A Botnet Detection Method Based on SCBRNN

Yafeng Xu, Kailiang Zhang(✉), Qi Zhou, and Ping Cui

Jiangsu Province Key Laboratory of Intelligent Industry Control Technology, Xuzhou University of Technology, Xuzhou 221018, China
zhangkailiang@xzit.edu.cn

Abstract. With the rapid development of the social network and Internet of things, the complex network environment has led to more serious network security issues. Botnets have always been one of the most important issues in network security. The continuous update of botnet technology has severely influence the network operation of Internet service providers, posing a huge threat to security. Effective detection of botnets is the focus of related security solutions. In the new environment, traditional solutions have become inefficient. In recent years, botnet detection results based on machine learning technology continue to emerge. From the perspective of small batch gradient sample collection, this article optimizes the two-way neural network model and adopts approximate entropy to determine the abnormality of the data, thereby effectively detecting botnets. Research data shows that the model has good performance and can accurately identify botnets. Compared with the traditional model method, when the small batch sampling range is reduced, the accuracy is significantly improved, which provides effective help for Internet service providers to accurately detect botnets, improves service security mechanisms, and improves core competition force.

Keywords: Botnet · SCBRNN · Small batch · ApEn

1 Introduction

With the development of emerging technologies such as artificial intelligence, big data and 5G, some emerging network technologies and communication methods have been proposed [1, 2], and network quality and experience quality have been continuously improved [3, 4]. The energy efficiency methods improve the performance of internet of things [5–8]. However, the large number of devices accessing to the network pose challenges in security. The security threats facing the network are increasing, and security problems and attacks related to global networks often occur [9]. In the global risk factor rankings of the World Economic Forum, cyber attacks ranked the top five, becoming the third largest risk factor, causing great damage to production and life [10]. As a common method of network attacks, botnets have caused huge damage to the normal operation of the network and the security of information [11]. Network operators provide the most important basic services including network data transmission services and DNS resolution services. However, botnets can take advantage of their characteristics to

D. Jiang and H. Song (Eds.): SIMUtools 2021, LNICST 424, pp. 123–131, 2022.
https://doi.org/10.1007/978-3-030-97124-3_11

easily launch targeted Various attacks and damages to basic network operations, including DDoS attacks and DNS attacks, which have caused huge damage to basic network operations and services, severely undermined network information security [12, 13]. Therefore, effective response to botnets is the core security task of each operator, and accurate detection is the key. In recent years, artificial intelligence and machine learning have greatly changed the network security industry. The traffic analysis methods [14, 15] and complex web data analysis on cloud platform can help recognition [16–21]. Because behavior characteristic statistical analysis and behavior simulation monitoring are difficult to extract and detect target data, this paper proposes an efficient optimization method SCBRNN (bidirectional recurrent neural network for small batch data acquisition), which optimizes based on the characteristics of network traffic data Improve the two-way recurrent neural network to improve learning efficiency, and finally judge whether the target network host is invaded by a botnet based on the characteristics of entropy.

2 Related Work

Among the existing botnet detection methods, emerging methods based on machine learning are constantly being proposed. Most of these methods are based on the characteristics of traffic for qualitative analysis. The literature [22] conducted a study on the network behaviors with potential hidden dangers, fully analyzed the characteristics of these network behaviors (ActBehavior, FailBehavior, ScanBehavior), found the difference signs from the network level, and then used the mean algorithm to compare the hosts in the botnet Members confirm. [23] From the perspective of communication characteristics and traffic classification, some scholars proposed a cluster correlation verification method. Through the clustering and judgment of malicious traffic, the host in the network was identified by the botnet. Literature [24] proposed a detection method based on P2P traffic feature extraction, which extracts the traffic of P2P-related applications, such as artificial intelligence, application software, entertainment terminals, and basic services, and then uses the Bayes method for detection. This method has Higher detection accuracy. Literature [25] proposed a detection algorithm based on the concept of random walk statistics. The feature of the algorithm is to target unstructured P2P botnets and use graph data in real-world scenarios for detection. The evaluation results show that better detection accuracy can be obtained. However, for the network robot targets below 5%, the detection effect needs to be studied. GetoarGallopeni et al. [26] proposed a traffic analysis method based on DNS attacks and Mirai. Its main feature is to capture characteristic traffic for analysis while performing simulated attacks. Although this method has better proactive defensive detection, it has better performance in experiments. Indoor hardware testing environment requirements are high, and social applicability needs to be further confirmed. S. Chen et al. [27] used feedforward artificial neural networks to extract effective traffic convolution from the target network, which can effectively identify botnets. Although this method has a certain accuracy, the confidence level needs to be further improved. In the literature [28], scholars analyzed the transmission path of botnets through TCP/HTTP protocol, confirmed the connection

characteristics based on TCP, and proposed a technology based on traffic behavior to detect botnets, because robots in the target network The nature of the difference is not suitable for feature-based detection. There are also scholars [29] using statistical learning methods, using a lightweight logistic regression model to identify the characteristics of the botnet traffic, and then confirm it through the Bro network monitoring framework, and classify malicious and benign traffic. In the literature [27, 30], a test method and system based on convolutional network are proposed, which is characterized by collecting, merging, and measuring data flow in a flow counter, which has good accuracy. In actual production and life, network operators, in order to be able to respond well to the threats and destruction brought by botnets, have deployed targeted solutions [31]. Existing solutions based on intrusion detection mainly use active detection methods to trigger targeted defense actions. Based on intrusion detection technology, it relies on various technologies to detect botnets, including request recognition [33], statistical recognition and entropy detection. The characteristics of these methods are all based on collecting botnet traffic characteristics to define them the behavior of. Due to their different development environments, the existing test technology has a challenging task, which is the reliability of operation in actual scenarios. Therefore, designing an intrusion detection method model for Internet service providers will provide adequate protection for end users and the underlying network, and can adapt to frequent changes. More accurate and effective botnet detection methods and mechanisms are of great significance for Internet service providers to reduce the harm of botnets to basic networks, improve service efficiency, protect the interests of end users, and enhance corporate competitiveness.

3 Model

Based on the characteristics of mini-batch gradient descent, this paper proposes an efficient optimization method SCBRNN, which uses a part of the sample to update the parameters in each iteration. Its advantage is that optimizing the neural network parameters of a part of the sample each time is not much slower than that of a single data, and using a part of the sample for each training can greatly reduce the number of iterations required for convergence, reduce the probability of gradient explosion, and perform effective convergence at the same time. To make the results more real and reliable, the relevant algorithms are shown in Table 1.

In Table 1, m represents the sample size, P is the sample collection amount, l and r are the sample collection boundaries, and g() and f() are the activation functions, thus showing the process of the bidirectional recurrent neural network based on small batch sample collection. The structure is shown in Fig. 1.

Table 1. SCBRNN Algorithm.

Heading level
1. Input layer X
2. Initialize the input layer data
3. While time node t has not ended
4. Computed hiding layer A and A′ with f()
5. If t < m * P
6. Take samples 1 to r(t + m * P)
7. Else if t > m − m * P
8. Take samples l(t − m * P) to m
9. Else
10. Take samples l(t − m * P) to r(t + m * P)
11. Calculate the target value Y with g()
12. Output layer Y
13. End

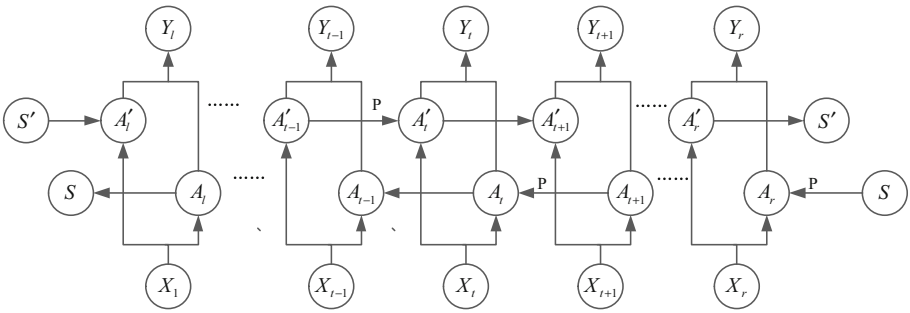

Fig. 1. Model structure of SCBRNN.

In Fig. 1, not every node has a connection between the hidden layers. For the node information at time t, part of the data before and after it will be collected instead of all of it. This is essentially different from the W in the model itself. Combining the recursive formula of the unidirectional cyclic neural network and the bidirectional cyclic neural network, adding the acquisition rate P, the information of the hidden layer is obtained as follows:

$$A_t = f(WPA_{t+1} + Ux_t) \tag{1}$$

$$A'_t = f\left(W'PA'_{t-1} + U'x_t\right) \tag{2}$$

Among them (1) is the At calculated by the forward input layer, and (2) is the At′ calculated by the forward input layer. P and WA are the vector product, and the total data

amount before and after the t node is collected is the data amount of P. In this way, the recurrence formula can be used to perform calculations based on part of the data, thereby increasing the fault tolerance rate and reducing the probability of gradient problems. The output formula (3) obtained by combining formulas (1) and (2) is:

$$Y_t = g(Vf(WPA_{t+1} + Ux_t) + V'f(W'PA'_{t-1} + U'x_t)) \tag{3}$$

Incorporating formulas (1) and (2) into (3) for continuous iteration, the sample collection formula (4) for the previous sequence of the t node is:

$$LA' = V'f(W'P(\ldots f(W'PA_{t-1+1\ldots}Ux_{t-1})) + U'x_t) \tag{4}$$

The sample collection formula (5) for the subsequent sequence of the t node is:

$$RA = Vf(WP(\ldots f(WPA_{t+r-1\ldots}Ux_{t+1})) + Ux_t) \tag{5}$$

4 Model Training

Select the network traffic data set for training. The data set spans three months and involves 10 local workstation IPs. According to the improved model, we will combine the data set and use RNN, BNRR and SCBRNN according to the changes in the data collection rate to determine whether the 10 local workstation IPs have been invaded and become members of the botnet. Existing research has collected data characteristics in the tuple of network information flow for identification, which is composed of interface index, source IP address, destination IP address, source port number, destination port number, and protocol number. The device performs network information flow statistics on past data packets according to the tuple information. However, in order to make better use of the time dynamic information of network traffic, this article does not care about the internal load information of network traffic, nor does it involve the privacy of network traffic. In this way, on the basis of improving security, the data characteristics are collected on the remote ASN (integer that identifies the remote ISP) and traffic (the number of connections in a day). The data set collected this time is based on network data traffic using ten different IP addresses in the past three months. The compiled data information statistics are shown in Table 2.

Table 2. Network data traffic statistics under different IP.

IP	Count	IP	Count
0	3980	5	1249
1	2159	6	1305
2	2416	7	2233
3	1186	8	2230
4	1308	9	2737

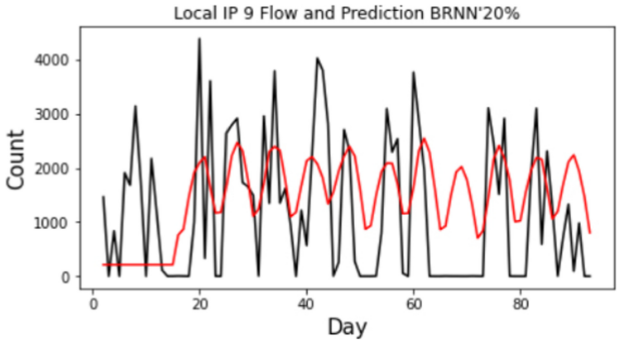

Fig. 2. Network traffic value of IP 9 predicted by SCBRNN.

It can be seen from Table 2 that in order to make more objective predictions and training, a total of 20,802 pieces of data have been prepared, and each piece of data contains Date information, that is, the time of data collection, which ensures the objectivity of the data. It also contains the IP number. For better data analysis, the IP address has been converted into a number. The ISP in the data bar is used to remotely identify the number of remote ASNs, and Stream is used to mark the IP at a specific time. How many network traffic connections have been made, these two data information we will use as the data characteristics of the information. Use Python's pandas library to perform statistics on the data and fill in the default values in the data information. The results of SCBRNN are shown in Fig. 2.

In Fig. 2, the sampling rate of SCBRNN is 20%. It can be seen that this method inherits the advantages of the two-way cyclic neural network, fully expresses the fluctuation trend of the data, and combines the characteristics of some data before and after each time. But not all. The data can better represent the data near the peak and more accurately predict the exact value of the network data flow.

5 Result Analysis

In order to test the accuracy of the model method, ApEn (Approximate Entropy) is introduced to evaluate the performance of the model. Approximate entropy is a nonlinear dynamic parameter used to quantify the regularity and unpredictability of time series fluctuations. The more complex the time series, the greater the approximate entropy. Approximate entropy has strong anti-interference ability. If the data contains abnormal values, ApEn can be compared with the abnormal level to determine the degree of expression of the true information in the original data. Taking IP6 as an example, the approximate entropy calculation process of setting the N-dimensional time series according to the statistical data is u(1), u(2) . . . u(N).

Define algorithm related parameters m, r, where m is an integer, which represents the length of the comparison vector, m in this article is the interval of days, which is 2, and r is a real number, which represents the measure of "similarity", in general, r = 0.2* std, where std is the standard deviation of the sample, r = 150 in this article.

Then, the approximate entropy can be expressed as shown in the following formula (6).

$$\mathrm{ApEn} = \Phi^{m}(r) - \Phi^{m+1}(r) \tag{6}$$

Combined with the keras and numpy libraries in Python, the approximate entropy time changes of IP6 in cyclic neural algorithms are shown in Fig. 3.

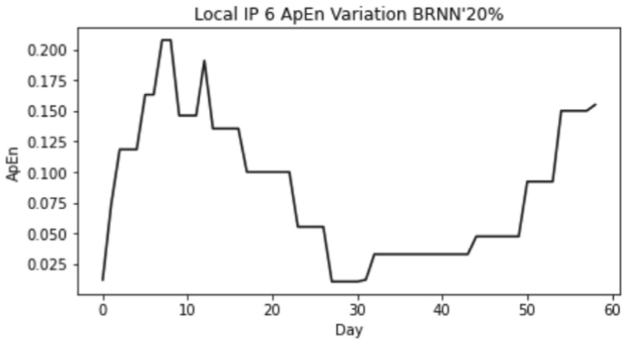

Fig. 3. Time trend of ApEn with IP 6 based on the predicted values of three algorithms.

It can be seen from Fig. 3 that in the approximate entropy calculation process, the vector that meets the condition will be extracted when the vector is reconstructed. Therefore, only 60 vectors are extracted in the ApEn timing diagram. On the whole, it can be found that the ApEn of the LocalIP6 host is relatively large, exceeding the level of 0.1, indicating that the statistical data has a large abnormal fluctuation during the model training process, indicating that it has been hacked. And become a member of the botnet.

6 Conclusion

This paper proposes a botnet detection method based on deep learning. From the perspective of small batch gradient sample collection, the two-way neural network model is optimized, and the approximate entropy method is used to judge data anomalies, thereby effectively detecting botnets. Research data shows that the model has good performance and can identify botnets more accurately. Compared with traditional model methods, this model method can learn the characteristics of botnet traffic more comprehensively and efficiently. When the small batch sampling range is reduced, the accuracy of botnet detection is significantly improved. The ability to detect unknown botnets can effectively solve the problem of accurate identification of botnets by Internet service providers, improve the quality of service security, and enhance the core competitiveness of enterprises.

Acknowledgment. This work is partly supported by Jiangsu technology project of Housing and Urban-Rural Development (No. 2019ZD041).

References

1. Zhang, K., Chen, L., An, Y., et al.: A QoE test system for vehicular voice cloud services. Mob. Netw. Appl. **26**, 700–715 (2019)
2. Chen, L., Jiang, D., Bao, R., Xiong, J., Liu, F., Bei, L.: MIMO scheduling effectiveness analysis for bursty data service from view of QoE. Chin. J. Electron. **26**(5), 1079–1085 (2017)
3. Chen, L., et al.: A lightweight end-side user experience data collection system for quality evaluation of multimedia communications. IEEE Access **6**(1), 15408–15419 (2018)
4. Chen, L., Zhang, L.: Spectral efficiency analysis for massive MIMO system under QoS constraint: an effective capacity perspective. Mob. Netw. Appl. **26**, 691–699 (2020)
5. Jiang, D., Wang, Z., Wang, W., et al.: AI-assisted energy-efficient and intelligent routing for reconfigurable wireless networks. IEEE Trans. Netw. Sci. Eng. **9**, 78–88 (2020)
6. Jiang, D., Huo, L., Zhang, P., et al.: Energy-efficient heterogeneous networking for electric vehicles networks in smart future cities. IEEE Trans. Intell. Transp. Syst. **22**, 1868–1880 (2020)
7. Jiang, D., Wang, Y., Lv, Z., Wang, W., Wang, H.: An energy-efficient networking approach in cloud services for IIoT networks. IEEE J. Sel. Areas Commun. **38**(5), 928–941 (2020)
8. Jiang, D., Huo, L., Lv, Z., Song, H., Qin, W.: A joint multi-criteria utility-based network selection approach for vehicle-to-infrastructure networking. IEEE Trans. Intell. Transp. Syst. **19**(10), 3305–3319 (2018)
9. Mohammadian, M.: Network security risk assessment using intelligent agents. In: 2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR), Putrajaya, pp. 1–6 (2018)
10. Huang, K., Yang, L., Fu, R., Zhou, S., Hong, Z.: HASN: a hierarchical attack surface network for system security analysis. China Commun. **16**(5), 137–157 (2019)
11. Vormayr, G., Zseby, T., Fabini, J.: Botnet communication patterns. IEEE Commun. Surv. Tutor. **19**(4), 2768–2796 (2017)
12. Shafi, Q., Basit, A.: DDoS botnet prevention using blockchain in software defined internet of things. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, pp. 624–628 (2019)
13. Li, W., Jin, J., Lee, J.: Analysis of Botnet domain names for IoT cybersecurity. IEEE Access **7**, 94658–94665 (2019)
14. Jiang, D., Wang, Z., Huo, L., et al.: A performance measurement and analysis method for software-defined networking of IoV. IEEE Trans. Intell. Transp. Syst. **22**, 3707–3719 (2020)
15. Jiang, D., Wang, W., Shi, L., Song, H.: A compressive sensing-based approach to end-to-end network traffic reconstruction. IEEE Trans. Netw. Sci. Eng. **7**(1), 507–519 (2020)
16. Yang, B., Bao, W., Huang, D.-S.: Inference of large-scale time-delayed gene regulatory network with parallel MapReduce cloud platform. Sci. Rep. **8**(1), 1–11 (2018). https://doi.org/10.1038/s41598-018-36180-y
17. Yang, B., Bao, W.: Complex-valued ordinary differential equation modeling for time series identification. IEEE Access **7**(1), 41033–41042 (2019)
18. Jiang, D., Huo, L., Song, H.: Rethinking behaviors and activities of base stations in mobile cellular networks based on big data analysis. IEEE Trans. Netw. Sci. Eng. **7**(1), 80–90 (2020)
19. Jiang, D., Wang, Y., Lv, Z., Qi, S., Singh, S.: Big data analysis based network behavior insight of cellular networks for industry 4.0 applications. IEEE Trans. Ind. Inform. **16**(2), 1310–1320 (2020)
20. Yang, B., Wang, G., Bao, W.: CSE: complex-valued system with evolutionary algorithm. IEEE Access **7**(1), 90268–90276 (2019)
21. Ghafir, I., et al.: BotDet: a system for real time botnet command and control traffic detection. IEEE Access **6**, 38947–38958 (2018)

22. Qiu, Z., Miller, D.J., Kesidis, G.: Flow based botnet detection through semi-supervised active learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, pp. 2387–2391 (2017)
23. Mai, L., Park, M.: A comparison of clustering algorithms for botnet detection based on network flow. In: 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN), Vienna, pp. 667–669 (2016)
24. Dhayal, H., Kumar, J.: Botnet and P2P botnet detection strategies: a review. In: 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, pp. 1077–1082 (2018)
25. Muhs, D., Haas, S., Strufe, T., Fischer, M.: On the robustness of random walk algorithms for the detection of unstructured P2P botnets. In: 2018 11th International Conference on IT Security Incident Management & IT Forensics (IMF), Hamburg, pp. 3–14 (2018)
26. Gallopeni, G., Rodrigues, B., Franco, M., Stiller, B.: A practical analysis on Mirai Botnet traffic. In: 2020 IFIP Networking Conference (Networking), Paris, France, pp. 667–668 (2020)
27. Chen, S., Chen, Y., Tzeng, W.: Effective botnet detection through neural networks on convolutional features. In: IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 372–378 (2018)
28. Kapre, A., Padmavathi, B.: Behaviour based botnet detection with traffic analysis and flow interavals using PSO and SVM. In: International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 718–722 (2017)
29. Bapat, R., et al.: Identifying malicious botnet traffic using logistic regression. In: Systems and Information Engineering Design Symposium (SIEDS), pp. 266–271 (2018)
30. Kant, V., Singh, E.M., Ojha, N.: An efficient flow based botnet classification using convolution neural network. In: International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 941–946 (2017)
31. Garg, S., Sharma, R.M.: Anatomy of botnet on application layer: mechanism and mitigation. In: International Conference for Convergence in Technology (I2CT), pp. 1024–1029 (2017)