameel, S., Xu, G.: Depressionnet: a novel summarization boosted deep framework for depression detection on social media. ArXiv **abs/2105.10878** (2021)

13. Gui, T., et al.: Cooperative multimodal approach to depression detection in twitter. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.3301110
14. Holt-Lunstad, J., Smith, T.B., Baker, M., Harris, T., Stephenson, D.: Loneliness and social isolation as risk factors for mortality: a meta-analytic review. Perspect. Psychol. Sci. J. Assoc. Psychol. Sci. **10**(2), 227 (2015)
15. Kessler, R.C., et al.: Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative. World Psychiatry **6**(3), 168 (2007)
16. Kohler, C.G., Hoffman, L.J., Eastman, L.B., Healey, K., Moberg, P.J.: Facial emotion perception in depression and bipolar disorder: a quantitative review. Psychiatry Res. **188**(3), 303–309 (2011)
17. Lewis, M., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.703, null ; Conference date: 05-07-2020 Through 10-07-2020
18. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced NLP tasks. arXiv preprint arXiv:1911.02855 (2019)
19. Lin, H., Jia, J., Nie, L., Shen, G., Chua, T.S.: What does social media say about your stress? In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, pp. 3775-3781. AAAI Press (2016)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
21. Malhi, G.S., Mann, J.J.: Depression. The Lancet **392** (2019)
22. Malhotra, A., Jindal, R.: Deep learning techniques for suicide and depression detection from online social media: a scoping review. Appl. Soft Comput. **130**, 109713 (2022)
23. Meng, Y., Li, M., Li, X., Wu, W., Li, J.: Dsreg: using distant supervision as a regularizer. arXiv preprint arXiv:1905.11658 (2019)
24. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in twitter. In: Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012, pp. 1–8 (2012)
25. Pyszczynski, T., Holt, K., Greenberg, J.: Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. J. Pers. Soc. Psychol. **52**(5), 994 (1987)
26. Ríssola, E.A., Losada, D.E., Crestani, F.: A survey of computational methods for online mental state assessment on social media. ACM Trans. Comput. Healthcare **2**(2), 1–31 (2021)
27. Salas-Zárate, R., Alor-Hernández, G., Salas-Zárate, M.D.P., Paredes-Valverde, M.A., Bustos-López, M., Sánchez-Cervantes, J.L.: Detecting depression signs on social media: a systematic literature review. In: Healthcare, vol. 10, p. 291. MDPI (2022)
28. Sekulić, I., Strube, M.: Adapting deep learning methods for mental health prediction on social media. arXiv preprint arXiv:2003.07634 (2020)

5 Conclusions

In this work, we have devised a versatile framework for processing Twitter data to facilitate the diagnosis of early-stage depression. Through an extensive study, we have ascertained that Twitter textual content, user profile information, and historical posting data all hold profound significance in diagnosing depression. Consequently, we have proposed a comprehensive model that amalgamates these inputs and conducted empirical validations to evince the efficacy of our approach. Moreover, we addressed the issue of imbalanced data pertaining to depression patients by exploring several diverse methodologies, culminating in commendable achievements.

Acknowledgement. This work is supported by the Natural Science Foundation of Guangdong Province of China (No. 2021A1515011905)

References

1. Ahmed, U., Mukhiya, S.K., Srivastava, G., Lamo, Y., Lin, J.C.W.: Attention-based deep entropy active learning using lexical algorithm for mental health treatment. Front. Psychol. **12**, 642347 (2021)
2. Beck, A.T.: Cognitive Therapy of Depression. Guilford Press, New York (1979)
3. Belmaker, R.H., Agam, G.: Major depressive disorder. New England J. Med. Mech. Disease **385**, 47–60 (2008)
4. Birmaher, B., Ryan, N.D., Williamson, D.E., Brent, D.A., Kaufman, J.: Childhood and adolescent depression: a review of the past 10 years. part ii. J. Am. Acad. Child Adolescent Psychiatry **35**(11), 1427–1439 (1996)
5. Brent, A.D.: Course and outcome of child and adolescent major depressive disorder. Child Adolescent Psych. Clin. North Am. **11**(3), 619–637 (2002)
6. Carlson, G.A.: The challenge of diagnosing depression in childhood and adolescence. J. Affect. Disord. **61**(supp-S1), S3–S8 (2000)
7. Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., De la Torre-Díez, I.: Suicide risk assessment using machine learning and social networks: a scoping review. J. Med. Syst. **44**(12), 205 (2020)
8. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
9. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Fine, A., Crutchley, P., Blase, J., Carroll, J., Coppersmith, G.: Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data. In: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pp. 50–54 (2020)
12. Ghosh, S., Anwar, T.: Depression intensity estimation via social media: a deep learning approach. IEEE Trans. Comput. Soc. Syst. **8**(6), 1465–1474 (2021)

Table 3. The impact of different information

User information	Prec	Rec	F1	Acc
Historical tweet	80.2%	**89.0**%	**84.3**%	83.5%
Historical tweet+BiGRU	**85.8**%	82.8%	84.2%	**84.5**%

the model's robustness. In practical scenarios, this inclusion can mitigate the risk of misidentifying healthy individuals as depressed patients. Furthermore, as user information tends to display greater individuality and lacks a discernible pattern, our user information extraction module can still capture its distinctive features, thereby contributing to incremental improvements in the overall model.

4.4 The Experiment on Data Imbalance

To investigate the efficacy of traditional methods for handling imbalanced data in depression detection, we randomly extracted two imbalanced datasets from the original dataset. The first dataset had a ratio of depressed patients to non-depressed patients of 1:4, while the second dataset had a ratio of 1:2. Subsequently, we conducted a comparative analysis of three methods on these two datasets: the direct usage of cross-entropy loss without any imbalance treatment, the employment of Focal Loss for handling imbalance, and the utilization of Dice Loss for the same purpose. The experimental outcomes are presented in Table 4, with the evaluation metric being the F1 score.

Table 4. The effects of different methods for handling data imbalance.

Imbalanced ratio	CE	Focal Loss	Dice loss
1:4	33.3%	**79.5**%	75.7%
1:2	66.3%	79.8%	**81.8**%

From Table 4, it can be observed that when the imbalance ratio reaches 1:4, the model fails to learn useful classification knowledge from the limited positive samples when using cross-entropy loss without imbalance handling. Even under a 1:2 imbalance ratio, the results are significantly unsatisfactory. Focal Loss, in comparison to cross-entropy, demonstrates a 46.2% and 13.5% increase in F1 scores at 1:4 and 1:2 imbalance ratios, respectively. Similarly, Dice Loss shows a 42.4% and 15.5% improvement in F1 scores at 1:4 and 1:2 imbalance ratios. This verifies the effectiveness of traditional data imbalance handling methods for depression classification in our model. Additionally, it is evident that under higher imbalance conditions (1:4), Focal Loss outperforms, achieving an F1 score of 79.5%, whereas when the imbalance ratio reduces to 1:2, the performance of Dice Loss is superior, achieving an F1 score of 81.8%.

Historical Tweet Model. We employed the large-scale language pretraining model BERT and attention-based bidirectional LSTM to construct a historical tweet feature extraction model. Subsequently, we conducted experimental comparisons on each module, and the results are presented in Table 2. Note: due to better performance with straightforward extraction during data processing, all experiments were performed on data obtained through straightforward extraction.

Table 2. Historical tweet model

model	Prec	Rec	F1	Acc
BERT	77.7%	82.8%	80.1%	79.5%
BERT+BiLstm	**81.6**%	83.3%	82.4%	82.3%
BERT+StackedBiLstm	80.2%	**89.0**%	**84.3**%	**83.5**%

According to Table 2, it can be observed that the utilization of BERT in conjunction with the StackedBiLSTM model yields the most favorable results when processing textual features. Following this, the employment of BERT in combination with BiLSTM ranks second in performance. We believe this is due to a certain temporal correlation in the user's tweet data. Due to BiLSTM's capacity to maintain "memory," the model with an added BiLSTM layer outperforms the classification model solely relying on BERT. Furthermore, it is evident that the StackedBiLSTM model outperforms the BiLSTM model in terms of recall, F1 score, and accuracy, surpassing it by 5.7%, 1.9%, and 1.2%, respectively. However, it should be noted that the accuracy is reduced by 1.4% . We posit that this could be attributed to an excessive focus on contextual information, leading to the inadvertent capture of some depression-irrelevant data and thereby increasing the likelihood of misidentifying depression patients.

The Impact of User Information. We investigated the impact of user information on the classification of users with depression in our experiment. We extracted personal information from users, including the number of individuals they follow, the count of their followers, the quantity of friends, and the number of tweets sent in the past month. For feature extraction, we utilized a Bidirectional Gated Recurrent Unit (BiGRU) to complement the historical tweet features. We explored the experimental outcomes achieved by solely using historical tweets and by amalgamating user information with historical tweets. The model employed in this study is depicted in the aforementioned model diagram in Fig. 1, and the results are presented in Table 3.

From Table 3, it can be observed that the incorporation of user information and historical text enhances the model's precision and accuracy, surpassing the historical tweet model by 4.6% and 1%, respectively. However, the recall rate decreased by 6.2%. We contend that the inclusion of user information enhances

4.2 Hyperparameter Configuration

We attempted to employ the BERT pre-trained model as our Encoder model. For the tweet extraction model (a BiLSTM classification model with attention mechanism), we utilized a 2-layer BiLSTM with hidden layer neural units set to 128. As for the user behavior model, we opted for a single-layer BiGRU with hidden layer neural units set to 128. All experiments were conducted on an RTX3090GPU using the Pytorch framework. We employed the SGD optimizer for training with specific parameters: learning rate (lr) = 0.001, momentum = 0.9, and weight decay = 0.0004. Additionally, we employed a warm-up strategy to reach the initial learning rate. We performed a total of 30 epochs, and during the 10th and 20th epochs, we applied a learning rate decay with a rate of 0.1.

We introduced two approaches to process tweets: simple tweet extraction and K-means-based tweet filtering. To assess the performance of our model, we employed metrics such as accuracy, precision, recall, and F1 score. Additionally, to examine the impact of various components in the model, we conducted an ablation analysis. In experiments involving imbalanced data, we set $\gamma = 2$ in the Focal Loss and $\alpha = 0.9$ in the Dice Loss, with $\epsilon = 1e^{-4}$.

4.3 Ablation Study

Method of Data Processing. We have presented three distinct approaches for data processing: a straightforward tweet extraction method and a tweet filtering based on K-means clustering. These methods were subjected to experimental comparison. All models employed the BERT pre-trained model in conjunction with a single layer of Bidirectional LSTM (BiLSTM). The results are presented in Table 1.

Table 1. Module for Data Processing

Data processing	Prec	Rec	F1	Acc
simplistic tweet extraction	**81.6**%	83.3%	**82.4**%	**82.3**%
K-means	76.6%	**83.5**%	79.9%	79.0%

As indicated in Table 1, the employment of a simplistic tweet extraction approach exhibits superior performance compared to the utilization of the K-means extraction approach, yielding improvements of 5%, 2.5%, and 3.3% in accuracy, F1 score, and precision, respectively, over the BiLSTM model. In terms of recall, the difference between the two methods is negligible. We hypothesize that this discrepancy could be attributed to K-means clustering, which identifies text closely related to the classification but, at the same time, disrupts contextual coherence to some degree, leading to the deterioration of results.

the model to pay less attention to them. From a derivative perspective, once the model correctly classifies the current sample (just passed the 0.5 threshold), Dice Loss leads the model to pay less attention to it, unlike cross-entropy, which encourages the model to approach the two endpoints, 0 or 1. This effectively prevents the training of the model from being dominated by numerous straightforward samples.

3.4 User Information Model

After considering user behavioral data, we extracted relevant features pertaining to their social interactions, such as the number of followers and friends. Furthermore, we took into account user-generated actions, including the quantity of tweets posted and tweets favorited. These extracted features were utilized as inputs for the Bidirectional Gated Recurrent Unit (BiGRU) [9].

Both GRU and LSTM employ gating mechanisms to capture interdependencies among inputs, with GRU being a simplified variant of LSTM. Given the relatively straightforward nature of user behavioral data, we posit that the Bidirectional GRU is better suited for capturing relationships among these features.

Subsequently, the features derived from the Bidirectional GRU were fed into a fully connected layer. The resulting output from this layer serves as a guiding factor for classifying users within the historical posting model.

4 Experimental Setup

4.1 Dataset

We have reprocessed the extensive publicly available depression dataset proposed by [29]. These tweets were collected and labeled by the authors on Twitter, while also retrieving the user's historical tweets within a month. The dataset consists of three parts: (1) **Depression Patient Dataset D1**, comprising 2506 labeled samples of depressed users and their tweets; (2) **Non-depression Patient Dataset D2**, comprising 4166 labeled samples of non-depressed users and their tweets; and (3) **Depression Patient Candidate Dataset D3**. The author constructed a large-scale unlabeled depression candidate dataset containing 58,810 samples. In our experiments, we only utilized the labeled datasets: D1 and D2. We preprocessed the datasets by removing users with fewer than ten posting histories, users without an anchor tweet, or users posting tweets in languages other than English. Additionally, we removed emojis from the data to eliminate any impact on the experimental results, thus ensuring that we have sufficient statistical information related to each user. Finally, for balanced data experiments, we considered only 4000 user samples, with 2000 samples each for depressed and non-depressed users. For unbalanced data experiments, we explored the ratios of depressed users to non-depressed users at 1:2 and 1:4, with 2000 samples for non-depressed users in both cases. For testing purposes, we randomly divided the dataset into a training set (80%) and a test set (20%).

The mechanism of attention allows assigning distinct weights to each input feature and reflects the correlation between features and outcomes.

Data Imbalance. The phenomenon of data imbalance is quite common within social media datasets. This imbalance gives rise to the following two issues:

(1) Disparity between training and testing procedures: Under the influence of imbalanced data, models tend to converge towards points that strongly favor classes with the majority labels. This, in effect, creates a disparity between the training and testing processes. During training, each training instance contributes equally to the objective function, whereas during testing, F1 equally weighs the contributions of positive and negative samples.
(2) Excessive impact of simple negative samples on the model: As pointed out by [23], an abundance of negative samples implies a large quantity of straightforward negatives. Consequently, an overwhelming proportion of the loss stems from these numerous simple negative samples, thereby dominating the gradients and hindering the model from adequately learning how to differentiate between positive samples and challenging negative samples. Both cross-entropy (CE) and maximum likelihood estimation (MLE), which are extensively utilized loss functions in machine learning, fail to address these two issues.

Focal Loss and Dice Loss are two deliberately designed loss functions aimed at mitigating the imbalance between positive and negative samples during the one-stage classification process.

The primary objective of Focal Loss is to diminish the weight of easy samples, thereby focusing the training on the negation of difficult samples. To be more precise, Focal Loss introduces a modulation factor $(1-p_t)^\gamma$ into the cross-entropy loss, where $\gamma \geq 0$ represents an adjustable focal parameter. The general form of Focal Loss can be expressed as follows:

$$FL(p_t) = -(1-p_t)^\gamma log(p_t). \tag{4}$$

Dice Loss, on the other hand, contemplates the classification task from a distinct perspective. In this framework, categorizing a sample as negative is contingent solely upon its probability being less than 0.5; there is absolutely no need to drive it towards 0. Furthermore, considering that the primary objective is to mitigate the data imbalance issue within the dataset and, consequently, enhance the effectiveness of the F1 evaluation metric, Dice Loss is designed to exert a direct impact on F1.

Consequently, the general formulation of Dice Loss has been derived as follows:

$$Dice(p_t) = \frac{2(1-p_t)p_t \cdot y_t + \epsilon}{(1-p_t)p_t + y_t + \epsilon}, \tag{5}$$

where p_t represents the estimated probability, incorporated ϵ acts as a smoothing term, and y_t denotes the true label. The term $(1-p_t)$ serves as a scaling factor. For uncomplicated samples (when p_t approaches 1 or 0), $(1-p_t)p_t$ prompts

of tweets articulating their emotions over several days. Considering the token length constraints inherent in BERT [10] and the acknowledged temporal impact on a user's narrative, we judiciously adopt a straightforward strategy: selecting a user's most recent 20 tweets as the primary method for tweet processing. In cases where a user has fewer than 20 tweets, we employ padding techniques to ensure completeness of the dataset.

Tweet Filtering Based on K-Means Clustering. We acknowledge that not all of a user's posts are necessarily relevant to the point we focus on. For instance, a depressed individual might also publish tweets expressing positive emotions, such as 'Today's weather is lovely!' In order to mitigate the influence of such tweets on our determination of a user's depressive status, we endeavor to filter out tweets that more accurately portray the user's identity. Since we cannot introduce user labels during the processing phase, we adopt an unsupervised approach to analyze users' historical posts. Consequently, we employ the K-means clustering method as our second approach to tweet processing. We select a user's most recent 50 posts, tokenize them using BERT, apply the K-means algorithm to cluster the tweets into two categories, and then extract the 20 tweets closest to the cluster centroids. Should there be an insufficient number of tweets remaining, we will once again utilize padding to complete the dataset.

3.3 Historical Posting Information Model

We employed a pre-trained BERT model and a bidirectional LSTM (BiLSTM) based on an attention mechanism to process the input, capturing sequential information such as sentence context. Moreover, in light of the minority representation of depression patients, we tackled the prevalent issue of imbalanced data in the depression dataset by adopting the Focal Loss and Dice Loss techniques, as introduced in the works by [18,20], respectively.

Classification Module Based on Pre-trained Bidirectional LSTM. From the embedding layer of BERT, the extracted features are passed to the Bidirectional Long Short-Term Memory (BiLSTM), which is an RNN designed to capture sequential information and the long-term dependencies within sentences. Comprising the Bidirectional LSTM are the forward and backward LSTMs, each one independently updating the input x_i at time t:

$$forward(h_t) = LSTM(x_t, forward(h_{t-1})). \tag{1}$$

$$forward(h_t) = LSTM(x_t, backward(h_{t-1})). \tag{2}$$

After BiLSTM processing, the hidden state h_t at time t is a concatenation of the states $\overrightarrow{h}$ and $\overleftarrow{h}$ obtained from the forward LSTM and backward LSTM, respectively. The representation of the i-th word is as follows:

$$h_t = forward(h_{t-1}). \oplus backward(h_{t-1}). \tag{3}$$

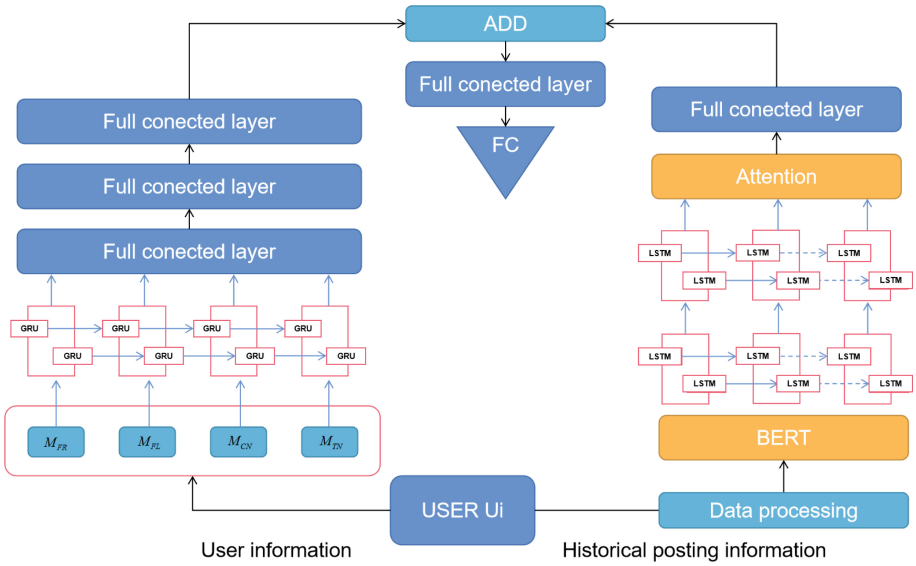


Fig. 1. We have integrated the user information and the user's past tweets of each user to forecast their labels. The user's historical tweets are processed through a tweet processing procedure and then handled by a classification model based on pre-training and bidirectional LSTM. The user's information, on the other hand, is processed using a model based on bidirectional GRU.

extraction, tweet filtering based on K-means clustering. Furthermore, we also incorporate user behavioral metrics, including the number of friends, followers, favorited contents, and the frequency of posts within a month, denoted as $M_{FR}, M_{FL}, M_{CN}, M_{TN}$ respectively.

3.2 Data Processing

The analysis of users' historical tweets constitutes a pivotal aspect of the depression assessment process. Consequently, our efforts are directed towards scrutinizing the past tweets of individuals experiencing depression, in comparison to those of mentally healthy users. This endeavor seeks to afford us a more profound understanding of the behavioral patterns manifested by individuals grappling with depression. Following this analysis, our objective is to leverage these insights to enhance the efficacy of depression detection methodologies.

Conducting Direct Tweet Extraction. Primarily, our initial consideration revolves around a streamlined approach to tweet handling, specifically involving the extraction of a defined quantity of tweets to encapsulate the entirety of a user's Twitter activity. We posit that the temporal dynamics of a user's posts exert a substantial influence on the thematic content of their tweets. To illustrate, an individual experiencing depression may consecutively share a series

to effectively capture both local contextual features and long-range dependencies [31,38].

Moreover, attention mechanisms [1,34,37] are employed to enable models to focus on the most salient aspects of the input. Additionally, multi-task learning is harnessed to jointly train models alongside other auxiliary tasks, such as statistical feature classification [35] and depression causation prediction [36], which yield supplementary insights for depression detection.

Recently, researchers have gathered data through online surveys and online social discourse, leveraging the substantial number of users and tweets on social platforms to obtain an ample and longitudinally sampled dataset. This approach effectively addresses the previously mentioned issue of collecting depression-related data. One study collected data from over 4,000 individuals, encompassing both depressed and non-depressed users on Twitter. [30] The dataset comprised over 2,000 samples from individuals diagnosed with depression, encompassing personal information and an anchor tweet to determine their depressive status. Additionally, all tweets posted within the 30 days preceding the anchor tweet were collected, based on the clinical characteristic of prolonged mood despondency among depression patients. For non-depressed users, relevant personal information and tweets within the same 30-day period were collected. However, this dataset did not account for the issue of data imbalance. Building upon this dataset, [39] further integrated personal information features and textual features, effectively removing redundant features through the utilization of k-means algorithm and the Bart summarization model [17], thereby improving the accuracy of depression identification. Our work builds upon the aforementioned research, focusing primarily on addressing data redundancy and data imbalance issues.

3 Method

In this section, we present our model for depression. Our model utilizes both the users' personal information and their historical posts as the foundation for detecting depression. Figure 1 illustrates our model diagram.

3.1 Task Definition

For social media datasets, users' posts often exhibit redundancy, irrelevance to their status, and may even contain unusable information, posing significant challenges for researchers to effectively extract user information. Herein, we have established the relevant symbols as defined in the article. We assume a user, denoted as U_i, has a total of n tweets in their history: $[T_1, T_2, \ldots, T_t, \ldots, T_n]$, where the t-th tweet represents the user's recent post. Our objective is to determine whether user U_i is a depression sufferer, for which we define the label as $y_i \in depression, undepression$. To achieve our goal, we amalgamate each user's profile information with their historical posts. We explore three approaches to process a user's historical tweets to address this issue: conducting direct tweet

extract activity information typically mandate real-time tracking for a duration exceeding two weeks. This extended timeframe, coupled with the requisite high level of patient cooperation, introduces significant costs and complexity into the research process.

Efforts to enhance the precision and applicability of AI-based depression identification must contend with these multifaceted challenges.

Our contributions are as follows:

1. We present a comprehensive depression assessment model that concurrently leverages users' historical tweets and their personal information. To achieve this, we meticulously devise distinct models for both historical tweets and user information. Our approach involves the integration of two discrete methodologies, one dedicated to handling multiple tweets and the other addressing the inherent challenge of imbalanced depression data.
2. We reprocess the depression tweet dataset to enhance its practical utility.
3. Through rigorous experimentation, we showcase the efficacy of our model in discriminating depression, yielding compelling empirical results and partially alleviating the associated challenges.

2 Related Work

In the field of psychology, early scholars have observed the theoretical correlations between mental health conditions and specific linguistic attributes, such as the presence of "depressive language" [2] advanced cognitive therapy and emphasized the significance of the frequency of negatively-valenced words, while other researchers [25] focused on the utilization of first-person pronouns and the patients' negative anticipations. Subsequent empirical investigations have validated these hypotheses and revealed associations between specific linguistic characteristics and the mental states of patients. Consequently, numerous studies utilize social media as a rich source of textual data, employing online user-generated posts for the manual analysis of mental health conditions. [7,26].

However, due to the burgeoning volume of online texts and the sensitivity of mental health conditions, manual text analysis and large-scale psychiatric interventions are no longer tenable. Consequently, Natural Language Processing (NLP) and text mining technologies have been harnessed to automate the analysis of mental health from social media data. While these approaches are not intended for definitive diagnoses, they do offer assistance in early detection [11,22,27]

Advancements in the realm of deep learning also bolster tasks related to mental health. The most recent methodologies employ deep learning models to automatically capture latent semantic information without the need for explicit feature engineering. Some studies utilize Convolutional Neural Networks (CNN) [33] or Recurrent Neural Networks (RNN) [8], including Long Short-Term Memory (LSTM) [12] and Gated Recurrent Unit (GRU) [28], to discern depression. Researchers also explore hybrid architectures combining CNN and RNN

1 Introduction

With the advancement and progress of society, people's material living standards are constantly improving, and psychological issues are receiving increasing attention. Psychological disorders are prevalent among young individuals, with approximately 75% of cases emerging during adolescence [15]. According to estimates by the World Health Organization, depression is one of the most prevalent psychological disorders, and by the year 2030, depression is projected to become the leading burden of disease globally [16,21].

Depression is characterized by significant and enduring mood melancholy, with symptoms encompassing sleep disturbances, appetite changes, and mental turmoil [3–5]. Despite its high prevalence, there is evidence indicating that 60% of severely depressed adolescents do not receive treatment.

Depression possesses a covert nature, and its occurrence is influenced by intricate factors such as heredity, gender, living environment, and physical ailments, rendering its diagnosis exceedingly challenging [4–6]. Presently, an accurate diagnosis of depression necessitates psychiatric practitioners to employ systematic inquiries, psychiatric examinations, and supplementary assessments, such as the Hamilton Rating Scale for Depression (HAMD) and the Patient Health Questionnaire-9 (PHQ-9) self-rating scale. Thus, the diagnostic evaluation heavily relies on patients' self-reported severity of depressive symptoms or clinical judgment regarding symptom severity. However, the advent of artificial intelligence-based approaches has presented the potential for objective diagnosis.

We focus on depression detection based on social networks. Recently, [14] discovered that a lack of social interaction increases the risk of depression. [24] analyzed the behavior and language usage of depressed users on Twitter. People's tweets on social networks such as Facebook, Twitter, and Weibo can be used to assess the risk of various mental health issues, such as depression and anxiety. [32] employed lemmatization tools to vectorize more recent tweets, reducing redundant features. [13] utilized a multimodal model and applied reinforcement learning to merge textual and image features of tweets, thereby improving the accuracy of depression identification [19].

The efficacy of depression identification through artificial intelligence remains suboptimal, encountering several noteworthy challenges. These challenges may be succinctly outlined as follows: **Limited Sample Size:** The recruitment of patients poses a substantial hurdle due to ethical concerns within the medical field. Consequently, a pervasive issue in depression studies is the constraint imposed by small sample sizes. This limitation complicates the attainment of definitive conclusions regarding individual-level depression diagnoses. **Data Complexity:** The datasets employed in these studies often exhibit a profusion of irrelevant features and noise. This characteristic not only augments the computational intricacy of algorithms but also compromises the predictive performance of models. Additionally, an inherent imbalance exists within the dataset, stemming from a lower representation of depression cases. **Temporal Constraints:** Extracting nuanced daily life characteristics of patients necessitates a protracted timeline. For instance, methodologies reliant on mobile devices to



Sentiment Analysis Based on Social Media - Early Stress and Depression Detection

Zixuan Li[1,3], Yuxuan Hu[1,3], Chenwei Zhang[1,3], Chengming Li[1,2(✉)], and Xiping Hu[1,2(✉)]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{lizx76,huyx55,zhangshw7}@mail2.sysu.edu.cn, {licm,huxp}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
[3] Sun Yat-sen University, Shenzhen, Guangdong, China

Abstract. Depression has recently gained significant attention as a condition marked by persistent and profound mood disturbances. Extensive research suggests that depression can influence individuals' online speech behavior, manifested through the use of depressive language and a reduction in posting frequency. Our system seamlessly integrates various sources of information, including historical tweets and user profile data. Concerning historical tweets, we propose two methods to navigate the extensive and intricate user tweet history. Our findings indicate that these methods yield more pertinent user information. Subsequently, we input this information into our meticulously constructed deep learning classification model. This model is built upon a pre-trained BERT (Bidirectional Encoder Representations from Transformers) and a bidirectional LSTM (Long Short-Term Memory) model that incorporates attention mechanisms. In the context of user information, we extract relevant details and directly incorporate them into a deep learning model based on bidirectional GRU (Gated Recurrent Unit) and MLP (Multi-Layer Perceptron). Concurrently, to address the challenge of imbalanced depression datasets, we introduce Focal Loss and Dice Loss. Our experimental results underscore the effectiveness of these loss functions in our model. To validate the efficacy of our system, we reprocess the depression tweet dataset and conduct a series of experiments. Through these experiments, we conclusively demonstrate the robustness of our model, effectively mitigating the challenge of sample imbalance to a considerable extent.

Keywords: Deep learning · Social network · Depression recognition · Data imbalance

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 26–39, 2024.
https://doi.org/10.1007/978-3-031-65126-7_3

26. Perkins, T.K., Kern, L.R.: Widths of hydraulic fractures. J. Petrol. Technol. **13**(09), 937–949 (1961)
27. Nordgren, R.P.: Propagation of a vertical hydraulic fracture. Soc. Petrol. Eng. J. **12**(04), 306–314 (1972)
28. Hudson, J.A.: A critical examination of indirect tensile strength tests for brittle rocks [Ph. D. Thesis]. University of Minnesota, Minneapolis (1984)
29. Wang, C., Wang, R., Wang, C.: Development of multiple-diameter core hydraulic fracturing machine to test tensile strength of rocks. Chin. J. Rock Mech. Eng. **36**(S1), 3321–3331 (2017)
30. Cuisiat, F.D., Haimson, B.C.: Scale effects in rock mass stressmeasurements. Int. J. Rock Mech. Min. Sci. Geomech. Abst. **29**(2), 99–117 (1992)
31. Hou, B., Chen, M., Wan, C., Sun, T.: Laboratory studies of fracture geometry in multistage hydraulic fracturing under triaxial stresses. Chem. Technol. Fuels Oils **53**(2), 219–226 (2017)
32. Park, J.Y., Tuell, G.: Conceptual design of the CZMIL data processing system (DPS): algorithms and software for fusing lidar, hyperspectral data, and digital images. Proc Spie **7695**(5), 731–739 (2010)
33. Qin, H.: Constructions of uniform designs with mixed levels. Acta Math. Appl. Sin. **28**(4), 704–712 (2005)
34. Montgomery, D.C., Peck, E.A.: Introduction to linear regression analysis (1982)
35. Schmitt, D.R., Zoback, M.D.: Diminished pore pressure in low-porosity crystalline rock under tensional failure; apparent strengthening by dilatancy. J. Geophys. Res. **97**(B1), 273–288 (1992)
36. Ito, T., Satoh, T., Kato, H.: Deep Rock Stress Measurement by Hydraulic Fracturing Method Taking Account of System Compliance Effect. Xie Furen. CRC Press, Boca Raton (2010)
37. Zhu, X., Zhang, J., Feng, J.: Multiobjective particle swarm optimization based on PAM and uniform design. Math. Probl. Eng. **2015**, 1–17 (2015)
38. Yang, L., Pan, F., Weifeng, J., Shengwei, S., Yong, Z., Tao, Z.: Predictive method of nonlinear system based on artificial neural network and svm. Oxidat. Commun. **39**(1Appa), 1226–1235 (2016)

4. Haimson, B.C.: Hydraulic Fracturing in Porous and Nonporous Rock and its Potential for Determining In Situ Stresses at Great Depth. University of Minnesota, Minneapolis (1968)
5. Haimson, B.C., Fairhurst, C.: Initiation and extension of hydraulic fractures in rocks. Soc. Petrol. Eng **9**, 310–318 (1967)
6. Von Schonfeldt, H., Fairhurst, C.: Field experiments on hydraulic fracturing. Soc. Petrol. Eng. AIME **12**(1), 69–77 (1972)
7. Wang, C.: Brief review and outlook of main estimate and measurement methods for in-situ stresses in rock mass. Geol. Bull. China **60**(5), 971–996 (2014)
8. Wang, J., Li, H., Wang, Y., Li, Y., Jiang, B., Luo, W.: A new model to predict productivity of multiple-fractured horizontal well in naturally fractured reservoirs. Math. Probl. Eng. **2015**, 1–9 (2015)
9. Xie, F., Chen, Q.: Study on the Crustal Stress Environment in China. Geological Press, Beijing (2003)
10. Jaeger, J.C., Cook, N.G.W., Zimmerman, R.W.: Fundamentals of Rock Mechanics. Blackwell Publishing, London (2007)
11. Wang, C., Song, C., Xing, B.: Compliance of drilling-rod system for hydro-fracturing in situ stress measurement and its effect on measurements at great depth. Geoscience **26**(4), 808–816 (2012)
12. Zoback, M.D., Pollard, D.D.: Hydraulic fracture propagation and the interpretation of pressure-time records for in-situ stress determinations. In: 19th US Symposium on Rock Mechanics (USRMS), pp. 14–22. American Rock Mechanics Association (1978)
13. Ito, T., Hayashi, K.: Physical background to the breakdown pressure in hydraulic fracturing tectonic stress measurements. Int. J. Rock Mech. Mining Sci. Geomech. Abst. **28**(4), 285–293 (1991)
14. Chang, C., Jo, Y., Oh, Y., Lee, T.J., Kim, K.: Hydraulic fracturing in situ stress estimations in a potential geothermal site, Seokmo Island, South Korea. Rock Mech. Rock Eng. **47**(5), 1793–1808 (2014)
15. Zhou, L., Ding, L., Guo, Q.: Experimental study of absolute rock stress measurements under different fracture media. Rock Soil Mech. **10**, 2869–2876 (2013)
16. Zhang, J.: Analysis of the Hydromechanics Factors Impact on Hydraulic Fracturing In-situ Stress Measurement. China University of Geosciences, Beijing (2018)
17. Matsunaga, I., Kobayashi, H., Sasaki, S.: Studying hydraulic fracturing mechanism by laboratory experiments with acoustic emission monitoring. Int. J. Rock Mech. Min. Sci. Geomech. Abst. **7**, 909–912 (1993)
18. Ishida, T., Chen, Q., Mizuta, Y.: Effect of injected water on hydraulic fracturing deduced from acoustic emission monitoring. In: Seismicity Associated with Mines, Reservoirs and Fluid Injections. Birkhäuser, Basel (1997)
19. Fang, K.: Uniform Design and Uniform Design Table. Science Press, Beijing (1994)
20. Wang, Z., Fang, K.: Measures of uniformity for uniform designs with qualitative factors. Math. Stat. Manag. **19**(3), 28–32 (2000)
21. Myers, R.H.: Classical and Modern Regression with Applications, 2nd edn. Duxbury Press, Belomont (1994)
22. Liu, Y., Wang, C., Wang, J., Ji, W.: Optimization research on thermal error compensation of FOG in deep mining using uniform mixed-data design method. Math. Probl. Eng. **2019**, 1–6 (2019)
23. Zhang, L., Cai, X.: Uniformity masks design method based on the shadow matrix for coating materials with different condensation characteristics. Sci. World J. **2013**, 1–4 (2013)
24. Khristianovich, S.A., Zheltov, Y.P.: Formation of vertical fractures by means of a highly viscous fluid. In: Proceedings 4th World Petroleum Congress, pp 579–586 (1955)
25. Geertsma, J., De Klerk, F.: A rapid method of predicting width and extent of hydraulically induced fractures. J. Petrol. Technol. **21**(12), 1571–1581 (1969)

experimental results. Meanwhile, neural networks and deep learning algorithms are also considered to analyze and predict rock fracturing values, verifying the accuracy of the rock fracturing model [38].

6 Conclusions

In this paper, we utilized the optimal uniform design method to optimize hydraulic fracturing simulation experiments. The results showed that this design method not only reduced the number of experiments but also improved the uniformity and test effect. It provides a fast and effective way to develop an error compensation formula for various influence elements, aiming to enhance measurement accuracy of the hydrofracture method of geostress measurement.

- The paper proposed an optimal approach for the hydrofracture method of geostress measurement using the uniform design method. Considering these unique properties of the drilling mud, which involves multiple hydraulic elements and values. This paper constructs an experimental plan that can simplify the testing procedure, and decrease implementation fees. As a result, the efficiency of the simulation hydraulic fracturing tests can be significantly enhanced.
- This study examined the impact of various hydrodynamic factors on rock fracturing pressure using test results. Through multivariate regression analysis, an optimal regression model for multiple influencing factors (fracturing fluid viscosity, density, axial load, and injection rate) was obtained. Additionally, the values of instantaneous rock splitting were discussed in depth, and the validity of Perkins-Kern-Nordgren (PKN) classical mechanical model in theoretical analysis was confirmed.

Acknowledgments. This work was supported by Project funded by National Natural Science Youth Foundation of China (41804089), and Geological survey Project of China Geological Survey (DD20230447).

Data Availability. The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest. The authors declare that there are no conflicts of interest regarding the publication of this paper, and the authors confirm that the mentioned received funding in the "Acknowledgment" section did not lead to any conflict of interests regarding the publication of this manuscript.

References

1. Clark, J.B.: A hydraulic process for increasing the productivity of wells. J. Petrol. Technol. **1**(1), 1–8 (1949)
2. Zhao, Z., Guo, J., Ma, S.: The Experimental investigation of hydraulic fracture propagation characteristics in glutenite formation. Adv. Mater. Sci. Eng. **2015**, 1–5 (2015)
3. Hubbert, K.M., Willis, D.G.: Mechanics of hydraulic fracturing. Trans. AIME **210**(1), 153–168 (1957)

To summarize, Eqs. (6) and (7) demonstrate a similar changing pattern as x_1 and x_4, but undergoes an opposite changing pattern in relation to x_2 (fracturing fluid density). The simulation experiments confirmed a strong correlation between rock fracturing pressure and the viscosity, density, and injection rate of the fracturing fluid. These findings support the conclusions of theoretical analysis in Sect. 2. This indicates a good distribution uniformity of experimental points. Additionally, this study validated the efficiency and suitability of the experimental method in establishing a fracture pressure correction formula for various hydrodynamic factors.

5 Discussion

Simulation experiments were conducted to analyze the impact of various factors such as injection rate, density, viscosity, and fracturing fluid medium on hydraulic fracturing. These experiments provide a fast and reliable way to understand the influence of hydrodynamic factors on hydraulic fracturing. A compensation model can be utilized to minimize the interference of hydrodynamic factors and enhance the accuracy of in-situ stress measurement during hydraulic fracturing in practical engineering applications. Consequently, simulation tests have the potential to improve the measurement accuracy of hydraulic fracturing methods. However, this study only considered three fluid mechanics parameters, namely injection rate, fracturing fluid density, and viscosity. Therefore, future research should explore the incorporation of additional hydrodynamic parameters such as hydraulic friction, liquid compressibility, and the effects of different types of fracturing fluid media on the effective fracturing pressure of rocks. These insights are valuable in advancing our understanding of hydraulic fracturing in practical applications.

- In terms of different fracturing fluids, such as hydraulic oil, mud, and aqueous solution, test results from Zhou Longshou (2013) [15] and Zhang Jie, Wang Chenghu et al. (2017) [16] indicate that mud and hydraulic oil lead to higher rock fracturing pressures compared to aqueous solutions. The combination of densities and viscosities of the fracturing fluids greatly affects the rock fracture pressure, while the compressibility of the fracturing fluid also influences the flexibility of the hydraulic fracturing measurement system (Wang Chenghu et al., 2012) [11], thereby affecting the measurement results. This study only considered mud as the fracturing fluid, so future studies should include more representative hydrodynamic parameters and different types of fracturing fluids for a comprehensive analysis of their influence on rock fracture pressure. Additionally, an appropriate correction formula and compensation model should be established for hydraulic fracturing errors under different working conditions.
- The experimental results confirmed that the injection rate of the fracturing fluid has a significant impact on the rock fracturing pressure, with a proportional increase. This finding aligns with the results of hydraulic fracturing tests conducted by several foreign researchers (Ito and Hayashi, 1991; Schmitt et al., 1992; Zo-back et al., 2007) [12–14, 36, 37]. To further enhance the accuracy of future simulation tests and reduce losses associated with hydraulic friction, especially head loss, it is recommended to install a high-precision pressure sensor in the fracturing test section. This enhancement will allow for a better analysis of the influence of injection rate on the

Table 4. Multi-factor VIF value

VIF	Value
x_1	1.0027
x_2	1.0009
x_3	1.0035
x_4	1.0021

polynomial. The Matlab Linear-Model.stepwise function was utilized to perform a multivariate quadratic polynomial regression of the data presented in Table 4. Based on this regression analysis, the stepwise regression method of the LinearModel class object was employed to establish the Eq. (6), to show the relationship between factors (x_1, x_2, x_3, x_4) and fracturing value (y).

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_3^2 + b_6x_1x_4 + b_7x_4^2 \tag{6}$$

Based on the stepwise regression calculation results, the resulting equation for multivariate polynomial regression can be expressed as follows:

$$y = 17.937 + 0.023x_1 - 10.266x_2 + 2.054x_3 + 1.598x_4 - 0.361x_3^2 - 0.067x_1x_4 + 18.535x_4^2 \tag{7}$$

Furthermore, Eq. (8) produces a $p\text{-}value_2 = 0.000405$, and $p\text{-}value_2 << 0.05$ (significance level). Figure 3 illustrates the regression fitting plots of Eqs. (6) and (7), demonstrating a higher degree of fit in the latter. Therefore, Eq. (7) is an optimal fitting formula for (y) and (x_1, x_2, x_3, x_4) in this design.

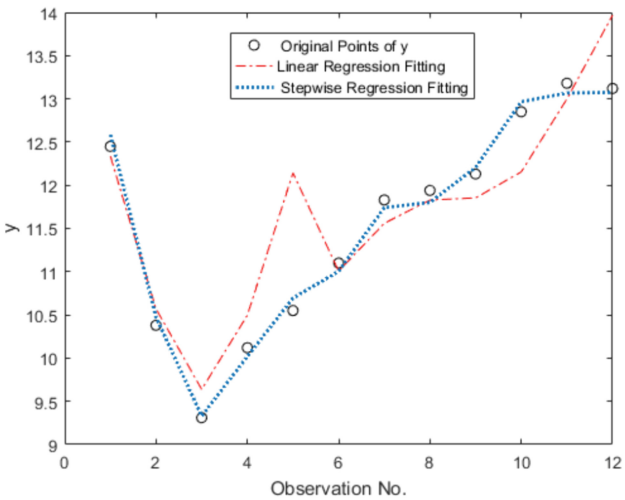


Fig. 3. The optimal fitting formula for linear regression and stepwise regression fitting

4 Results Analysis

4.1 Experimental Results

Table 3 presents results of the experiments using the uniform design method involved in Sect. 3, showcasing the obtained effective rock fracturing value.

Table 3. Table of hydraulic fracturing values of the simulation experiments

No.	Influencing factors				Results
	Viscosity(g/cm^3)	Density(mPa•s)	Axial Compression(MPa)	Injection Rate(MPa/s)	Fracturing pressure(MPa)
1	280	1.4	1.2	0.1	12.46
2	170	1.6	3.6	0.1	10.35
3	70	1.6	1.2	0.55	9.34
4	70	1.2	2.4	0.05	10.12
5	70	1.0	4.8	0.1	10.58
6	150	1.2	4.8	0.4	11.12
7	280	1.6	4.8	0.1	11.85
8	130	1.2	2.4	0.2	11.91
9	150	1.2	3.6	0.2	12.1
10	170	1.2	2.4	0.4	12.88
11	170	1.0	1.2	0.05	13.19
12	280	1.0	3.6	0.55	13.15

4.2 Multivariate Polynomial Regression

Regression analysis is a method used to establish the relationship between the dependent variable y and the independent variables (x_1, x_2,…, x_i) [34–36]. In Eq. (6), y represents actual demonstration value of the rupture pressure, x_1 represents the density, x_2 represents the viscosity, x_3 represents the axial pressure, and x_4 represents the injection speed. Table 4 underwent multiple linear regression and multivariate polynomial regression to determine the respective fitting models. These models were then compared to obtain the optimal fitting formula.

A regression model was subjected to a multicollinearity diagnosis using the variance inflation factor (VIF) method, which resulted in Table 4. Generally, if VIF < 5, there is no collinearity. The independent variables in Table 4 had VIFs below 5, indicating the absence of multicollinearity in the model.

In order to enhance the non-linear terms in the model, a stepwise regression approach was employed to conduct a generalized linear regression analysis using a quadratic

Table 1. Elements and their numerical value

Element	Level	Parameter value
Viscosity	4	70; 150; 170; 280
Density	4	1.0; 1.2; 1.4; 1.6
Axial compression	4	1.2; 2.4; 3.6; 4.8
Injection rate	6	0.05; 0.1; 0.2; 0.25; 0.4; 0.55

In order to account for the numerous elements and their values, using mud as fracturing fluid, an optimized experimental scheme based on mixed-level uniform was developed. Using the Data Processing System (DPS) software [33], a total of 12 experiments were conducted, as shown in Table 2. The constructed optimal mixed-level uniform design table $U_{12}^{*}\left(6 \times 4^{3}\right)$ was subjected to a maximum of 1000 iterations.

Table 2. Table of Influencing elements in $U_{12}^{*}\left(6 \times 4^{3}\right)$

No.	Influencing elements			
	x_1	x_2	x_3	x_4
1	4	3	1	3
2	3	4	3	3
3	1	4	1	5
4	1	3	2	1
5	1	1	4	3
6	2	3	4	6
7	4	4	4	2
8	2	2	2	4
9	2	2	3	1
10	3	2	2	6
11	3	1	1	2
12	4	1	3	5

A smaller D implies better uniformity of the experimental design [19, 20]. By calculating the Eq. (6), we obtained that $D^{*} = 0.1713$ for the $U_{12}^{*}\left(6 \times 4^{3}\right)$, which exhibits a good distribution uniformity.

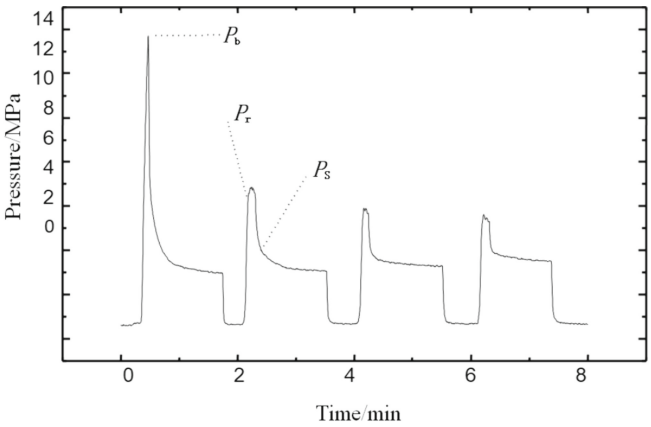


Fig. 2. Typical Pressure-Time Record Curve in Hydraulic Fracturing

and the pressure value at this time is recorded as the instantaneous closure pressure P_s. After releasing the pressure, reloading causes the fracture to reopen, and the pressure value at this time is recorded as the reopening pressure P_r.

According to the elastic theory and the PKN mechanical model, as shown in Fig. 3, the fracturing pressure of the rock in the fracturing section is:

$$P_b = 3\sigma_h - \sigma_H + T \tag{5}$$

Among them, σ_H and σ_h are the maximum and minimum horizontal principal stresses, respectively, and T is the tensile strength of the rock. The fractures induced by hydraulic fracturing are vertical fractures and perpendicular to the direction of the minimum horizontal principal stress. Equation (5) indicates that the fracturing pressure of rocks is independent of the size of the borehole and the elastic modulus of the rock, and is mainly determined by the tensile strength of the rock and the magnitude of the in-situ stress around the borehole.

3 Optimal Design of the Testing

The high pressure fluids are commonly applied in hydraulic fracturing simulation experiments, including clean water, hydraulic fluid, carboxymethyl cellulose (CMC) aqueous solution, and drilling mud [30–32]. The density and viscosity of the mud medium can be adjusted according to the requirements of the simulation experiment. For these fracturing fluid media, only a small number of factors and tests are required, so conventional comprehensive experimental methods can be used for their respective simulation experiments. In contrast, there are more parameters and theirs values in the mud medium, so an optimal design based on uniform design method is suitable for the testing.

The theoretical analysis of the PKN model revealed that when using mud as the fracturing fluid medium in the simulation test, it requires three hydrodynamic factors: density, injection rate, and viscosity, as well as a factor of loading axial compression. Different numerical values of each element are presented in Table 1.

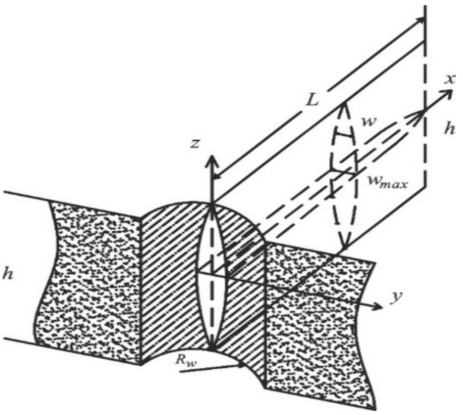


Fig. 1. PKN classical mechanical model

Here, *q(x, t)* represents the volume of fluid flowing through the cross-section of the crack, $q_l(x, t)$ represents the volume of fluid lost per unit length of the crack, and *A(x, t)* represents the cross-sectional area of the crack. When there is no fluid leakage, the length of the crack L, its local width w, and the pore pressure P_w can be calculated [28, 29]:

$$L = 0.68\left[\frac{GQ^3}{(1-\upsilon)\mu h^4}\right]^{\frac{1}{5}} t^{\frac{4}{5}} \tag{2}$$

$$w = 2.5\left[\frac{(1-\upsilon)\mu Q^2}{Gh}\right]^{\frac{1}{5}} t^{\frac{1}{5}} \tag{3}$$

$$p_w = 2.5\left[\frac{G^4\mu Q^2}{(1-\upsilon)^4 h^6}\right]^{\frac{1}{5}} t^{\frac{1}{5}} \tag{4}$$

The following variables are used in this context: G (shear modulus), υ (Poisson ratio), h (length), Q (injection rate), and μ (viscosity).

2.2 Principles of Hydraulic Fracturing Measurement

The basic principle of in-situ stress measurement based on hydraulic fracturing involves placing drill rods and packers into a borehole using a drilling rig to measure their positions. Fluid is injected into the packers through a loading control system, isolating a test section within the borehole, and the fluid is further injected into the test section until fracturing occurs.

As shown in Fig. 2, the first highest pressure value is recorded as the fracturing pressure P_b. Then the pressure drops rapidly to a state of fluid seepage into the fracture and remains constant. At this point, the pump is turned off to stop loading, and the pressure in the fracturing section decreases rapidly, causing the fracture to close quickly. When the fracture is in the near-closed state, the rate of pressure decrease slows down,

of error include: (1) the drill pipe and the packer deformation [9–11]; (2) variations in the determination method of the measurement curve during data analysis [12]; and (3) different category and the associated performance factors of the fracturing fluids [13–17].

Many researchers have dedicated themslves to explore the fracturing fluids impact on rock fracturing. For instance, Ito (1991) and Chang (2014) suggested that increasing the injection rate of fracturing fluid and considering factors like flow rate, viscosity, and density can enhance the tensile strength of the rock. Zhou et al. (2013) and Zhang (2018) conducted tests using different density mud media as fracturing fluids and observed significant variations in rock fracturing behavior. Matsagaga (1993) and Ishida et al. (1997) verified the impact of fracturing fluid viscosity on rock fracturing through oil drilling experiments. Wang (2012) and Zhou (2013) analyzed the error in stress measurement caused by the compressibility of clear water used as a fracturing fluid and its effect on system flexibility. These studies contribute to a better understanding of fluid mechanics factors in accurate rock fracturing measurements.

In summary, the hydrodynamic factors that influence rock fracturing during hydraulic fracturing include flow velocity, viscosity, density, and compressibility. Conducting simulation experiments based on these factors is crucial for understanding their impact on rock fracturing. However, these experiments can be destructive to the testing core, making them complicated and costly to design. To address this, the uniform design method has been proposed as an experimental design approach that evenly spreads test points throughout the range of variables, requiring fewer trials compared to other methods [19, 20]. In particular, the design aims to conduct trials with many experimental factors and a large number of levels, with fewer trials required compared with the orthogonal design or comprehensive design methods [21–23]. In this study, a uniform table for experiment design is used to combine selected hydrodynamic factors of the fracturing fluid with the factor of horizontal pressure. This approach reduces the test times while ensuring their effectiveness and significantly improving efficiency. The results of these experiments are then analyzed to determine the effects of the factors on rock fracturing value.

2 Error Analysis of Hydraulic Fracturing Theory

2.1 PKN Mechanical Model

The borehole used to measure hydraulic fracturing in-situ stress was typically vertical and primarily influenced by the maximum horizontal principal stress, and the minimum stress, and minimum is same as it is [24]. The fracturing crack was vertical because it was perpendicular to the minimum horizontal principal stress plane [25, 26]. The authors used the PKN classical mechanical model [27, 28] to analyze how fluid mechanics affects the fracture crack and its fracture pressure in this paper.

Figure 1 illustrates the PKN fracturing crack model [27, 28]. Nordgren (1972) obtained the fluid's continuity equation in the crack, ignoring the compression properties of the fracturing fluid [28]:

$$\frac{\partial q}{\partial x} + q_t + \frac{\partial q}{\partial t} = 0 \tag{1}$$

Optimal Design of Hydraulic Fracturing Simulation Experiments for In-Situ Stress Measurement

Yang Li[1], Daji Zhang[3], and Yimin Liu[2(✉)]

[1] Institute of Exploration Technology, CGS, Chengdu 611734, China
[2] Chengdu Technological University, Chengdu 611730, China
153973418@qq.com
[3] Chengdu Rail Transit Group Co., Ltd., Chengdu 610036, China

Abstract. In order to account for a large number of hydrodynamic influencing factors with multiple levels in rock fracturing experiments, the uniform design method is frequently utilized instead of conventional methods like comprehensive and orthogonal designs, as they significantly impact the experimental effects. Based on the Perkins-Kern-Nordgren (PKN) model, the influencing factors of injection rate, viscosity, and density of the fracturing fluid, along with their corresponding parameter values or levels, were taken into consideration to construct an optimal table $U_{12}^{*}\left(6 \times 4^{3}\right)$ for experiment design. Subsequently, an optimized experimental scheme was developed. The experimental results based on this design were analyzed using multiple regression analysis to establish an optimal regression equation for the influencing factors (x_1, x_2, x_3, x_4, representing fluid viscosity, density, loading axial compression, and injection rate, respectively) and to determine the corresponding rock fracturing value (y). This indicates a good distribution uniformity of experimental points. Additionally, this study validated the efficiency and suitability of the experimental method in establishing a fracture pressure correction formula for various hydrodynamic factors, and it is also a precise approach for geostress measurements.

Keywords: Uniform design method · Mixed-level · In-situ stress measurement · Rock fracturing

1 Introduction

The stress stored in the interior of a rock mass without disturbance is referred to as in-situ stress, which has multiple sources and is influenced by various factors, resulting in a complex and variable distribution of stress in the Earth's crust [1]. Hydraulic fracturing is a crucial technique for measuring in-situ stress in various geological structures such as hydropower stations, tunnels, chambers et al. [2–4]. This approach offers an efficient test procedure, along with straightforward data analysis procedures. However, several factors can influence the accuracy of rock fracturing measurements [5–7]. The primary sources

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 15–25, 2024.
https://doi.org/10.1007/978-3-031-65126-7_2

33. Xia, R., Liu, Y.: A multi-task learning framework for emotion recognition using 2d continuous space. IEEE Trans. Affect. Comput. **8**(1), 3–14 (2017). https://doi.org/10.1109/TAFFC.2015.2512598
34. Xie, B.: Research on key technology of Mandarin speech emotion recognition. Ph.D. thesis, Zhejiang University (2006)
35. Xu, Y., Su, H., Ma, G., Liu, X.: A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context. Complex Intell. Syst. **9**(1), 951–963 (2023)
36. Yu, W., et al.: Ch-sims: a chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3718–3727 (2020)

16. Li, Y., Zhao, T., Kawahara, T., et al.: Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In: Interspeech, pp. 2803–2807 (2019)
17. Liu, A.T., Yang, S.W., Chi, P.H., Hsu, P.C., Lee, H.V.: Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6419–6423. IEEE (2020)
18. Liu, Y., et al.: Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and AV-Mixup consistent module. In: Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 247–258 (2022)
19. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north American English. PLoS ONE **13**(5), e0196391 (2018)
20. Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In: Ismir, vol. 270, p. 11. Plymouth, MA (2000)
21. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F.: Multimodal emotion recognition on ravdess dataset using transfer learning. Sensors **21**(22), 7665 (2021)
22. McFee, B., et al.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25 (2015)
23. Nwe, T.L., Foo, S.W., De Silva, L.C.: Classification of stress in speech using linear and nonlinear features. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP 2003), vol. 2, pp. II–9. IEEE (2003)
24. Oh, K.J., Lee, D., Ko, B., Choi, H.J.: A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In: 2017 18th IEEE International Conference on Mobile Data Management (MDM), pp. 371–375. IEEE (2017)
25. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
26. Pravena, D., Govind, D.: Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. Int. J. Speech Technol. **20**(4), 787–797 (2017)
27. Singh, A., Liu, H., Plumbley, M.D.: E-panns: sound recognition using efficient pre-trained audio neural networks. arXiv preprint arXiv:2305.18665 (2023)
28. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21–25. IEEE (2021)
29. Triantafyllopoulos, A., Schuller, B.W.: The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7268–7272. IEEE (2021)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. **30** (2017)
31. Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., Ma, C., Huang, Y.: Video sentiment analysis with bimodal information-augmented multi-head attention. Knowl.-Based Syst. **235**, 107676 (2022)
32. Xi, Y., Li, P., Song, Y., Jiang, Y., Dai, L.: Speaker to emotion: domain adaptation for speech emotion recognition with residual adapters. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 513–518. IEEE (2019)

5 Conclusions

In this paper, we effectively applied transfer learning to fine-tune the English pre-trained model, achieving a notable improvement in the F1 score to 0.46, significantly surpassing the baseline of 24%. For future research, we will continue exploring the cross-corpus SER domain and further investigating other deep learning techniques to enhance the performance of the transfer learning models in emotion recognition.

References

1. Amiriparian, S., et al.: Snore sound classification using image-based deep spectrum features (2017)
2. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
3. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
4. Gemmeke, J.F., et al.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 776–780. IEEE (2017)
5. Gong, Y., Chung, Y.A., Glass, J.: Ast: audio spectrogram transformer. arXiv preprint arXiv:2104.01778 (2021)
6. Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: Ssast: self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 10699–10709 (2022)
7. Haq, S., Jackson, P.J.: Multimodal emotion recognition. In: Machine Audition: Principles, Algorithms and Systems, pp. 398–423. IGI Global (2011)
8. Hossain, M.S., Muhammad, G., Song, B., Hassan, M.M., Alelaiwi, A., Alamri, A.: Audio-visual emotion-aware cloud gaming framework. IEEE Trans. Circuits Syst. Video Technol. **25**(12), 2105–2118 (2015)
9. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 801–804 (2014)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894 (2020)
12. Koolagudi, S.G., Murthy, Y.S., Bhaskar, S.P.: Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. Int. J. Speech Technol. **21**(1), 167–183 (2018)
13. Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., Stober, S.: Transfer learning for speech recognition on a budget. arXiv preprint arXiv:1706.00290 (2017)
14. Lee, M.C., Chiang, S.Y., Yeh, S.C., Wen, T.F.: Study on emotion recognition and companion chatbot using deep neural network. Multimedia Tools Appl. **79**, 19629–19657 (2020)
15. Li, P., Song, Y., McLoughlin, I.V., Guo, W., Dai, L.R.: An attention pooling based representation learning method for speech emotion recognition (2018)

Table 1. Quantitative evaluation of the different strategies on speech emotion recognition. In bold, the best model.

Classification	train F1 score	val F1 score	train Acc	Val Acc
original frozen CNN10(baseline)	0.2760	0.3125	0.3658	0.3994
original frozen CNN14(baseline)	0.2528	0.2413	0.3644	0.3778
original CNN10	0.6612	0.4432	0.6831	0.4657
original CNN14	0.8822	0.4542	0.8901	**0.4774**
multihead CNN10	0.6297	0.4536	0.6576	0.4684
multihead CNN14	0.9144	0.4642	0.9208	0.4674
multilayer CNN10	**0.7694**	**0.4636**	**0.7871**	0.447
multilayer CNN14	**0.9516**	**0.4652**	**0.9613**	**0.4722**

also observed a large performance gain for valence and a lesser gain for other aspects. The results suggest that while fine-tuning does incur additional computational costs, the benefits it yields in terms of improved performance make it a worthwhile endeavor. The validation set F1 scores for both CNN10 and CNN14 models, when employing the multilayer multi-head attention module, surpass those of the baseline, with the CNN14 model also demonstrating higher accuracy on the validation set. A comparison between different structures reveals that the multilayer multi-head attention modules generally outperform their single-layer counterparts. Specifically, the 'multilayer CNN14' model delivered the best results, achieving optimal performance with the least amount of epochs.

4.4 Future Work

Compared to the baseline, we believe there is ample room to improve the accuracy of the validation set. There are several areas for future improvements. First, we did not adjust the architecture of the pre-trained model, and the limited number of CNN layers may have hindered its ability to recognize emotions arising from emotional correlations in the data fully. Thus, further adjustments to the model architecture and hyperparameters are necessary for better generalization. In addition, we should further explore the linguistic and cultural differences in the datasets. Our target dataset is in Mandarin Chinese, while the baseline dataset is in English. Cross-language disparities may impede significant performance improvements.

Revealing these potential differences between languages and cultures requires further research in multi-task learning and exploring the fields of language and cultural studies. These areas offer significant potential for future research efforts, helping bridge the cross-linguistic gap and improving the performance of deep learning algorithms in specific tasks.

attention layer is passed into the next layer, it first goes through an additional transformation via a fully connected layer. The potential benefit of this could be to provide an additional means to capture and transform more complex patterns in the data.

4 Experiment

4.1 Dataset Setup

In our experiments, we utilized the CH-SIMS v2.0 [18] dataset, which is partitioned into three sets: the training set (80%), the test set (10%), and validation set (10%). The obtained output is categorized into five distinct labels. For feature extraction, the librosa library [22] is employed to extract log mel spectrograms from raw audio data.

4.2 Experimental Setting

In the training experiments, we leverage a pre-trained model on the AudioSet dataset to facilitate transfer learning on an existing dataset. During the fine-tuning phase, we employed both single and triple multi-head self-attentive layers, with the results being labeled as 'multihead' and 'multilayer' respectively. The training process was utilized the Adam optimizer [10] and cross-entropy loss with a batch size of 16.

Results from the two original models (CNN10 and CNN14), with frozen parameters, were served as the baseline for our benchmark. In the fine-tuning phase, the models with multi-head and multi-layer were trained for 200 epochs with an initial learning rate of 1e-4. Each experiment set were conducted ten times with the average results recorded. The best-performing model is selected and saved, conducting experiments on both test sets and validation sets. The recorded results are presented in Table 1.

4.3 Results and Discussion

Table 1 presents the results of experiments conducted on the CH-SIMS2.0 dataset, with the primary evaluation metrics being the F1 score and accuracy (Acc). Remarkably, the fine-tuned models consistently outperform their counterparts with frozen parameters. When compared to other models, our approach delivers highly competitive results. The findings indicate that fine-tuning of parameters significantly enhances the accuracy of audio classification. Therefore, we firmly advocate for implementing parameter fine-tuning as an effective strategy to elevate output performance.

Table 1 provides a summary of the performance exhibited by the various speech emotion recognition models that were tested. When considering the experiments with the frozen initial parameters as the baseline, improvements are observed across all tested results in comparison to the baseline. Notably, we

every convolutional layer, and then ReLU non-linearity is applied to facilitate better training convergence. For CNN10 and CNN14, the convolutional blocks are used in pairs before an average pooling layer is applied. Specifically, CNN10 is composed of 8 convolutional blocks (4 pairs), while CNN14 consists of 12 convolutional blocks (6 pairs). All networks include a penultimate fully connected layer to enhance representation capability, succeeded by a final fully connected layer with 527 units. A sigmoid activation function is applied at this stage to derive the probabilities of each class.

3.3 Multi-head Attention Block

To capture the semantic relevance embedded within the speech signal, the multi-head self-attention mechanism [30] is employed to focus on emotional information from various subspaces. In the multi-head attention mechanism, there are H parallel attention heads, and each of these attention heads calculates a set of attention weights:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q^\top \cdot K}{\sqrt{d_k}}\right) \cdot V^\top \tag{1}$$

where: Q, K, and V are the query, key, and value matrices for calculating the multi-attention mechanism. The Softmax function is commonly used to normalize the attention scores and ensure that they represent a valid probability distribution, where the sum of all attention weights is equal to 1.

We use the optimization algorithm Noam for learning rate tuning to achieve better model solutions. By computing similarities between Q and K, the mechanism assigns weights to each query position, determining the significance of the corresponding values.

$$lr = factor \cdot modelsize^{-0.5} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5}) \tag{2}$$

where: *factor* refers to the initial learning rate size, *model size* denotes the hidden layer dimension, *step* represents the number of optimization steps, and *warmup* denotes the value of the step when the learning rate reaches its maximum value.

Between each layer, we introduced a gate function incorporating the sigmoid function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

x is the input variable. It converts the output of the model into a probability value between 0 and 1. This gated mechanism helps the model to dynamically adjust the importance of different layers and enables more flexible and adaptive information processing. Furthermore, the use of the sigmoid function ensures a smooth gating operation, avoiding abrupt changes and maintaining stability during the learning process.

After applying several layers of multi-head attention, we performed an fully connected layer (comprised of a linear layer and a ReLU activation function) following each gating mechanism. This means that before the output from each

3.1 CH-SIMS Dataset

To conduct Speech Emotion Recognition (SER) on Chinese speech, we utilized the CH-SIMS v2 training dataset [18]. This dataset comprises 60 original videos, resulting in 2,121 video segments. It offers a diverse range of character backgrounds, covering different age groups, and featuring high-quality recordings. Only Mandarin Chinese speech is included in this dataset.

Compared to version 1.0 [36], the video data in this updated version includes a broader range of scenarios, and the focus is on acoustic and visual features rather than text, encompassing a wider variety of emotional expressions. This aspect serves as a valuable inspiration for our research.

Each video segment in the CH-SIMS v2 dataset has undergone multimodal annotations, further categorized into five emotion categories:

$$\begin{aligned} \text{negative} &: \{-1.0, -0.8\}, \\ \text{weakly negative} &: \{-0.6, -0.4, -0.2\}, \\ \text{neutral} &: \{0.0\}, \\ \text{weakly positive} &: \{0.2, 0.4, 0.6\}, \\ \text{positive} &: \{0.8, 1.0\}. \end{aligned}$$

3.2 Pre-trained Block

Our approach aims to leverage a pre-trained speech recognition network to extract meaningful features from the samples of CH-SIMS. The CNN architectures utilized in our study are adapted from those presented in reference [11]. The PANNs framework houses a diverse of pre-trained models, encompassing various versions of CNN models. These models are trained on extensive audio datasets, empowering them with ability to capture intricate audio feature representations. This capability allows PANNs to efficiently capture and analyze patterns and recognizable features within audio data. We applied its subsample since, within PANNs [11], the CNN-14 model achieves the best performance, and also uses the pre-trained model corresponding here. Following the preprocessing phase, the vocal data is fed into the framework which then internally constructs a frequency-based representation of the recordings. Interestingly, in a related study [27], it was observed that CNN-10 model performs well with some smaller datasets. Consequently, in our experiments, we employed both CNN-10 and CNN-14 models for the feature extraction and embedding.

The audio data undergoes the following preprocessing steps: first, the audio is resampled to 32kHz. Then, a Short-Time Fourier Transform (STFT) is applied with a window size of 1024 frames and a hop size of 320 frames. This process is to obtain spectrograms from the standard time-domain waveforms. Subsequently, Mel filter banks are utilized to the obtained spectrograms. After this, a logarithm operation is performed to derive log Mel spectrograms.

Each of CNN architectures is composed of convolutional layers with a kernel size of 3×3 for CNN10 and CNN14. Batch normalization is applied after

In deep learning research, various studies have adopted transfer learning methodologies, using techniques like embedding extraction and fine-tuning of pre-existing models [13,21], instead of training models from scratch. Both PANNs [11] and DeepSpectrum [1] are highly influential modern libraries designed for audio-based tasks. Among them, PANNs introduce pre-trained audio neural networks for sound event detection. The ability to fix hyperparameters in PANNs provides flexibility to use it as a transfer learning module with pre-existing knowledge. Singh et al. [27] simplified the original PANNs model using a pruning algorithm to remove redundant parameters and reduce computational effort.

To reduce the computational cost, researchers often use pre-trained models with fixed parameters to extract features, and training output layers on the generated embeddings. However, fine-tuning certain layers has been found necessary for specific tasks to achieve excellent performance [11,17]. Earlier layers in convolutional neural networks (CNNs) generally have stronger generalization capabilities than subsequent layers [29], explaining why fine-tuning all layers is essential for achieving good performance. In our study, we aim to explore whether a similar operation is necessary for the model under consideration. We will conduct two experimental designs, freezing the parameters of pre-trained layers or fine-tuning all output layers, to compare the effects of these approaches empirically.

3 Methodology

In our proposed architecture, we have designed two key modules: the pre-trained block and the multi-head attention block. The pre-trained block is a convolutional neural network (CNN) model that we have encapsulated within the PANNs [11] framework. The system's overall structure and the interconnections between these modules are depicted in Fig. 1. In this section, we provide comprehensive explanations of the datasets utilized and the specific application strategies employed for each module.

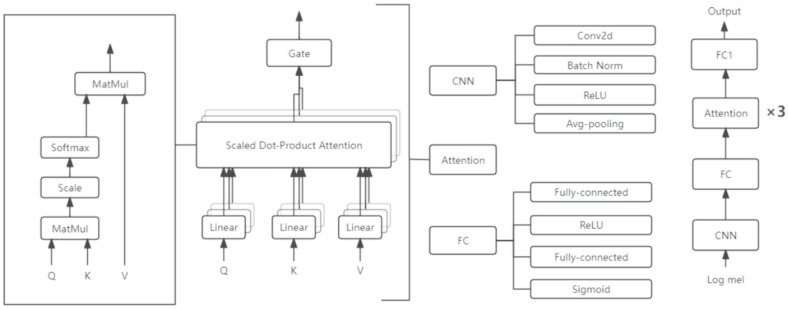


Fig. 1. The structure of proposed multi-head attention block. The number of attention layers will be adjusted to the specific task. 1 and 3 were applied in our experiments.

models, deep learning models do not require handcrafted features but directly learn feature representations from raw data, making them more powerful for processing complex and large-scale datasets and often exhibiting superior performance in specific tasks.

Xu et al. [35] proposed a framework for dual-modal (audio-text) emotion recognition. The framework consists of a parallel convolution module (Pconv) and an attention-based BLSTM [30], with a specific focus on single-modal processing of audio data from the CH-SIMS dataset. By combining Pconv and attention-based BLSTM, the Tensor Fusion Network effectively captures the complementary information from audio and text modalities, enabling more powerful multimodal sentiment analysis. The multiple self-attention mechanism is also a method of sentiment analysis that can enhance modal information [31]. In this paper, we apply transfer learning to a pre-trained CNN model with a multi-head attention mechanism and evaluate the performance of the system in terms of classification accuracy and training time.

2.2 Transformer

The Transformer model possesses several advantages, including its ability to effectively handle long sequences, capture long-range dependencies, and its parallel computing capabilities, making it highly suitable for processing large-scale data. Initially, the transformer model was mainly used in the field of machine translation, but because of its properties, it has gradually been generalized to the field of audio recognition.

In 2015, Chorowski [3] proposed to utilize an attention-based architecture, where the encoder side is a BiRNN structure. This was followed by a study on how transformers can replace RNNs for computation. The combination of CNN and attention mechanism is also a trend in audio emotion recognition, and the self-attention mechanism can express the salient regions of emotion in audio very well [16]. In 2021, Gong et al. Li et al. [15] proposed an Attention pooling method to avoid overfitting of convolutional features input to the fully connected layer. [5]introduced the Audio Spectrogram Transformer (AST), an audio classification model that canceled CNNs. Applying the Transformer encoder output to an audio spectrogram representation. They then proposed a semi-supervised framework [6] that improved the performance of AST by an average of 60.9%.

2.3 Transfer Learning

Transfer learning leveraging knowledge and models learned from one task to improve performance on another related task, reducing the need for extensive training data. It can effectively bypass the time-consuming task of data tagging when discrepancies exist in the feature space or data distribution [25], significantly increasing data mining efficiency. Transfer learning is crucial for multilingual or cross-lingual datasets due to the correlation between languages and speech, enabling the discovery of implicit connections parallelization [28].

collection of data covering Chinese text, images, audio data, and detailed annotations of modality.

In our study, we employed Pretrained Audio Neural Networks (PANNs) that were trained on the comprehensive AudioSet dataset. PANN is a deep learning model architecture crafted for audio data processing, built on the convolutional neural network (CNN) structure. Through fine-tuning on our unique dataset and integrating a multi-head self-attention mechanism, PANNs became more attuned to the specific features of the task and emotional nuances present in speech data, leading to enhanced emotion recognition performance. Our primary contributions include:

- We fine-tuned the pre-trained model on the AudioSet dataset and applied it to CH-SIMS for data preprocessing, yielding results with remarkable generalization capabilities.
- We introduced an architecture that merges CNN with a multi-head attention mechanism, enhancing the model's downstream performance.

2 Related Works

2.1 Speech Emotion Recognition

Over the past nearly three decades, researchers have tried to give machines the ability to understand and express emotions. Currently, the mainstream emotion recognition methods are extracting features that can accurately express emotions and detecting them, either manually or with the help of machines. This field encompasses a wide range of literature and utilizes various English datasets, such as RAVDESS [19], SAVEE [7], and IEMOCAP [2]. AudioSet [4] records a collection of 10-second sound clips including 632 audio event classes and over two million human-tagged clips drawn from YouTube videos. For Chinese language datasets, CH-SIMS [18] is notably prevalent, offering sentiment labels such as Strong Negative, Weak Negative, Neutral, Weak Positive, and Strong Positive. This study contributes to advancing multimodal sentiment analysis and capturing richer representations of sentiment within Chinese language data.

Emotion detection of sound relies on the integration of classical machine learning methods and deep learning techniques. Acoustic features, such as loudness, pitch, and timbre, are extracted and utilized in the algorithm to achieve accurate emotion detection. Spectral features, including Mel Frequency Cepstral Coefficients (MFCC) and their associated features, are also widely used [20]. The demarcation between machine learning and deep learning methodologies primarily resides in their respective approaches to data representations. In machine learning, a set of values is extracted from temporal, frequency, and perceptual domains and then fed into the machine learning algorithm as manually selected or predefined features to establish patterns and relationships for tasks like classification or regression. On the other hand, deep learning employs more complex and elusive algorithms, for example, CNN and attention mechanisms, to automatically learn intricate correlations within data. Unlike the traditional

speeches, such as happiness, sadness, and more. Emotion recognition systems leverage machine learning and deep learning techniques to extract relevant features from speech data, enabling accurate classification of emotions. High-performance SER systems hold significant value across various domains, including human-machine interaction [24], voice assistants [14], and psychological research [8]. They not only help computers better recognize the emotional states of inter-actor, but also pave the way for more personalized and effective human-computer interactions. Advancing SER is one of key objectives in emotion recognition system research. To improve accuracy, researchers employ techniques such as data augmentation and transfer learning, complemented by the use of larger and more diverse speech datasets. These strategies aid in training models proficient at accurately capturing and identifying emotional cues from speech data.

In the task of SER, the objective is to correlate input speech signals with specific emotion categories, thereby determining the underlying expressed emotions. Traditional classification techniques usually rely on probabilistic models, such as the Gaussian mixture model (GMM) [12], hidden Markov model (HMM) [23], and support vector machine (SVM) [26]. However, with the progression of research, various artificial neural network architectures have also been widely utilized, ranging from the simplest multilayer perceptron (MLP) [33], convolutional neural networks (CNNs) [9], to deep architectures like residual neural networks (ResNets) [32] and recurrent neural networks (RNNs) [17] [18]. Particularly, long short-term memory (LSTM)and gated recurrent units (GRU)-based neural networks, which are state-of-the-art solutions in time-sequence modeling, have been ubiquitously applied in speech signal modeling. Additionally, researchers have also proposed various end-to-end architectures aiming to jointly learn both feature extraction and classification [16]. These architectures intensively optimize the identification and association of emotions in speech signals, enhancing the overall performance of SER systems.

SER in Chinese involves identifying and analyzing emotions in Chinese speech data. Chinese-specific speech datasets are used to create diverse databases covering various emotional states. Techniques such as sound signal processing and feature extraction are employed to capture emotion-related features from speech. Machine learning algorithms, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are used for emotion classification. Recent advancements like transfer learning and data augmentation have shown promising results. [34]

The attention mechanism imitates human attention, selectively focusing on different parts of input data and assigning varying levels of importance. Self-attention, used for sequential data, treats each input element as a query, key, and value. Multi-head self-attention extends this concept by introducing multiple attention heads, enabling the model to capture diverse feature representations and enhancing its expressive power.

Our research focuses on Speech Emotion Recognition (SER) in Chinese. We leveraged the CH-SIMS dataset for our study, which provides a comprehensive

Transfer Learning for Audio-Based Speech Emotion Recognition in Chinese: Leveraging Pretrained Models for Improved Performance

Lanke Zhu[1,2], Xinyue Ma[1,2], Rui Zhang[1,2], and Jianbo Zheng[1,2(✉)]

[1] Artificial Intelligence Research Institute, Shenzhen MSU -BIT University, Shenzhen 518172, Guangdong, China
jianbo.zheng@smbu.edu.cn

[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU -BIT University, Shenzhen 518172, Guangdong, China

Abstract. In the field of Speech Emotion Recognition (SER) research, there is a growing emphasis on strengthening model generalization, stepping beyond the traditional classification accuracy metrics. Recent progress in cross-corpus SER has allowed machines to explore relationships among languages from diverse regions. In this paper, we propose an audio emotion recognition model which leverages a pretrained CNN model with a multi-head attention block. To adapt the model for the Chinese dataset CH-SIMS employed in our experiments, we fine-tuned it from a pre-trained English model. The data are categorized into five valence states: negative, weakly negative, neutral, weakly positive and positive. Remarkably, our top-performing model (multi-layer-CNN14) achieves a 24% improvement in accuracy over the baseline. The results highlight the effectiveness of fine-tuning in enhancing speech emotion recognition performance. This study contributes to improving model generalization in transfer learning, nudging us toward a deeper understanding and more accurate recognition of emotions expressed in speech.

Keywords: speech emotion recognition · transfer learning · fine-tuning · attention mechanism · Pretrained audio neural network

1 Introduction

Speech Emotion Recognition(SER) is a vital task in Natural Language Processing (NLP). It aims to detect and recognize the emotions conveyed through

L. Zhu, X. Ma, R. Zhang—These authors contributed equally to this work.

J. Zheng—This work was supported in part by the Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 3–14, 2024.
https://doi.org/10.1007/978-3-031-65126-7_1

E-Health Networks I

Autonomous Vehicles

Efficient Joint Deployment of Multi-UAVs for Target Tracking 409
Jiashuai Wang, Lu Sun, Liangtian Wan, Jibin Zheng, and Xianpeng Wang

Joint User Scheduling and UAV Height Control for Smart Wearable Device Charging Network 422
Hongjing Ji, Xiaojie Wang, and Zhaolong Ning

Studies on Vehicle Object Detection and Tracking in UAV Aerial Data 431
Ting Cao, Xinrong Zhang, Penghui Wang, and Chenle Wang

Task Prediction Based Computation Offloading over Multi-UAV MEC Network 438
Xi Cheng, Zhenquan Qin, Ruixin Liu, Jiong Lu, and Jianbo Zheng

TraMap: SLAM-Based Trajectory Generation and Optimization for Emergency Scenarios 453
Yuqing Sun, Lei Wang, Sunhaoran Jin, Jian Fang, and Bingxian Lu

Bandwidth Resource Allocation and Uplink Optimization in MEC System Based on Multi-UAV Collaboration 471
Na Yu and Xuehe Wang

Visible Light Two-Way Communication Method for Vehicle-Road Collaboration 484
Caipeng Gu, Jijing Cai, Meilei Lv, Jiefan Qiu, Chenzhuo Jin, and Kai Fang

Author Index 495

Dual-Branch Differentiated Similarity Network for Semi-supervised Medical Image Segmentation 264
Weixian Yang, Jing Lin, Wentian Cai, and Ying Gao

PIDNet: Prohibited Items Detection Network and Fine-Coarse Encoder Module 279
Yu Yao, Boliang Zhang, H. K. Kan, and Chan Tong Lam

Multiple People Tracking Based on Improved SiameseFC Combined with Lightweight YOLO-V4 291
Lu Shen, Zhiwen Chen, Boliang Zhang, Su-Kit Tang, and Silvia Mirri

Research on Image Stitching for Parking Assistance System 306
Sheng Liu, Yiqing Yang, and Ting Cao

Unsupervised Multi-source Adaptive Pedestrian Re-recognition: Based on Target Domain Prioritization and Multi-dimensional Edge Features 315
Jia He, Xiaofeng Zhang, Tong Xu, Mingchao Zhu, Kejun Wang, Pengsheng Li, and Xia Liu

Proactive Hybrid Autoscaling for Container-Based Edge Applications in Kubernetes 330
Kaile Zhu, Shihao Shen, Shizhan Lan, Xiaofei Wang, Cheng Zhang, Chao Qiu, and Victor Leung

Hybrid Platoon Control Based on Driving Characteristics 346
Jingpeng Hu and Zhiguo Xiong

Solving Traveling Salesman Problem with Deep Reinforcement Learning and Knowledge Distillation 360
Xiaowen Li, Xiaofeng Gao, Shaoyao Niu, Wenxuan He, Wanru Gao, and Qidong Liu

Wireless Parallel Reinforcement Learning: An Actor-Critic Approach 375
Ke Xing, Xinyue Ma, and Yanjie Dong

Research on a High Efficiency Work Flow for Automated Mail Sending 385
Zhongbing Tan and Huibin Luo

Artificial Neural Network Approach for Estimating Operating Parameters for Predictive Maintenance of Hydraulic Circuit 391
Ivan Kuric, Daria Fedorova, Ivan Zajačko, Vladimír Tlach, Vladimír Stenchlák, and Andrej Bencel

A Multi-factor Water Quality Prediction Method Based on Wavelet Transform and LSTM 130
Mingxia Yang, Lianghuai Tong, Aiping Xia, and Kai Fang

Quality of Service (QoS) and Quality of Experience (QoE)

Optimizing Computing Job Scheduling and Path Planning with Multi-objectives 147
Haoran Song, Chengxiao Yu, Kang Liu, Deyun Gao, Xuening Shang, and Hanxiao Yan

Research on Task Scheduling Algorithms for Cloud-Edge Collaboration 158
Shuai Lu, Haibo Zhou, Shuaishuai Zhao, Wangbei Xu, and Kai Fang

Stress Analysis of Welding Seam of Throttling Flowmeter Used in Power Plant Boiler 167
Lianghuai Tong, Chengwei Huang, Yuliang Zhang, Fan Hua, Aiping Xia, and Wen Zhou

An Improved Model for Sap Flow Prediction Based on Linear Trend Decomposition 179
Bo Li, Yane Li, Hailin Feng, Bin Wu, Qiang Zhu, Xiang Weng, and Yaoping Ruan

Prediction of the Short-Term PM2.5 Concentration Based on Informer 197
Jijing Cai, Chen Wang, Le Yu, Meilei Lv, and Kai Fang

Monte Carlo Reinforcement Learning for Cooperative Spectrum Sensing in Decision Fusion 211
Qingying Wu, Benjamin K. Ng, Han Zhu, and Chan-Tong Lam

Artificial Intelligence and Machine Learning II

Research on Fine-Grained Classification of Small Sample Marine Organism Images 229
Huibin Luo and Zixin Lin

A Dual Attention-Based Task Offloading Approach in Computing Power Networks for Object Detection 244
Kang Huang, Chao Qiu, Hong Zhu, Lisha Gao, Qizhe Zhang, Guozheng Peng, Nan Xiang, and Xiaofei Wang

Contents – Part II

Robustness

Research and Design of Hidden Trouble Target Reconfirmation and Repeated Hidden Trouble Target Filtering Technology in Transmission Line Online Monitoring 3
Yi Yang, Zhengheng Li, Nanhao Liu, Yu Su, Xifeng Yan, and Huabo Tao

UCAT: User Centric Adaptive Transmission for Meeting Diverse Network Demands 15
Hanxiao Yan, Chengxiao Yu, Kang Liu, Deyun Gao, Du Chen, and Haoran Song

A New Rolling Bearing Work Condition Monitoring Method Based on Back Propagation Network 26
Qilu Wu, Yuxi Chen, and Wenxin Hu

Vectorized Colorization of Icon Line Art Based on Closed Contour Extraction 38
Ning Wang, Sen Ning, Yifei She, Bin Liu, Haojie Li, and Zhihui Wang

Adaptive Control Scheme for Clustering of Nodes Based on the Signs of Connections in Dynamical Signed Networks 54
Qi Wang, Yinhe Wang, Zilin Gao, Peitao Gao, Jianbin Xiong, Jian Cen, and Ying Gao

Network Security and Privacy

Entrofuse: Clustered Federated Learning Through Entropy Approach 79
Kaifei Tu, Wenhao Yuan, and Xuehe Wang

M2F: Multi-centered Fairness-Aware Federated Learning Framework 95
Jing Deng, Handi Chen, Yunhin Chan, and Edith Ngai

Federated Learning Optimization Algorithm Based on Dynamic Client Scale 109
Luya Wang, Wenliang Feng, and Ruoheng Luo

Research on Domain Specific Chinese Named Entity Recognition Based on RTBC Algorithm 118
Xiaohua Ke, Xiaobo Wu, Zexian Ou, and Binglong Li

Multi-objective Deployment of WSNs in Underground Sheltered Space 521
Liangtian Wan, Caiyun Wang, Lu Sun, Boyu Chen, Jibin Zheng, and Xianpeng Wang

Author Index 535

Short Text Data Mining Based on Incremental AP Clustering 363
Fuyu Lu, Ying Guo, Peiyi Qu, and Yonglin Leng

A Novel Method for Semantic Segmentation on Lidar Point Clouds 374
Fei Wang, Liangtian Wan, Yan Zhu, Lu Sun, Xiaowei Zhao, Jianbo Zheng, and Xianpeng Wang

Forward Secure Searchable Encryption over Medical Cloud Data 384
Yang Mi, Cheng Guo, Qianqian He, and Xinyu Tang

Wireless Network Topology Discovery Based on Spectrum Data by Convolutional Neural Network 398
Xinfeng Deng, Zhihui Xie, and Li Zhou

Semi-Supervised Learning Based Trust Evaluation for Underwater Wireless Sensor Networks 411
Weicheng Meng, Zhenquan Qin, Yuxin Cui, Hao Lu, Bingxian Lu, and Jianbo Zheng

Wireless Charging Based Sensor Network Information Collection Through Unmanned Aerial Vehicles (UAVs) 426
Guoxin Xu, Jiawen Zhao, and Xuehe Wang

Joint Symbol-Level Precoding and Reflecting Design for Heterogeneous Networks with Intelligent Reflecting Surface 441
Haoran Pang, Fei Ji, and Miaowen Wen

SOH Prediction in Li-ion Battery Energy Storage System in Power Energy Network 457
Xiaofen Fang, Kai Fang, Lihui Zheng, Han Zhu, Qichang Zhuo, and Jianqing Li

Load Balancing in Software-Defined Networks Based on Particle Swarm Optimization 472
Haiyan Zhang, Liren Zou, and Yilong Xie

An Improved Genetic Algorithm for College Course Scheduling 481
Chenle Wang and Bin Wang

KNN-Based Collaborative Filtering for Fine-Grained Intelligent Grad-School Recommendation System 494
Jinfeng Xu, Jiyi Liu, Zixiao Ma, Yuyang Wang, Wei Wang, and Edith Ngai

RPBFT: A Scalable Consensus Mechanism for Large Blockchain Systems 509
Weizhe Wang, Daxin Tian, Xuting Duan, and Jianshan Zhou

Identifiable EEG Embeddings by Contrastive Learning from Differential Entropy Features 227
Zhen Zhang, Feng Liang, Jiawei Mo, and Wenxin Hu

Contrastive Learning Consistent and Identifiable Latent Embeddings for EEG 236
Feng Liang, Zhen Zhang, Jiawei Mo, and Wenxin Hu

SEVGGNet-LSTM: A Fused Deep Learning Model for ECG Classification 245
Tongyue He, Yiming Chen, Bo Fang, and Junxin Chen

Fast Convergence Federated Learning with Adaptive Gradient: An Application to Mental Healthcare Monitoring System 255
Junqiao Fan, Xuehe Wang, and Yuzhu Hu

Artificial Intelligence and Machine Learning I

Research on Handover Technology for 5G LEO Satellite Network Based on ns-3 279
Zheng Wang, Li Zhou, and Yankun Wang

Joint Delay and Energy Optimization for WPT-MEC System Based on Immune Algorithm 290
Lu Sun, Dianju Li, Hao Lu, Liangtian Wan, Jianbo Zheng, and Xianpeng Wang

An Abnormal Detection Method Based on the Device Interaction Behavior in the Internet of Things 302
Wenjing Jin, Xiaofei Cui, Chengsheng Zhou, Hanxue Li, and Jianbo Zheng

Trusted Personalized Federated Learning Based on Differential Privacy 320
Ruixin Liu, Zhenquan Qin, Xi Cheng, Rui Zhang, and Jianbo Zheng

Networks and Applications

An Online Big-Data Driven Design of Reading and Writing Test 335
Yuwei Sun, Yongcheng Wen, and Yazhen Zhu

Research on Feature Extraction and Recognition of Inverter Fault Data Based on Neural Networks 354
Jingpeng Hu and Zhiguo Xiong

Transportation Networks

Design and Implementation of Traffic Flow Prediction Model Based on Short and Long Time Memory Network 99
Sheng Liu, Xinyue Li, Ting Cao, and Shuxiao Chang

Research on Traffic Sign Image Recognition Algorithms Under Complex Weather Conditions 109
Sheng Liu, Liming Qi, and Ting Cao

Deep Neural Network Based on Sparse Auto-Encoder for Road Extraction 120
Sheng Liu, Shuxiao Chang, Ting Cao, and Xinyue Li

DECS: A Decentralized and Efficient Cross-Chain Scheme in IoT System 128
Ying Gao, Peihao Zhang, Qiaofeng Pan, and Xianfeng Qiu

Artificial Intelligence Model Based Security Protection Method for IoT Applications 143
Xiaolong Luo, Xiaoli Chen, Jie Wei, Liang Zhang, Luping Xu, and Bijun Zhao

Extraction of Frequently Active Areas of Ships Based on Advanced Grid Density Peak Clustering 158
Xuanrui Xiong, Han Shen, Lanke Zhu, and Jianbo Zheng

Cyber Physical System Modeling and Analysis in Typical Scenarios Based on the Theory of Autonomous Transportation System 165
Zi-Sheng Zhou, Ming Cai, Zhuo-Lin Deng, and Chen Xiong

Reliability and Scalability

PathBit: A Bit Index Based on Path for Large-Scale Knowledge Graph 181
Yonglin Leng, Peiyi Qu, Ying Guo, and Chaoliang Xi

Skeleton Prototype Contrastive Learning with Multi-level Graph Relation Modeling for Unsupervised Person Re-Identification 196
Haocong Rao and Chunyan Miao

E-Health Networks II

Investigating the EEG Embedding by Visualization 221
Yongcheng Wen, Jiawei Mo, Wenxin Hu, and Feng Liang

Contents – Part I

E-Health Networks I

Transfer Learning for Audio-Based Speech Emotion Recognition in Chinese: Leveraging Pretrained Models for Improved Performance 3
Lanke Zhu, Xinyue Ma, Rui Zhang, and Jianbo Zheng

Optimal Design of Hydraulic Fracturing Simulation Experiments for In-Situ Stress Measurement 15
Yang Li, Daji Zhang, and Yimin Liu

Sentiment Analysis Based on Social Media - Early Stress and Depression Detection 26
Zixuan Li, Yuxuan Hu, Chenwei Zhang, Chengming Li, and Xiping Hu

Automatic Depression Detection Using Attention-Based Deep Multiple Instance Learning 40
Zixuan Shangguan, Xiaxi Li, Yanjie Dong, and Xiaoyan Yuan

Analysis of Factors Related to Anxiety and Depression in Medical Students ... 52
Zheng Jinfang, Pan Jiachen, Zhang Peiyi, Xiao Yi, and Wang Wei

Structural Health Monitoring of Carbon Fiber Composite Lamination Using Electrical Resistance 61
Guiping Lu, Xiaofeng Zhang, Shan Lu, Binghua Su, Kejun Wang, and Jiaran Liang

Identification of Economic Factors for Mass Depression Based on Panel Study and Machine Learning 72
Iaroslava Pravolamskaya, Jian Chen, and Wei Wang

Review of Sleep Monitoring Research Based on Wireless Sensor 79
Yuzhu Hu, Jian Chen, Shen Zhao, Kexin Tan, Kuai Yu, and Wei Wang

Understanding Obsessive-Compulsive Disorder Through Human Skin Textures 85
Yazhen Zhu, Jian Chen, Yuwei Sun, and Wei Wang

Workshops Chairs

Ying Gao	South China University of Technology, China
Amit Kumar Singh	National Institute of Technology Patna, India
Laisen Nie	Northwestern Polytechnical University, China

Publicity and Social Media Chairs

Edith Ngai	The University of Hong Kong, China
Yu Wu	Chongqing University of Posts and Telecommunications, China

Publications Chair

Kai Fang	University of Electronic Science and Technology of China, China

Web Chairs

Jianbo Zheng	Shenzhen MSU-BIT University, China
Marcin Wozniak	Silesian University of Technology, Poland

Panels Chairs

Jinyu Tian	Macau University of Science and Technology, China
Chengchao Liang	Chongqing University of Posts and Telecommunications, China

Tutorials Chairs

Han Liu	Dalian University of Technology, China
Lu Sun	Dalian Maritime University, China

Conference Organization

Steering Committee

Bo Li	Hong Kong University of Science and Technology, Hong Kong

Organizing Committee

General Chairs

Victor C. M. Leung	University of British Columbia, Canada
Hezhang Li	Shenzhen MSU-BIT University, China

General Co-chairs

Xiping Hu	Shenzhen MSU-BIT University, China
Zhaolong Ning	Chongqing University of Posts and Telecommunications, China

TPC Chairs and Co-chairs

Wei Wang	Shenzhen MSU-BIT University, China
Xiaojie Wang	Chongqing University of Posts and Telecommunications, China
Chengming Li	Shenzhen MSU-BIT University, China
Liangtian Wan	Dalian University of Technology, China

Sponsorship and Exhibit Chair

Jia Duan	Jd.Com, Inc., China

Local Chair

Yanjie Dong	Shenzhen MSU-BIT University, China

Preface

We are delighted to introduce the proceedings of the 19th European Alliance for Innovation (EAI) International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness. This conference brought together researchers, developers and practitioners around the world who are addressing the challenges from both trending application requirements and communication technologies. The theme of QShine 2023 was heterogeneous networking-related subjects, particularly quality, experience, reliability, security and robustness of heterogeneous networking.

The technical program of QShine 2023 consisted of 78 full papers, including 2 invited papers in oral presentation sessions at the main conference tracks. The conference tracks were: E-Health Networks; Transportation Networks; Reliability and Scalability; Artificial Intelligence and Machine Learning; Networks and Applications; Robustness; Network Security and Privacy; Quality of Service (QoS) and Quality of Experience (QoE). Aside from the high-quality technical paper presentations, the technical program also featured four keynote speeches. The four keynote speakers were Song Guo from Hong Kong University of Science and Technology, Hong Kong SAR, China Panos Nasiopoulos from The University of British Columbia, Canada, F. Richard Yu, Carleton University, Canada, and Jiangchuan Liu from Simon Fraser University, British Columbia, Canada.

Coordination with the steering chair Prof. Bo Li was essential for the success of the conference. We sincerely appreciate his constant support and guidance. It was also a great pleasure to work with such an excellent organizing committee team for their hard work in organizing and supporting the conference. In particular, the Technical Program Committee, led by our TPC Co-Chairs, Xiaojie Wang, Chengming Li, and Liangtian Wan, have completed the peer-review process of technical papers and made a high-quality technical program. We are also grateful to the Conference Manager, Ivana Bujdakova, for her support and to all the authors who submitted their papers to the QShine 2023 conference.

We strongly believe that QShine provides a good forum for all researchers, developers and practitioners to discuss all science and technology aspects that are relevant to heterogeneous networking-related subjects. We also expect that future QShine conferences will be as successful and stimulating, as indicated by the contributions presented in this volume.

October 2023

Xiping Hu
Wei Wang

Editors
Victor C. M. Leung
Shenzhen MSU-BIT University
Shenzhen, China

Hezhang Li
Shenzhen MSU-BIT University
Shenzhen, China

Xiping Hu
Shenzhen MSU-BIT University
Shenzhen, China

Zhaolong Ning
Chongqing University of Posts
and Telecommunications
Chongqing, China

ISSN 1867-8211 ISSN 1867-822X (electronic)
Lecture Notes of the Institute for Computer Sciences, Social Informatics
and Telecommunications Engineering
ISBN 978-3-031-65125-0 ISBN 978-3-031-65126-7 (eBook)
https://doi.org/10.1007/978-3-031-65126-7

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Victor C. M. Leung · Hezhang Li · Xiping Hu · Zhaolong Ning
Editors

Quality, Reliability, Security and Robustness in Heterogeneous Systems

19th EAI International Conference, QShine 2023
Shenzhen, China, October 8–9, 2023
Proceedings, Part I

Springer

The LNICST series publishes ICST's conferences, symposia and workshops.

LNICST reports state-of-the-art results in areas related to the scope of the Institute. The type of material published includes

- Proceedings (published in time for the respective event)
- Other edited monographs (such as project reports or invited volumes)

LNICST topics span the following areas:

- General Computer Science
- E-Economy
- E-Medicine
- Knowledge Management
- Multimedia
- Operations, Management and Policy
- Social Informatics
- Systems

Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 573

Editorial Board Members

Ozgur Akan, *Middle East Technical University, Ankara, Türkiye*
Paolo Bellavista, *University of Bologna, Bologna, Italy*
Jiannong Cao, *Hong Kong Polytechnic University, Hong Kong, China*
Geoffrey Coulson, *Lancaster University, Lancaster, UK*
Falko Dressler, *University of Erlangen, Erlangen, Germany*
Domenico Ferrari, *Università Cattolica Piacenza, Piacenza, Italy*
Mario Gerla, *UCLA, Los Angeles, USA*
Hisashi Kobayashi, *Princeton University, Princeton, USA*
Sergio Palazzo, *University of Catania, Catania, Italy*
Sartaj Sahni, *University of Florida, Gainesville, USA*
Xuemin Shen, *University of Waterloo, Waterloo, Canada*
Mircea Stan, *University of Virginia, Charlottesville, USA*
Xiaohua Jia, *City University of Hong Kong, Kowloon, Hong Kong*
Albert Y. Zomaya, *University of Sydney, Sydney, Australia*

Victor C. M. Leung
Hezhang Li
Xiping Hu
Zhaolong Ning (Eds.)



573

Quality, Reliability, Security and Robustness in Heterogeneous Systems

LNICST

19th EAI International Conference, QShine 2023
Shenzhen, China, October 8–9, 2023
Proceedings, Part I

Part 1

EAI

Springer