EURASIAN JOURNAL OF ACADEMIC RESEARCH
Innovative Academy Research Support Center
UIF = 8.1 | SJIF = 5.685 www.in-academy.uz

ANDROID MALWARE CLASSIFICATION APPROACH BASED ON HOST-LEVEL ENCRYPTED TRAFFIC SHAPING

[1]Latipova Nodira Xalimovna
Teacher of the Department, "Systematic and Practical Programming", Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, UZBEKISTAN
[2]Ibragimov Jaloliddin Obidjon o'g'li
Teacher of the Department, "Systematic and Practical Programming", Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, UZBEKISTAN
https://doi.org/10.5281/zenodo.7478822

ARTICLE INFO

Received: 13th December 2022
Accepted: 21th December 2022
Online: 23th December 2022

KEY WORDS

Android malware classification; Host-level traffic; Encrypted traffic analysis; Machine learning; Confusion classifier.

ABSTRACT

With the development of mobile terminals, smartphones have attracted a very huge number of users with their powerful functions. Among them, Android system is famous for its opensource and convenience, which occupies a large market share. But this also leads many attackers to use their malware to gain benefits quickly, which make it necessary to design a practical android malware detection approach. At present, there are not many pieces of research on detecting malware by analyzing Android malicious traffic. This paper examines the characteristics of malicious traffic on the host computer to construct a traffic fingerprint. It combines machine learning algorithms to build a practical detection approach which is also suitable for encrypted traffic. To distinguish similar fuzzy traffic, an additional layer named confusion classifier is added to help further malware classification. This paper uses a realworld dataset called CICAndMal2017 and simulates two classification scenarios: malware binary detection and malware category classification. The experimental results show that the accuracy of the malware binary detection reached 98.8% while the accuracy rate of malware category classification is 95.2%.

1. Introduction

With the development of the digital age, smartphones have become an indispensable part of people's daily lives. According to the statistics of Statista, the number of global smartphone users is expected to reach 3.8 billion in 2021 [1]. And in the report of International Data Corporation [2], Android became the most widely used smartphone system in 2019, which occupied 86.7% market share. However, Android has also become the primary target of attackers. According to the report by 360 Mobile Security Research Office [3], there are 1.048 million new mobile malware samples in the first half of 2020, with almost an average of 6 thousand new samples per day. As user's privacy and

experience are greatly threatened by Android malware, we urgently need some effective methods to detect and defend against them. Abnormal network traffic detection [4] is a useful malware detection method that has attracted the attention of more and more researchers in recent years. According to the survey, there are four methods of detecting malware [5]: port-based methods, DPI-based methods, statistics-based methods, and behaviorbased methods. In this paper, we combine methods based on statistics and behavior, use three machine learning methods to classify malware. We have made the following contributions in this paper:

1). We analyze host-level network traffic to detect malware and further identify specific types of them.

2). Our approach weakens the timing features in traffic to improve the accuracy of classification by 5%-10%.

3). We use confusion classifier to deal with data similarities in multiple classifications, which eventually increased the accuracy to 95.2%.

This paper has been organized into five sections. Section 2 describes the related work. Section 3 demonstrates the detection method in detail. Section 4 conducts comparison experiments and analyzes the results. Section 5 provides a summary of the paper.

2. Related work

With the rapid development of malware, a variety of malware variants are emerging, and attackers have made some confusion measures against malware detection methods. There are some researches find that malware traffic fingerprint is an effective detection method. Jiang *et al.* [6] noted that 93% of the Android malware required a network connection to connect with the attacker and receive commands from the C&C server. Besides, Yerima *et al.* [7] analyzed the permissions of 2,000 applications, and more than 93% of malicious applications required network connectivity. Therefore, more and more researchers are now engaged in the analysis of abnormal network traffic.

In 2011, Iland *et al.* [8] proposed a method of detecting malware and user privacy disclosure by analyzing network traffic. The authors parse HTTP traffic to build a static feature library to detect information leaks and malicious commands. Fakhroddin *et al.* [9] analyzed the network traffic of ransomware in detail, tested the effectiveness of various machine learning models on binary classification and ransomware family classifications. Mohammad *et al.* [10] extracted session-level network traffic characteristics, applied it with a supervised model and conducted experiments in three scenarios. The effect with the same dataset has been significantly improved.

With the spread of HTTPS, more and more researchers have put research centres on the analysis of encrypted traffic. Arora *et al.* [11] proposed an Android malware detection method that does not require the resolution of payloads. They identified seven statistical characteristics for an experiment whose accuracy reached 95%. In addition, Li *et al.* [12] proposed a network traffic monitoring system, which consists of four parts: traffic monitoring, traffic anomaly identification, response processing, and cloud storage. The system parses the packet protocol, extracts the characteristic data (process ID, start time and end time, etc.), and then seed the feature vector to the SVM classifier to find

abnormal classes. The experimental results show that the monitoring system is useful in detecting Android malware.

However, there are not many pieces of researches about the malware category classification. Because of this situation, this paper analyzes the differences in data traffic generated by different types of malware, combined with the current machine learning classification algorithm, to find out the specific types of the malware. And to improve the classification accuracy, the confusion classifier is used to assist, the particular approach will be described in section 3.

3. Methodology

This section describes the proposed malware classification approach. The approach mainly analyzes the statistical characteristics and behavioral characteristics of network traffic, combines three machine learning methods for malware classification, and adds a confusion classifier to improve the accuracy of classification. Fig.1 shows the complete structure of the approach, and a detailed description of each part is given below.

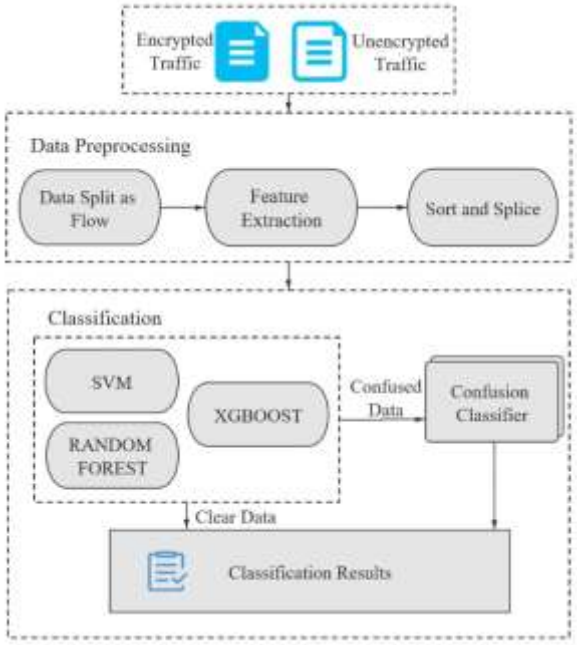

Fig.1 The overall architecture of the approach.

3.1. Traffic preprocessing

Firstly, we divide these complex files into multiple discrete units according to a certain granularity. There are currently five packet segmentation methods [13]: TCP connection, flow, session, service, and host. This paper combines flow and host methods to process PCAP files.

A flow is a collection of data packets with the same fivetuple information including source IP, destination IP, source port, destination port and protocol. However, these malicious behaviors may hide in the usual traffic, so it is unconsidered to directly mark all flows with the same label of the PCAP source file. We can only guarantee that the label of the PCAP source file is reliable, so we merge the feature

vector of the flow as the fingerprint of the PCAP file.

3.2. Feature extraction

This paper extracts the initial feature of the flow based on statistics and behavioral methods, including the size of the packet, the number of packets, and the interval between sending packets. Fortunately, the CICFlowMeter [14] tool can help to process the data. The tool divides the PCAP file into units of flows and extracts 76 features for each flow.

All the flows are divided into *N* groups while the aggregation functions of the feature vectors are calculated in each group. In this paper, maximum, minimum, average and standard deviation is chosen. Then, the feature vector of the *N* flow sets are spliced together to obtain a long feature vectors of the entire PCAP file whose dimension is *N**4*76.

Since each malware has different traffic, the time of malicious traffic has a particular impact on the classification. Considering this situation, this paper focuses on whether a certain shaped flow exists, and sorts the value of the same features, which can weaken the influence of the timing features and keep the characteristic information of the flow.

3.3. Machine learning model

We select three machine learning algorithms: SVM, Random Forest, and XGBoost. In the multi-classification experiment, we found that Adware and Scareware are always difficult to distinguish, so we used confusion classifier to help. Firstly, the instances which were classified as Adware and Scareware by the first classifier will be marked as confused. Then confused instances were sent to confusion classifier whose training data is only Adware and Scareware.

4. Experiment and evaluation

All the experiments have been carried out on the Microsoft Windows 10 Professional (64-bit) version with a processor which is i5-7400, 3.00GHz and 16GB of memory. Python 3.7.3 has been chosen for data preprocessing, feature extraction and construction of detection models.

This paper uses the public dataset of the Canadian Institute of Cyber Security which called CICAndMal2017 [15]. CIC was captured in a real mobile phone environment and generated from 2126 Android applications (426 of which were malware and the remaining 1,700 were benign). The PCAP files were divided into five types: Adware, Benign, Ransomware, Scareware, and SMS Malware. Therefore, two classification scenarios were simulated for experiments:

1). Scenario A: divide Android applications into benign and malicious

2). Scenario B: classify Android application software into specific types including Adware, Benign, Ransomware, Scareware and SMS Malware.

Our experiments used 80% of the dataset as the training set and 20% as the test set. And three classifiers: Random Forest, SVM and XGBoost were used for detecting. The experiment adopted the accuracy to evaluate our approach.

First of all, we verified the effectiveness of sorting the same type of features to weaken the effect of segmented position. *N* values were 7, 10, 15, and experiments were conducted in both scenarios. The result is the accuracy of the three machine learning models. As shown in Fig.2, it is not difficult to find that it has a good improvement effect on our detection model to weaken timing features by sorting, and

EURASIAN JOURNAL OF ACADEMIC RESEARCH
Innovative Academy Research Support Center
UIF = 8.1 | SJIF = 5.685 www.in-academy.uz

the accuracy rate has increased by 5%-10%.

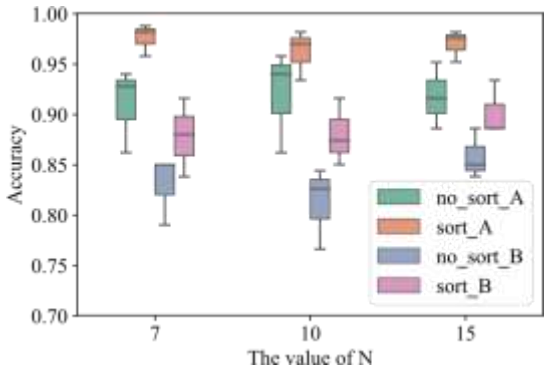

Fig.2 Experiment of weakening timing features.

Then, we change the values of N to find its effect. As shown in Fig.3, the best value is 9 in Scenario A with 98.8% score, and when it turns into Scenario B, the best value is 15.

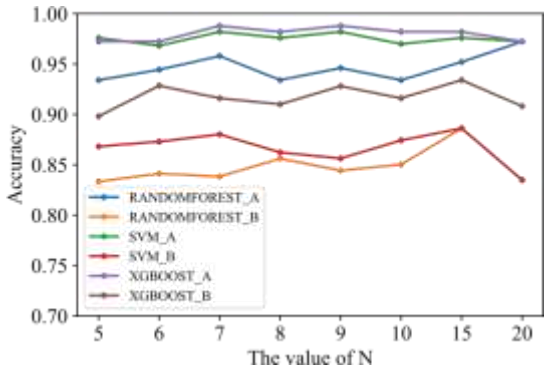

Fig.3 Experiment of different values of N.

The experiment for confusion classifier only works on Scenario B. We choose XGBOOST as the confusion classifier algorithm. The result is shown in Figure 4. Obviously, for three algorithms, the accuracy of confusion classifier is improved. Eventually, our classifier has reached 95.2% accuracy in the work of judging malware categories.

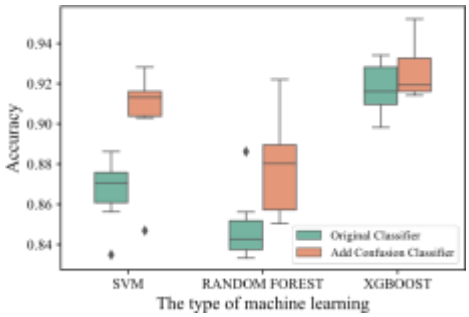

Fig.4 Experiment of confusion classifiers.

Finally, we compared the effect of our approach with CIC [14] and Mohammad [10]. Our model is superior to these two models in both scenario A and scenario B.

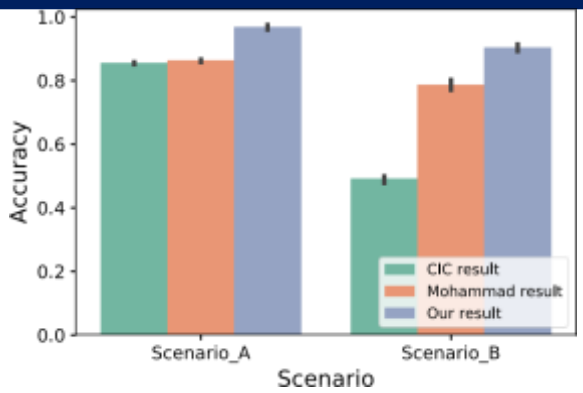

Fig.5 Performance comparison with other studies.

5. Conclusions

This paper proposes an approach to classify malware by analyzing host-level traffic. We use CICAndMal2017 dataset, extract flow characteristics from the flow level, merge the feature vectors to form the fingerprint. Then we use three machine learning algorithms for classification, and design confusion classifier to improve accuracy. Our approach is suitable for encrypted traffic and can effectively detect malware whose accuracy rates reached 98.8% and 95.2% in two scenarios. As part of future work, we plan to combine with other malware detection approaches based on usage permissions and code auditing to obtain a stronger effect.

Acknowledgements

This work was partially supported by the National Key Research and Development Program of China (Grant No. 2016QY13Z2302), the National Natural Science Foundation of China (Grant no. 61902262), the National Defense Innovation Special Zone Program of Science and Technology (Grant no. JG2019055), the Special Central Finance project (Grant Y030202059018060).

References:

1. “(Online) Smartphone Users Report 2020”. Available: https://www.statista.com/statistics/330695/number-ofsmartphone-users-worldwide/
2. “(Online) IDC Smartphone Report 2019”. Available: https://www.idc.com/getdoc.jsp?containerId=prUS454 87719
3. “(Online) 360 Smartphone Security 2020”. Available: https://zt.360.cn/1101061855.php?dtid=1101061451& did=610637885
4. Ahmed, Mohiuddin, Abdun Naser Mahmood, and Jiankun Hu, "A survey of network anomaly detection techniques." Journal of Network and Computer
5. Applications, Vol 60, pp. 19-31, Jan. 2016.
6. Biersack, Ernst , C. Callegari , and M. Matijasevic . Data Traffic Monitoring and Analysis. Springer Berlin Heidelberg, 2013.
7. Y. Zhou and X. Jiang, "Dissecting Android Malware: Characterization and Evolution," 2012 IEEE Symposium on Security and Privacy, San Francisco, CA, pp. 95-109, May 2012.
8. Yerima, Suleiman Y., Sakir Sezer, and Gavin McWilliams. "Analysis of Bayesian classificationbased approaches for Android malware detection, " IET Information Security, Vol 8, pp. 25-36, Jan. 2014.

9. Iland, Danny, Alexander Pucher, and Timm Schauble. "Detecting android malware on network level, " University of California, Santa Barbara, Vol 12, Dec. 2011.
10. F. Noorbehbahani, F. Rasouli and M. Saberi, "Analysis of Machine Learning Techniques for Ransomware Detection," 2019 16th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC), Mashhad, Iran, 2019, pp. 128-133, Aug. 2019.
11. Abuthawabeh M, Mahmoud K, "Enhanced Android Malware Detection and Family Classification, using
12. Conversation-level Network Traffic Features, " International Arab Journal of Information Technology, Vol 17, pp. 607-614, Jul. 2020.
13. A. Arora, S. Garg and S. K. Peddoju, "Malware Detection Using Network Traffic Analysis in Android Based Mobile Devices," 2014 Eighth International Conference on Next Generation Mobile Apps, Services and Technologies, Oxford, pp. 66-71, Sept. 2014.
14. J. Li, L. Zhai, X. Zhang and D. Quan, "Research of android malware detection based on network traffic monitoring," 2014 9th IEEE Conference on Industrial Electronics and Applications, Hangzhou, pp. 17391744, Jun. 2014.
15. A. Dainotti, A. Pescape and K. C. Claffy, "Issues and future directions in traffic classification," in IEEE Network, vol. 26, no. 1, pp. 35-40, Jan. 2012.
16. H. Lashkari, A. F. A. Kadir, L. Taheri and A. A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification," 2018 International Carnahan Conference on Security Technology (ICCST), Montreal, QC, pp. 1-7, Oct. 2018.