Research on Federated Sharing Methods for Massive Data in Blockchain

Bing Wu and Haiyan Kang(✉)

School of Information Management, Beijing Information Science and Technology University, Beijing 100192, China
kanghaiyan@126.com

Abstract. Data storage with the help of blockchain can ensure the transparency, non-tampering and autonomy of data information holders. However, the on-chain storage of massive data will seriously affect the performance of blockchain, and some private sensitive data and so on are not suitable for public storage in blockchain. To address the above problems, a trusted federated learning method based on local differential privacy mechanism, Loosely Coupled Local Differential Privacy Blockchain Federated Learning (LL-BCFL) is proposed for blockchain that realizes secure and efficient processing of massive user data. Firstly, a client selection mechanism is proposed and designed with the help of blockchain, which mainly includes two operations, namely, verification update and reputation calculation, to ensure the correctness and effectiveness of global model aggregation as well as the honesty and motivation of clients participating in training. Secondly, federated learning is used to realize the joint training of massive data distributed stored in each terminal device, so as to alleviate the phenomenon of "data silos" caused by privacy and security issues. In addition, a local differential privacy mechanism is designed in this method to solve the inference attack problem in the training process of federated learning. Finally, experiments are conducted on the MNIST dataset for both balanced and unbalanced datasets to verify the effectiveness of the proposed method LL-BCFL.

Keywords: Federated learning · Local differential privacy · Privacy protection · Blockchain storage · Massive data utilization

1 Introduction

In today's digital era, data is the core support and basic elements for the breakthrough development of emerging technologies and field industries such as artificial intelligence, cloud computing, mobile Internet and other big data industries. China has become the world's first data producer by virtue of the large amount of data generated from social networks, the Internet of Things, hospitals, banking systems, social networks, and other fields [1, 2], and the growth of data volume still shows explosive growth. The existence of massive data has provided fundamental guarantees for the development of various

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
D.-J. Deng and J.-C. Chen (Eds.): SGIoT 2023, LNICST 557, pp. 12–27, 2024.
https://doi.org/10.1007/978-3-031-55976-1_2

advanced technologies, but the increasing emphasis on data security and privacy protection by countries, enterprises, and individuals has led to the inability of massive data to be effectively shared and fully utilized [3, 4]. For example, data leakage, data theft and other data security issues caused by incidents such as the illegal and excessive collection and use of users' personal information by Didi Company, as well as the introduction of legal documents related to data privacy protection, such as the "Data Security Law of the People's Republic of China" and the "Personal Information Protection Law of the People's Republic of China", have all reflected the phenomenon of massive data being difficult to centrally process.

With the security attributes of blockchain such as traceability and tamperability, the secure storage of distributed data is effectively realized [5]. However, the on-chain storage of massive data will lead to a series of problems such as higher storage cost, slower transaction efficiency, and higher probability of node failure in blockchain. In view of the above blockchain massive data storage and the "data silos" problem caused by the distributed storage of massive data, it is proposed to adopt the idea of combining "on-chain and off-chain" to realize the efficient utilization and safe processing of massive data by combining with the federated learning technology [6]. In recent years, the model construction method combining blockchain technology and federated learning technology has been widely studied. For example, Qi et al. [7] proposed a federated learning framework based on federated blockchain to solve the security vulnerabilities and data privacy breaches caused by single point of failure, and ultimately realize the traffic flow prediction under privacy protection.

Federated learning is an effective solution to the privacy protection problem in machine learning, but there is still a risk of privacy leakage during its model training process [8–10]. With the help of federated learning technology to solve the "data silos" problem caused by industry competition, privacy security, and aggregation cost, while combining with the corresponding privacy protection methods to ensure the security in the model training process. In recent years, there are 2 main research directions on privacy protection techniques in federated learning, which are perturbation mechanism and encryption mechanism. The perturbation mechanism is mainly implemented by centralized differential privacy (CDP) and local differential privacy (LDP), which directly adds noise to the original data and perturbs sensitive information such as data features so that even if the data is leaked or stolen by a malicious attack, the valid information of the original data cannot be accurately inferred. Zhang et al. [11] applied the local differential privacy technique to the clustering problem and proposed AGCluster, a privacy-preserving grid clustering method based on LDP, to improve the quality of clustering under the premise of guaranteeing data privacy and security. Encryption mechanism is an indirect data privacy protection method acting in the process of data exchange, which is realized by combining security techniques such as cryptography tools, such as homomorphic encryption and secret sharing techniques. Yu et al. [12] proposed an efficient and secure federated aggregation scheme based on homomorphic encryption, which effectively solves the problems of federated learning data security as well as increased communication overhead after encryption.

The above related researches have made important breakthroughs in data processing and data privacy protection, but at the same time, there are still the following three urgent

problems to be solved, which are (1) federated learning under massive data suffers from the problem of inefficient information retrieval from distributed data terminal and is not easy to be managed, (2) there may be dishonest and unresponsive malicious clients participating in the training of federated learning models, and (3) only the validity verification is considered on the balanced data sets, without considering the actual prevalence of unbalanced datasets, which lacks the universal validation of the proposed method. Through in-depth research on the above issues, the main contributions of this paper are as follows.

(1) Propose a trusted federated learning method based on local differential privacy mechanism for blockchain, i.e., Loosely Coupled Local Differential Privacy Blockchain Federated Learning (LL-BCFL), to achieve the effective utilization and efficient processing of distributed massive user data.
(2) Propose a client selection method for participating in model training to ensure the correctness and effectiveness of global model aggregation, as well as the honesty and enthusiasm of participating clients in training.
(3) Design a local differential privacy mechanism to act in the federated learning parameter passing process, and perturb the data by adding noise to solve the privacy leakage problem in federated learning model training.
(4) Considering both balanced and unbalanced datasets, a large number of experiments are performed on MNIST real datasets to evaluate the effectiveness of the proposed method.

2 Related Work

2.1 Blockchain

Blockchain, as a distributed ledger technology with security attributes such as transparency, non-tamperability, and non-repudiation [13], is widely used in various fields such as finance, healthcare, and public services [14, 15]. While blockchain technology shows great potential in various fields, it also faces the following 2 key issues [16], which are (1) security issue, and (2) scalability issue. The security of blockchain systems is ensured by cryptography and consensus algorithms, however, theoretical weaknesses in security mechanisms can lead to the possibility of malicious attacks on blockchain systems such as malware attacks, distributed denial of service attacks, and other malicious attacks. The three main reasons that contribute to the scalability issues [17] of blockchain systems include the following (1) low throughput, (2) excessive data load, and (3) inefficient query engines.

Massive data storage with the help of blockchain may cause problems such as transaction, query, and other functions become inefficient due to data overload. Therefore, it is considered to combine with federated learning to directly store massive data locally at each terminal, which solves the problem of data security and privacy protection while solving the problem of decentralized massive data usage.

2.2 Federated Learning

Federal learning (FL) [18] adheres to the core idea of "data does not move, the model moves, and the data is available but not visible", and trains machine learning models by

combining data from multiple parties under the premise of not sharing local data. Combined with deep learning, privacy protection technology and other domain technologies, it solves the problem of maximizing the utilization of massive data under decentralized storage. The definition of federated learning is as follows.

Definition 1 Federated learning [19]. Define N participants $\{F_1, ..., F_N\}$ to hold their respective datasets $\{D_1, ..., D_N\}$ and collaborate to train a global model M_{FED}. Compared with the centralized traditional machine learning model M_{FED}, the federated learning model M_{FED} has a certain degree of accuracy loss. Let V_{FED} be the accuracy of the federated learning model and V_{SUM} be the accuracy of the traditional machine learning model, the loss of accuracy is

$$|V_{FED} - V_{SUM}| < \delta(\delta \text{ is a non} - \text{negative real number}) \tag{1}$$

Ideally, the loss of accuracy of the federated learning model is small, i.e., the non-negative real number δ is small.

Federated learning better solves the problems of "data silos" and data privacy and security, but it still has the defects of reliability and security. Therefore, it is an important challenge for federated learning to solve the security problem of parameter transfer between the center server and the clients and the honest reliability problem of the clients participating in the model training by combining with relevant privacy protection techniques.

2.3 Local Differential Privacy Technique

In 2008, Dwork proposed the concept of differential privacy (DP), which mainly relies on a randomization algorithm. Differential privacy mechanism can be utilized to protect against differential attacks and inference attacks in order to prevent the attacker from successfully obtaining the specific information of a particular piece of data based on some small differences in information. In this paper, in order to ensure the availability of data, we utilize the relaxation differential privacy which is widely used in real world scenarios and is defined as follows.

Definition 2 (ε, δ)-Differential privacy [20].Given n users, for any randomized algorithm M, take as input any two neighboring datasets D and D' that differ by at most one record, such that any subset of the output of algorithm M be Y $(Y \in R)$ and satisfy.

$$Pr[M(D) = Y] \leq e^{\varepsilon} \times Pr\left[M\left(D'\right) = Y\right] + \delta \tag{2}$$

Then the algorithm M is said to satisfy (ε, δ)-differential privacy. Where parameter ε denotes the magnitude of the degree of privacy protection, the smaller the value of ε the higher the degree of privacy protection. δ is a relaxation parameter to ensure the effectiveness of relaxed differential privacy, which is usually taken to be a small positive number, e.g., 0.1 or less.

Differential privacy is mainly realized by adding random noise to the input parameters or output results to perturb the data. Common data perturbation mechanisms [21] are Gaussian, Laplace and random response mechanisms. In particular, the Gaussian mechanism achieves (ε, δ)-differential privacy by adding normally distributed Gaussian noise with mean 0 and variance $\sigma^2 I$ to the output $f(t)$, i.e., $M(t) = f(t) + M\left(\sigma^2 I\right)$. The introduced Gaussian noise satisfies a Gaussian distribution and is a random number in the range between $(0, \sigma^2 I)$ and I for the unit matrix. The Laplace mechanism achieves (ε, δ)-differential privacy by adding to the output result $f(u)$ a Laplace noise generated according to the Laplace distribution of the probability density function $p(x|\lambda|) = \frac{1}{2\lambda} e^{-|x|/\lambda}$, i.e., $M(t) = f(u) + Laplace(\ \Delta f/\varepsilon)$, $Laplace(\bullet)$ is Laplace noise.

3 LL-BCFL Method Design

3.1 Description of the Problem

Storing massive data in the blockchain will exacerbate the problems of scalability, low throughput, and high latency efficiency of the blockchain itself, and distributed storage of large amounts of data across local users will exacerbate the phenomenon of "data silos", which leads to the inability to effectively utilize massive data. Federated learning technology is widely used to solve the centralized data collection and processing problems of traditional machine learning, but there are still privacy and security issues such as dishonest clients uploading false update parameters or intermediate parameter leakage in the model training process.

Therefore, in order to realize the effective utilization and efficient processing of massive data in distributed terminal devices while ensuring privacy and security, there is an urgent need to solve the above problems. The related symbols and parameters involved in this paper are shown in Table 1.

Table 1. Related symbols and parameters

Notation	Meaning
M	Number of clients
N	Number of federal learning participants
T	Total number of federated learning exchange rounds
ε	Privacy budget in local differential privacy definition
δ	Relevant parameters in local differential privacy definition
w_i	Model parameters relevant for client evaluation in federated learning
u_i	Client-trained local model

3.2 Local Data Query Privacy Protection Mechanism

To address the above problems, this paper combines blockchain technology to propose a trusted federated learning method based on local differential privacy mechanism, i.e.,

Loosely Coupled Local Differential Privacy Blockchain Federated Learning (LL-BCFL) to realize efficient processing of massive user data. The method adopts the core idea of combining "on-chain and off-chain", and is constructed by storing data summary information on the chain and storing and processing massive data off the chain, and its architecture is shown in Fig. 1, which mainly contains the following three methods, namely, (1) loosely coupled BCFL method, (2) federated learning method based on local differential privacy, and (3) massive data storage and processing method. The proposed method in this paper provides an effective solution to the problem of using massive data in industries such as healthcare, finance, and security, where data privacy requirements are extremely high.

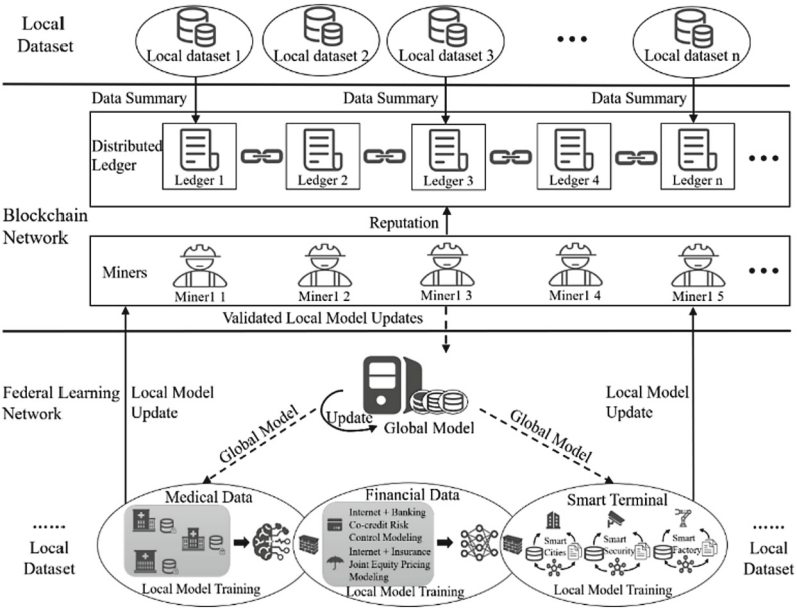

Fig. 1. System architecture diagram

3.2.1 Loosely Coupled BCFL Method

In this paper, the loosely coupled federated learning [22] approach is used to combine blockchain technology and federated learning mechanism, and its detailed coupling is reflected in Fig. 1. In particular, blockchain is used to validate model updates and manage the reputation of the client, which joins the federated learning network by sending down the reputation value. Federated learning based on server-client model is used for local model training and global model aggregation, joining the blockchain network by uploading local model parameter updates. In the loosely coupled BCFL, the distributed ledger is used to record information such as client data digests and reputation values, and the miners are used to provide a client selection mechanism for federated

learning to select appropriate clients to participate in the model training process. The client selection process is described as follows.

(1) Client self-assessment: the client evaluates itself based on its data type, data size, and data type.
(2) Server evaluation: The server weights the client's self-assessment value and historical reputation value, and calculates the client's reputation index with the help of subjective logic model (SLM) method, and finally generates the client's comprehensive reputation value.
(3) Client selection: Select the clients as the participants of the federal learning model training according to the comprehensive reputation value in descending order.

The specific implementation of the client selection mechanism is detailed in algorithm 1.

Algorithm 1: Select_Client
Input: number of clients M, number of participants in federated learning model training N, number of federated learning exchange rounds T.
Output: List of clients participating in federated learning model training $client\_select\_list$.

Step 1: Define the list $client\_eval\_list$, $client\_score$, $client\_select$.
Step 2: for $i \leftarrow 1$ to M do
Step 3: Iterate through M clients, weight and sum each client's assessment of its own data size, quality, and category to get the client self-assessment value $client\_eval$, and add it to the list of $client\_eval\_list$.
Step 4: $client\_eval \leftarrow s_1 * w_1 + s_2 * w_2 + s_3 * w_2$
Step 5: if $t \leftarrow 1$ then
Step 6: Select N clients to participate in the model training based on their evaluation values from largest to smallest and add them to the $client\_select$ list.
Step 7: else
Step 8: Obtain the reputation value $client\_rep\_list$ from the list $client\_rep$ of client's historical reputation value, weight and sum the self-assessment value and the reputation value to get the client's comprehensive assessment value, and add it to the $client\_rep\_list$ list.
Step 9: $client\_rep \leftarrow client\_eval * w_4 + client\_rep * w_5$
Step 10: Select N clients to participate in the model training based on their comprehensive evaluation values and add them to the $client\_select\_list$ list.
Step 11: end for
Step 12: return $client\_rep\_list$

The client selection mechanism acts during each round of federated learning model training, iterating in a loop until the model converges. The server, with the help of the blockchain, selects reliable participants with high reputation for the federated learning task based on the client selection mechanism. The high-reputation participants bring with them high-quality data for modeling training, which can significantly improve the learning efficiency of federated learning. At the same time, the validation mechanism is provided for federated learning with the help of miners to verify the validity of local model updates. The validation update process is described as follows.

Obtain local model updates: the miner obtains the local model updates uploaded to the blockchain network by the client.

Validate model parameters: the miner validates the received local model updates according to predefined rules and constraints.

The implementation of the local model update validation mechanism is detailed in algorithm 2.

Algorithm 2: Valid_Model
Input: Updated client local model u_i, test dataset $valid\_data$, number of correctly predicted labels cor_num.
Output: Client local model accuracy acc.

Step 1: Prediction on updated local model u_i using validation dataset.
Step 2: $prediction \leftarrow u_i.predict(valid\_data)$
Step 3: Iterate through the prediction results and real labels, record the number of correctly predicted labels using accuracy as a performance metric, and thus judge the effectiveness of local model updates.
Step 4: $if\ prediction == valid\_data$:
Step 5: $cor\_num\ +\!= 1$
Step 6: $acc = cor\_num\ /\ len(valid\_data)$
Step 7: return acc

After the miner completes the local model update validation mechanism, the validated model parameters are sent down to the server to ensure the validity and compliance of the local model update, to prevent malicious or invalid model parameters from entering the process of global aggregation, and to improve the reliability and overall performance of the models in the BCFL system.

3.2.2 A Federation Learning Method Based on Local Differential Privation

In this paper, we incorporate the local differential privacy mechanism into federated learning to solve the problem of possible inference attacks during intermediate parameter exchanges between server-side and client-side in federated learning. The method architecture is shown in Fig. 2, and its round of complete training process can be summarized as follows.

The central server is responsible for client selection and global parameter initialization, distribution, aggregation, and global model update. The selected clients get the global parameters and train the model locally by themselves. After that, the resultant parameters after adding noise perturbation are sent to the miners, who perform transaction validation to filter unqualified or even malicious local model updates [23], and calculate the reputation value to evaluate the reputation of the clients as the basis for the next round of client selection. The miner sends the verified local model updates and the reputation values of the clients to the server, which completes the final global model aggregation. This process iterates in a loop until the model converges.

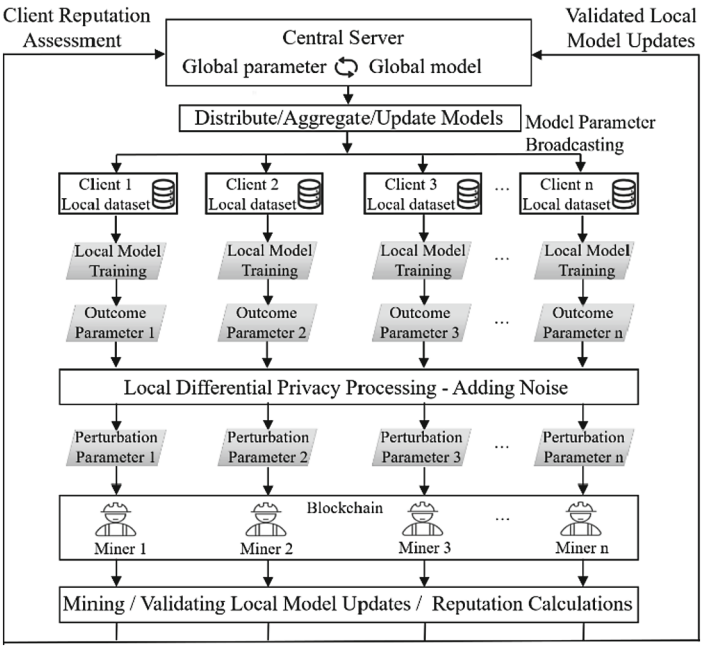

Fig. 2. A model structure for federated learning data sharing based on local differential privacy

3.2.3 Massive Data Storage and Processing Method

The storage of data in this system mainly includes the following two parts: (1) storage of simple data summary information on the blockchain, including local user information, dataset information, and client reputation value and other summary information, to ensure that the data stored on the chain is "visible and immutable", so as to facilitate the completion of the reputation calculation and client retrieval operations, and effectively improve data security and model training efficiency. (2) The storage of local datasets in the user terminal, which is directly managed by each data holder, provides a basic guarantee for data privacy and security. The processing of data in this system consists of the following three parts, namely, (1) the client uses the local data for model training and adds differential privacy noise to the process parameters, (2) uploads the trained local model updates to the miner for validation, and (3) sends the validated local model updates to the server for aggregation and sends the aggregated model parameters to the client.

3.3 Method Analysis

In the LL-BCFL method, the relevant data summary information is stored with the help of blockchain, and the unique properties of blockchain can ensure the security of the information. Secondly, the designed client selection mechanism can ensure the reliability and motivation of clients participating in model training. In addition, the introduction of local differential privacy mechanism can effectively avoid the problem of inference

attacks that may be suffered during the training process of federated learning models. Therefore, the LL-BCFL method proposed in this paper has high security.

For the time complexity of the algorithm, the LL-BCFL method mainly consists of the aggregation algorithm of the server and the local model training algorithm of the client, and the local model training algorithm of the client as well as the Select_Client algorithm and the Valid_Model algorithm proposed in this paper are nested in the aggregation algorithm of the server. Denote the overall number of iterations of the LL-BCFL method is $T.$, the number of participants is M, and the time complexity of the server's aggregation algorithm is $O(log(M))$ at each iteration, the time complexity of the LL-BCFL method is equal to that of the server's aggregation algorithm, which is $O(Tlog(M))$.

4 Experiment and Analysis

4.1 Experimental Environment and Dataset

Experimental Environment

This section evaluates the effectiveness of the modeling approach proposed in this paper and designs comparative experiments. The experiments are conducted under the operating system Windows 10 (64-bit), and the experimental code is implemented in the Pycharm development environment based on the programming language Python 3.8 for collaborative model training for federal learning. The hardware configuration is Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz, NVIDIA GeForce MX150 GPU, 8 GB RAM. Deep learning models are trained using Pytorch 1.7.1 and differential privacy noise is added. The neural network model used for the network architecture of this experiment is conventional neural network (CNN) through which the global model is trained iteratively. The stochastic optimization algorithm used in this experiment is stochastic gradient descent (SGD) method, with which the model parameters are adjusted to iteratively optimize the local model.

Experimental Dataset

The MNIST dataset was chosen for training and testing in the experiment, which contains 10 grayscale images of handwritten digits 28×28 with 60,000 training samples and 10,000 test samples. The MNIST dataset is loaded automatically by code during the experiment to provide training data for the local model training of each participant as a way to simulate the horizontal federated learning process. The multi-party data used for federated learning suffers from variability issues due to factors such as domain-industry characteristics, which can have a direct impact on the accuracy and validity of the final model. Therefore, it is necessary to balance the data samples in the dataset before model training begins.

4.2 Model Effectiveness Evaluation Experiment

In this section, the effectiveness of the proposed method is evaluated by comprehensively considering both balanced and unbalanced datasets, using convolutional neural network (CNN) as the client's training model. Two parameter variables, namely the noise type and the number of training rounds, are mainly targeted to explore their effects on the global accuracy of the model, respectively. The two experiments are conducted on the MNIST dataset, and the main implementation methods are as follows.

(1) Explore the effect of noise type on the global accuracy of the model.

Under the differential privacy parameter C = 10% and participant N = 100, the added noise is set to be Laplace noise, Gaussian noise, respectively, and no differential privacy without adding any noise is used as a control to derive the impact on model accuracy produced by the addition of privacy noise.

1) Under the above conditions, the model is trained for 100 rounds using the balanced dataset and the results are shown in Fig. 3.

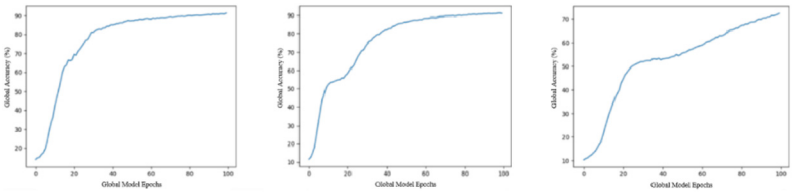

Fig. 3. Global accuracy with 100 rounds of IID distribution (left to right No-DP, Laplace-DP, Gaussian-DP)

2) Under the above conditions, the model is trained for 100 rounds using the unbalanced dataset and the results are shown in Fig. 4.

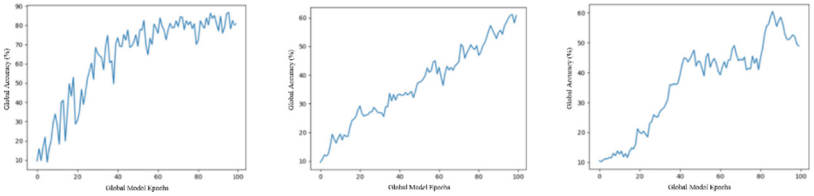

Fig. 4. Global accuracy with 100 rounds of Non-IID distribution (left to right No-DP, Laplace-DP, Gaussian-DP)

The following conclusions can be drawn from the curves obtained from the experiment, Figs. 3 and 4.

(1) Using a balanced dataset trained for 100 rounds with the same number of participants, adding noise affects the accuracy of model training. In particular, adding Laplace noise has little effect on the model accuracy and the final model accuracy reaches 90%. However, adding Gaussian noise has a greater effect on the model accuracy and the model accuracy reaches only 70% at the highest.
(2) Using the unbalanced dataset to train 100 rounds under the premise of the same number of participants, the curve fluctuates greatly compared to Fig. 3, and the final accuracy is reduced, and the addition of Laplace noise and Gaussian noise all appear

to be unable to converge. Therefore, under the condition of fewer training rounds using unbalanced dataset, adding noise can not complete the training of the model.

(3) Under the same training conditions, the model accuracy fluctuates greatly during model training using the unbalanced dataset, and its final accuracy decreases compared to model training using the balanced dataset. At the same time, the Gaussian mechanism differential privacy method for low latitude dataset under less rounds of learning has too much influence on the perturbation of the data, which seriously affects the model effectiveness and cannot be used.

(2) Explore the effect of the number of training rounds on the global accuracy of the model.

In the case of differential privacy parameter C = 10% and participant N = 100, the number of training rounds is set to 100, 200, and 500, respectively. Three cases of no differential privacy, Laplace mechanism differential privacy, and Gaussian mechanism differential privacy are considered to derive the impact of the number of training rounds on the accuracy of the model.

1) Under the above conditions, 100 rounds of training the model with balanced dataset and the results are shown in Fig. 3; 100 rounds of training the model with unbalanced dataset and the results are shown in Fig. 4.
2) Train the model for 200 rounds under the above conditions.

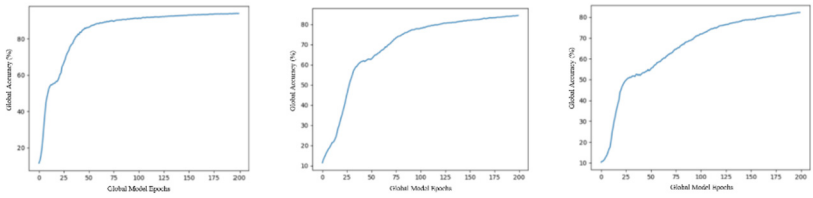

Fig. 5. Global accuracy with 200 rounds of IID distribution (left to right No-DP, Laplace-DP, Gaussian-DP)

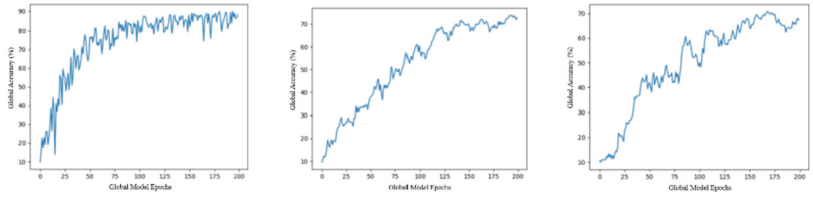

Fig. 6. Global accuracy with 200 rounds of Non-IID distribution (left to right No-DP, Laplace-DP, Gaussian-DP)

Through the curves obtained from the experiment, Figs. 5 and 6, and combined with Figs. 3 and 4, the following conclusions can be obtained.

(1) Using a balanced dataset trained for 200 rounds with the same number of participants, the global model accuracy curves with the addition of two noises respectively have

roughly the same trend, and the final model accuracies are all between 80% and 90%.

(2) Using the unbalanced dataset with the same number of participants for 200 rounds of training, the final accuracies of the models are all reduced compared to Fig. 5. The federated learning curve graph with Laplace noise added in Fig. 6 has a 10% increase in final accuracy compared to the federated learning curve graph with Laplace noise added in Fig. 4. The federated learning plot with Gaussian noise added in Fig. 6 shows a significant improvement in training results compared to the federated learning plot with Gaussian noise added in Fig. 4.

3) Train the model for 500 rounds under the above conditions.

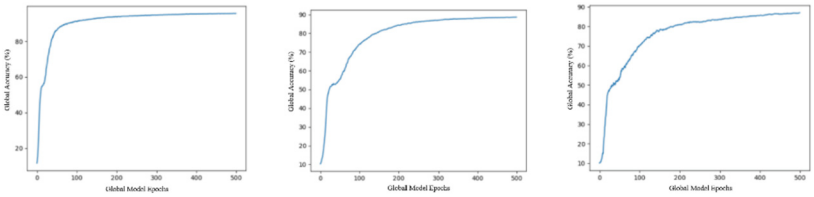

Fig. 7. Global accuracy with 500 rounds of IID distribution (left to right No-DP, Laplace-DP, Gaussian-DP)

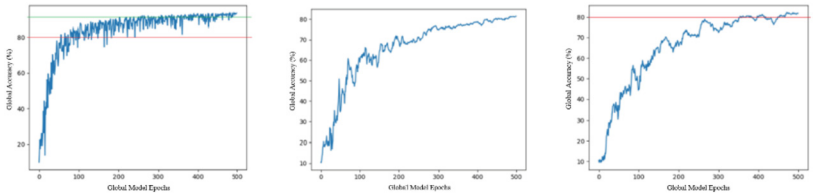

Fig. 8. Global accuracy with 500 rounds of Non-IID distribution (left to right No-DP, Laplace-DP, Gaussian-DP)

Through the curves obtained from the experiment, Figs. 7 and 8, and combined with Figs. 3, 4, 5 and 6, the following conclusions can be obtained.

(1) After 500 rounds of training with the same number of participants using the balanced dataset, the differences in the curves of the global model accuracy after adding the two noises separately become very small, and all of them have the characteristics of the curves obtained from the model training using the no-difference privacy technique at the early stage of training.

(2) Using the unbalanced dataset for 500 rounds of training with the same number of participants increases the final accuracies of the models compared to Fig. 6 for all of them, and the final model accuracies are almost the same compared to the results of Fig. 5 for 200 rounds of training using the balanced dataset.

(3) In the comparison of different number of training rounds using balanced dataset, the bounding curves of the federated learning accuracy under Laplace mechanism and

Gaussian mechanism both illustrate that increasing the number of training rounds of federated learning can improve the accuracy of the trained model. In the comparison of different numbers of training rounds using unbalanced datasets, unbalanced datasets that are prevalent in reality can be well aggregated into high-accuracy models as long as the number of rounds of training using federated learning is high enough.

5 Conclusions

In this paper, we combine blockchain technology and federated learning mechanism to propose and design a trusted federated learning method LL-BCFL based on local differential privacy mechanism to achieve secure and efficient processing of massive data. The blockchain is utilized to store simple summary information such as the personal information and reputation value of each local user, the type and number of entries of the data, to avoid storing too much data that affects the performance of the blockchain while improving the retrieval efficiency of user-related information. Design a selection mechanism for clients participating in model training, whereby clients are evaluated by miners in the blockchain through the verification of model update parameters and reputation calculation operations to ensure that honest and active clients are selected to participate in each round of model training, improving the efficiency and accuracy of model training. Federated learning is used to achieve centralized "sharing" of distributed massive data, and a local differential privacy mechanism is introduced to effectively ensure the security of parameter transmission during model training. Finally, the proposed LL-BCFL method is validated on the real dataset MNIST by conducting experiments on balanced and unbalanced datasets, and comparing with the original federated learning without differential privacy. Future work will focus on the integration of federated learning with advanced technological fields, as well as the study of privacy-preserving techniques in federated learning, in order to realize the improvement of the global accuracy of the resulting learning model while ensuring data privacy and security.

Acknowledgment. This work is partially supported by the National Social Science Foundation. China (No. 21BTQ079), the Humanities and Social Sciences Research Foundation of the Ministry of Education, China (No. 20YJAZH046), Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing Fund, and Scientific Research Project of Beijing Educational Committee (KM202011232022).

References

1. Hu, J., Vasilakos, A.V.: Energy big data analytics and security: challenges and opportunities. IEEE Trans. Smart Grid **7**(5), 2423–2436 (2016)
2. Kang, H., Ji, Y., Zhang, S.: Enhanced privacy preserving for social networks relational data based on personalized differential privacy. Chin. J. Electron. **31**(4), 741–751 (2022)
3. Meng, X., Zhu, M., Liu, J.: Quantitative research on privacy risk of large-scale mobile users. J. Inform. Secur. Res. **5**(9), 778-788 (2019). (孟小峰,朱敏杰,刘俊旭.大规模用户隐私风险量化研究.信息安全研究, 2019, **5**(09): 778–788.)

4. Xiaofeng, M., Minjie, Z., Lixin, L., Junxu, L., et al.: Research on data monopoly and its governance modes. J. Inform. Secur. Res. **5**(9), 789–797 (2019). (孟小峰,朱敏杰,刘立新,刘俊旭.数据垄断与其治理模式研究.信息安全研究, 2019, **5**(09): 789-797.)
5. Wang, L.P., Guan, Z., Li, Q.S., et al.: Survey on blockchain-based security services. J. Softw. **34**(01), 1–32 (2023). (王利朋,关志,李青山等.区块链数据安全服务综述.软件学报, 2023, 34(01): 1–32.)
6. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. **10**(2), 12 (2019)
7. Qi, Y., Shamim Hossain, M., Nie, J., Li, X.: Privacy-preserving blockchain-based federated learning for traffic flow prediction. Future Gener. Comput. Syst. **117**, 328–337 (2021)
8. Yin, X., Zhu, Y., Hu, J.: A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions. ACM Comput. Surv. **54**(6), 1–36 (2021)
9. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. **10**(2), 1–19 (2019)
10. Zhu, L., Han, S.: Deep leakage from gradients. In: Yang, Q., Fan, L., Yu, H. (eds.) Federated Learn. LNCS (LNAI), vol. 12500, pp. 17–31. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63076-8_2
11. Zhang, D.Y., Ni, W.W., Zhang, S., Fu, N., Hou, L.H.: A local differential privacy based privacy-preserving grid clustering method. Chinese J. Comput. **46**(02), 422-435 (2023). (张东月,倪巍伟,张森等.一种基于本地化差分隐私的网格聚类方法.计算机学报, 2023, **46**(02): 422–435.)
12. Yu, S.X., Chen, Z.: Efficient secure federated learning aggregation framework based on homomorphic encryption. J. Commun. **44**(01), 14–28 (2023). (余晟兴,陈钟.基于同态加密的高效安全联邦学习聚合框架.通信学报, 2023, 44(01): 14–28.)
13. Zhou, Y., Wang, C., Xu, J., Hu, K., Wang, J.: Privacy-preserving and decentralized federated learning model based on the blockchain. J. Comput. Res. Dev. **59**(11), 2423–2436 (2022). (周炜,王超,徐剑,胡克勇,王金龙.基于区块链的隐私保护去中心化联邦学习模型.计算机研究与发展, 2022, 59(11): 2423-2436.)
14. Sergii, K., Silvio, R., Giada, S.: Blockchain Tree for eHealth. In: The Internet of Things, pp. 1–5 (2019)
15. Kushch, S., Castrillo, F.P.: Blockchain for dynamic nodes in a smart city. In: The Internet of Things, pp. 29–34 (2019)
16. Muhammad, N.M.B., et al.: A survey on blockchain technology: evolution. Architect. Secur. IEEE Access **9**, 61048–61073 (2021)
17. Wei, Q., Li, B., Chang, W., Jia, Z., Shen, Z., Shao, Z.: A survey of blockchain data management systems. ACM Trans. Embed. Comput. Syst. **21**(3), 1–28 (2022). https://doi.org/10.1145/3502741
18. Kairouz, P., McMahan, H.B., Avent, B., et al.: Advances and open problems in federated learning. Found. Trends Mach. Learn. **14**(1–2), 1–210 (2021)
19. Dwork, C., Lei, J.: Differential privacy and robust statistics. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, pp. 371–380. Association for Computing Machinery, Bethesda (2009)
20. Wang, X.-S., Kang, H.-Y.: Research on noise addition and precision analysis in differential privacy. J. Lanzhou Univ. Technol. **49**(3), 94–103 (2023). (王骁识,康海燕.差分隐私中噪声添加与精度分析研究.兰州理工大学学报, 2023, **49**(03), 94–103.)
21. Tang, L.T., Chen, Z.N., Zhang, L.F., Wu, D.: Research progress of privacy issues in federated learning. Chinese J. Comput. **34**(01), 197–229. (汤凌韬,陈左宁,张鲁飞等.联邦学习中的隐私问题研究进展.软件学报,2023,34(01):197-229.)

22. Gu, T.L., Li, L., Chang, L., Li, J.J.: Fair federated machine learning and its design: a comprehensive survey. Chinese J. Comput. **46**(09), 1991–2024 (2023). (古天龙,李龙,常亮,李晶晶.公平联邦学习及其设计研究综述[J/OL].计算机学报, 2023, 46(09): 1991-2024.)
23. Kim, Y.J., Hong, C.S.: Blockchain-based node-aware dynamic weighting methods for improving federated learning performance. In: Asia-Pacific Network Operations and Management Symposium, pp. 1–4 (2019)