Data Mining Method of Malicious Attack Based on Characteristic Frequency

Jia Luo(✉) and Chan Zhang

School of Information Technology, Guangdong Industry Polytechnic, Guangzhou, China
luojia3l4@l63.com

Abstract. Aiming at the problem of high false alarm rate and failure rate in traditional data mining methods of malicious attacks, a data mining method of malicious attacks based on characteristic frequency is designed. Preprocess the original data in the data set, select the minimum attribute subset, use the discretization to process the unified data format, take the new subset as the input of feature frequency extraction of malicious attack data, extract the feature frequency according to the different protocols of malicious attack data transmission, integrate it into the value data mining algorithm, and use the spatial mapping principle to realize the malicious attack data excavate. The experimental results show that: compared with the traditional data mining method, the false alarm rate and failure rate of the designed malicious attack data mining method based on the feature frequency are reduced by 0.3 and 0.2 respectively, which shows that the method is more suitable for practical projects.

Keywords: Feature frequency · Malicious attack · Data mining · Spatial mapping principle

1 Introduction

Various attacks against Android are emerging in an endless stream. Malware detection for Android has become a very important link in the field of mobile security in recent years. Among them, web browsing accounts for a very large proportion of these activities. Because of this, many lawbreakers and hackers aim at the vulnerability of people's weak awareness of network security and deliberately carry out malicious activities attacking and intruding into the user's system, launching malicious attacks from different terminals used by the user, no matter web page or software, are extremely vulnerable to serious harm, which is the most serious problem of network security at present, greatly endangering the data security of the user using the Internet, and even causing serious economic losses [1].

Due to the continuous improvement of network speed and database technology, the number of access and operation log data generated in the Internet is increasing [2]. In the early stage, only relying on human power to extract the characteristic data of malicious attacks and normal behaviors from the log data has been unable to cope with the current large and complex network environment, so it is necessary to make

Y.-D. Zhang et al. (Eds.): ICMTEL 2020, LNICST 326, pp. 403–414, 2020.
https://doi.org/10.1007/978-3-030-51100-5_36

corresponding improvement according to the actual needs. Using data mining method is a simple and effective countermeasure.

Data mining refers to the process of extracting effective and potentially useful knowledge and patterns from a large number of uncertain and rough data, which is suitable for extracting behavioral features from a large number of historical data [3]. The technical characteristics of data mining meet the application requirements of building rule base. This can not only improve the accuracy of mining malicious attack data, but also effectively update the rule base intelligently, greatly ensuring the data security [4]. However, the traditional data mining methods can not extract the characteristics of malicious attack data completely, which makes it easy to mine out the normal security data when mining malicious attack data, resulting in data loss. When using the relevant data, there are different failures, and even the property loss of users. Therefore, the data mining method of malicious attack based on the characteristic frequency is designed, and the problems existing in the traditional data mining method are solved by using the characteristic of the characteristic frequency.

2 Design of Data Mining Method for Malicious Attacks Based on Characteristic Frequency

2.1 Data Preprocessing

The main task of data preprocessing stage is to randomly sample N records of normal data and malicious attack data specified in data set D, then filter all attribute $Q = \{Q_1, Q_2, \cdots Q_n\}$ of the selected records, and then filter out the required sub attribute set $q = \{Q_1, Q_2, \cdots Q_n\}$ according to the expression meaning and processing needs of each attribute for the next stage of processing.

Assuming that there are T normal data and I malicious attack data in database D, what we need to do in this stage is to sample N_1 normal data set D_1 from T normal data and N_2 attack connection data set D_2 from T malicious attack data. Take normal data extraction as an example, open the file containing data set D, read a record in the file as *line*, judge *line* as empty, then close the file, otherwise generate a random number between [0, T] *temp*, judge whether *temp* is less than or equal to N_2, if the result is "yes", then add the record to D_1, otherwise, continue to read the data in the file, cycle the above operations until all the numbers according to the completion cycle [5].

According to the above process, the algorithm is executed once for malicious attack data, and the required normal data subset D_1 and malicious attack data subset D_2 are selected. The process flow chart is shown in Fig. 1:

The purpose of the above filtering process is to filter out the relevant attributes with important meanings from the alternate attribute set. At the same time, ensure that the filtered data is the minimum set of attributes that can fully describe the network behavior [6].

After the completion of data filtering, discretization is carried out to deal with the data format. The continuous variables in the data are discretized. According to the defined functions, the functions are divided into specific linear intervals and given membership degree. In the very complex network connection data, the field types are

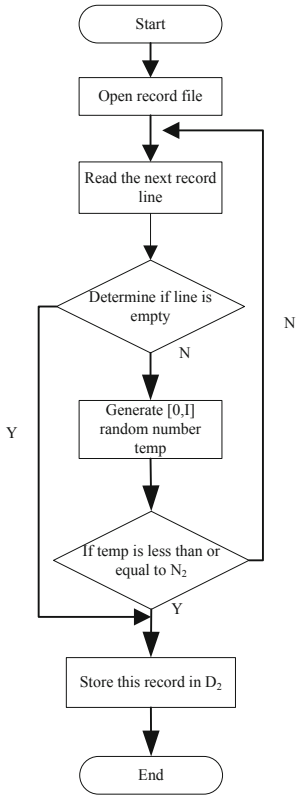

Fig. 1. Data filtering flow chart

binary, discrete and continuous [7]. In many previous studies, the treatment of continuous variables is to divide them into specific intervals according to different interval settings, which is the division of Cartesian sets [8]. The filtered data q in the data set is shown in Table 1:

Table 1. q data example table

Serial number	TID	q
1	001	500
2	002	400
3	003	600
4	004	700
5	005	800

The conversion function of variable q is defined as:

$$f(x) = \begin{cases} low & x \leq 400 \\ mid & 400 \leq x \leq 700 \\ high & x \geq 700 \end{cases} \quad (1)$$

According to the above functions, q of all records is transformed into linear interval q', and the division results are shown in Table 2.

Table 2. q data example discretization table

TID	q	q'
001	500	Mid
002	400	Low
003	600	Mid
004	700	High
005	800	High

In the above contents, mid, low and high represent that the q-transform of data set is divided into three linear intervals. The membership functions of dividing q to these three linear intervals are given as follows:

$$f_{low}(x) = \begin{cases} 1.0 & x \leq \alpha \\ \frac{x-\alpha}{\eta-\alpha} & \eta \leq x \leq \alpha \\ 0 & x \geq \alpha \end{cases} \quad (2)$$

$$f_{mid}(x) = \begin{cases} 0 & x \leq \alpha \\ \frac{x-\mu}{\alpha-\eta} & \eta \leq x \leq \alpha \\ 1.0 & x = \alpha \\ \frac{x-\omega}{\alpha-\omega} & \alpha < x < \omega \\ 0 & x \geq \omega \end{cases} \quad (3)$$

$$f_{high}(x) = \begin{cases} 0 & x \leq \alpha \\ \frac{x-\alpha}{\omega-\alpha} & \alpha \leq x \leq \omega \\ 1.0 & x \geq \omega \end{cases} \quad (4)$$

The membership functions defined by the above three formulas are shown in Fig. 2.

The database information processed by the above membership function is shown in Table 3.

After discretizing the malicious attack data set D_2 and normal training data set D_1 processed in the previous stage with the above membership functions, a new attack training data set D_{2c} and normal training data set D_{1c} are formed as the input of feature frequency extraction in the next stage.

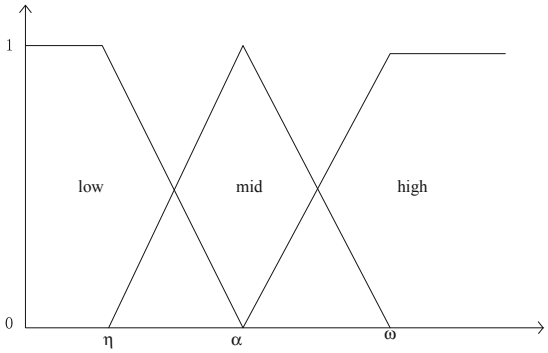

Fig. 2. Membership function

Table 3. Discrete processing data information

TID	q		
	q_{low}	q_{mid}	q_{high}
001	0.5	0.5	0
002	1.0	0	0
003	0	1.0	0
004	0	0.5	0.5
005	0	0	1.0

2.2 Extract Characteristic Frequency of Malicious Attack Data

Although there are many kinds of malicious attacks, they can be roughly divided into three forms (because Internet protocol is TCP/IP protocol, network layer is IP datagram, network layer is mainly two protocols, TCP protocol and UDP protocol, one is connection oriented protocol, the other is datagram service) [9].

The first is network packet data itself. Generally speaking, network data packets of different protocols are generally divided into two parts: Protocol head and data. Protocol head includes relevant management and control information for network packet transmission [10]. The information of data part load is generally divided into two kinds: one is protocol control and management information, the other is data information. The protocol control and management information of the protocol header and the data part load are strictly regulated by different protocols. Data information is produced in various network applications, which can be ordinary binary data sequence or string sequence [11]. At the IP network packet level, IP network packet encapsulation mainly includes Internet control message protocol, IGMP (Internet Group Management Protocol), TCP (transmission layer control protocol) and UDP (User Datagram Protocol). Most of the network malicious attacks are also realized through these four protocols. In the attack process, the network packet from the protocol header data and content may be different from the normal access characteristics [12].

The second form is the relationship between network packet and packet (packet sequence). It can be found that in some malicious attacks, although a single data packet conforms to RFC standard, the sequence of consecutive packets is different from the normal sequence. Because the data packets transmitted on the network are not arbitrary but according to the corresponding protocol, the sequence of packets must have its regularity, showing a specific frequency [13].

The third form is aimed at the statistical characteristics of network connection. In terms of the performance of data packets, some malicious attacks are similar to normal packets on the surface of a single packet, and the short-term packet sequence also conforms to the normal network transmission protocol. However, the statistics of various ways for a certain period of time or a certain length of network connection will form unique statistical characteristics, which can be distinguished from normal network access.

According to the above characteristics of malicious attacks on the transmission data, we use different data mining algorithms to analyze the huge amount of network data stored in the database, extract the appropriate characteristic frequency of malicious attacks data, and ensure that the extracted characteristic frequency does not take any subjective factors.

Malicious attack data is transmitted according to certain rules and protocols, and the characteristic frequency is shown in two aspects: one is reflected in the protocol header of each layer from the data link layer to the application layer, and the other is reflected in the fixed byte string TCP flow of the data part of the data part of the network data packet. Through the above analysis, the feature frequency is extracted.

First of all, historical data is classified: each kind of malicious attack data as a category, normal access data as the background class Z, because the main network protocol on the Internet is TCP/IP protocol, most of the malicious attack data is also carried out by TCP/IP protocol, so the main parsing method of TCP /IP protocol is different. For the common TCP protocol, because the TCP protocol is a connection oriented transport layer protocol, there is no connection between the whole data generated by the application program and the real single TCP packet, so the protocol data part takes all the data transmitted in the complete connection as the analysis object, and the data part of the single TCP packet as the analysis object. UDP protocol is a datagram protocol in the transport layer, which is not connection oriented. The data generated by the application program is completely reflected in each UDP packet, so the UDP protocol takes a single UDP packet as the protocol data analysis object.

For protocol header data, ICMP and IGMP protocols treat each value of each specified field as an item in the item set. All values of all specified fields constitute the item set $z = \{z_1, z_2, \cdots z_n\}$, and all header field values of each network packet constitute a transaction, that is, the item set. The historical database is transformed into a transaction database. The Apriori algorithm is used to find the frequent k-term set which satisfies the minimum support degree for the transaction database of protocol head type of network data transformation in the attack process. For these k-item sets in the transaction database which is the normal access data grouping transformation of the background class Z, the support degree of the background class Z is calculated as $S(x)$, and then the support degree of the background class Z is calculated as $Q(x)$. When $Q(x)$ is greater than or equal to the minimum support degree, it indicates that this frequent

k-item set has specific malicious attack data in the process of malicious attack, and the change of the support degree of malicious attack data is the characteristic frequency of malicious attack data, the algorithm of data mining based on feature frequency is used to mine malicious attack data.

2.3 Implementation of Malicious Attack Data Mining

Assuming that the malicious attack data obey the spherical distribution, the kernel function is used to map the data space to the high-dimensional kernel space, find a hypersphere that can contain all the data in the kernel space, find a minimum containing sphere that contains normal data, and solve the minimum containing sphere (sphere center, radius) by the minimum maximization method [14]. When identifying a new data, if the data is inside the sphere, it is considered normal, otherwise it is abnormal.

Firstly, the input space is mapped to a high-dimensional space by kernel function, in which a sphere containing all training data is constructed; the data points on the sphere are the obtained support vectors [15]. Assuming that model $r(x:\alpha)$ represents a kind of closely bounded data set, the goal of optimization is to find a minimum sphere with a center of x and a radius of α, and make all samples of the training set fall in the sphere. Define a minimization problem:

$$R(x,\alpha,\eta) = x^2 + E\sum_i \eta \tag{5}$$

Make this sphere satisfy:

$$\begin{cases} (x^2 - \alpha) \le \eta_i \\ \eta_i > 0, \forall i \end{cases} \tag{6}$$

In formulas 5 and 6, x represents the center of the sphere, α represents the radius of the sphere, η represents the relaxation variable, and E represents the influence of the adjustment relaxation variable. By using Lagrange function to solve the minimum optimization problem under the above constraints, it can be judged that the data belongs to this class and should meet the following requirements:

$$(z - x) \le \alpha^2 \tag{7}$$

In the formula, z represents a normal class, otherwise it is an abnormal class, that is, a class with malicious attack data. The criteria for determining whether the data belongs to this class are:

$$(z \cdot z) - 2\sum_i zx + \sum_i x_i \le \alpha^2 \tag{8}$$

In the formula, α represents the distance from any support vector to the spherical center x. When the data in the input space does not meet the spherical distribution, the input space can be mapped to the high-dimensional space first through the kernel

technique, and then solved in the mapped high-dimensional space. If the above conditions are rewritten as follows:

$$J(z \cdot z) - 2\sum_i Jzx + \sum_i Jx_i \leq \alpha^2 \tag{9}$$

In the formula, J represents a linear kernel function. In normal data, the number of malicious attacks is more, which means that the malicious attack data itself is active and the behavior is complex. Therefore, the feature frequency is integrated into the data mining algorithm to find the most suitable hypersphere for data mining, and the malicious attack data is included in the hypersphere as much as possible, so as to reduce the false alarm rate of data mining.

The characteristic frequency of malicious attack is integrated into the algorithm: the normal data activity frequency matrix $y = [y_1, y_2, \cdots, y_n]$ is defined, and n is the number of training data.

$$y_i = \frac{\sum_{j=1}^{m} b_{ij}}{\max\left(\sum_{j=1}^{m} b_{ij}\right)} \tag{10}$$

In the formula, i and j are variables, n represents the number of data, m represents the characteristic frequency, p represents a matrix of $m \times n$, then the problem of function minimization is as follows:

$$\min R(x, \alpha, \eta) = x^2 + E\sum_i y_i \eta_i \tag{11}$$

Adjust the corresponding constraints through the above formula:

$$\sum_i x_i = 1, 0 \leq x_i \leq y_i \tag{12}$$

By constraining the center of data sphere, the feature frequency is introduced into the algorithm. According to the principle of spatial mapping, the sphere of normal data and malicious attack data are constructed respectively, so as to realize the mining of malicious attack data.

3 Simulation Experiment of Data Mining Method of Malicious Attack

3.1 Data Extraction

Because of the large number of software users in the Android software market, the data is reliable. The data selection method in reference Android malware outlier detection

based on feature frequency is used to extract the data. Therefore, use data of Google store, app store, app Bao and other software were collected as normal sample test sets, a total of 1,823. 1619 malware was extracted from the above data sources as a negative sample data test set. In order to ensure the data quality, the missing data processing, repeated data processing and abnormal data processing are carried out. Then decompile these apps with apktool and other tools, run them to obtain manifest and small files, and complete the data extraction. Then we extract the feature frequency of attack data, and get 874 feature frequencies. Then we integrate the feature frequency of each sample into the mining algorithm as a new feature.

3.2 Evaluation Index

In order to accurately describe these two concepts and avoid unnecessary misunderstanding, 4 parameters and 2 evaluation criteria were introduced into the experiment. Four parameters are as follows:

FP: determine the data as normal data, which is actually the number of malicious data.

TN: determine the data as malicious data, which is actually the number of malicious data.

TP: determine the data as normal data, which is actually the number of normal data.

FN: determine the data as malicious data, which is actually the number of normal data.

2 evaluation criteria are as follows:

False positive rate: normal data is mined as malicious attack data. The calculation formula of false positive rate M is:

$$M = \frac{FN}{FN + TP} \times 100\% \tag{13}$$

Failure rate: malicious attack data is not successfully mined. The calculation formula of failure rate O is:

$$O = \frac{FP}{TN + FP} \times 100\% \tag{14}$$

3.3 Test Steps

Randomly select 1000 normal data as training set and other normal data and malicious attack data as test set. The frequency is the number of features contained in each data. Normalize it and then use it as data mining feature. Traditional data mining methods based on clustering analysis and data mining methods based on feature frequency are used respectively. For the convenience of description, a group using the traditional data mining method is called the control group, and a group using the feature frequency

based data mining method is called the experimental group. Two groups of data mining experiments at the same time, according to the experimental results for comparative analysis.

3.4 Experimental Results and Analysis

Through the above process, the data mining is completed, and the statistical software is used to count the false alarm rate and mining failure rate results. The statistical results are as follows.

As can be seen from Fig. 3, the overall trend of the experimental results of the false alarm rate of the data mining method based on the characteristic frequency is stable at about 0.1, with small fluctuation; the false alarm rate of the traditional data mining method is initially around 0.2, with the increase of the number of experiments, the false alarm rate gradually rises, and finally stabilizes at about 0.4; compared with the two, the data mining method based on the characteristic frequency of malicious attacks the false alarm rate of the method is lower than that of the traditional data mining method, which shows that in the data mining method, the feature frequency is used to effectively reduce the false alarm rate, which proves that the method is better.

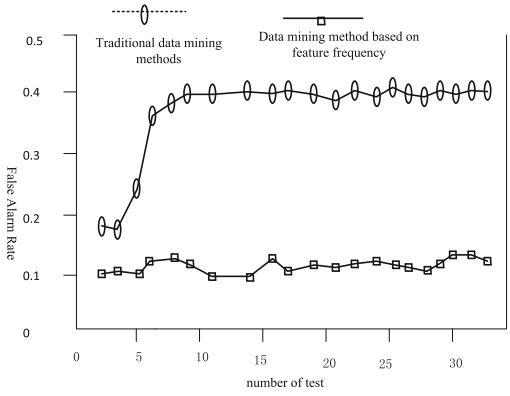

Fig. 3. False alarm rate results of different methods

It can be seen from Fig. 4, the failure rate of data mining method based on the characteristic frequency tends to 0 in many experiments, while in the traditional data mining method, the fluctuation range is large, the lowest is 0.1, the highest is 0.4, and the overall trend is high. Compared with the two methods, the failure rate of malicious attack data mining method based on feature frequency is lower, which shows that this method is better than the traditional data mining method.

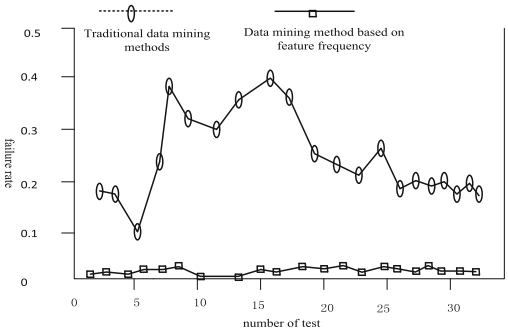

Fig. 4. Failure rate results of different methods

4 Conclusions

At present, with the rapid development of network, all kinds of network channels used by users are vulnerable to malicious attacks, resulting in data loss, even economic loss. Traditional malicious attack data mining methods have high false alarm rate and failure rate. In order to solve this kind of problem, a malicious attack data mining method based on the characteristic frequency is designed, which uses the stability and regularity of the characteristic frequency to reduce the false alarm rate and failure rate of the data mining method. Through many comparative experiments, it is proved that the data mining method of malicious attacks based on feature frequency effectively improves the problems existing in traditional data mining. However, because of the inherent characteristics of data extraction, the data experimental samples often mix with malicious behaviors, resulting in the deviation of experimental results. These problems need more in-depth research and discussion.

References

1. Liu, Y., Yuan, H.H.: Parallel discretization of data preparation optimization in data mining. J. Sichuan Univ. (Nat. Sci. Edn.) **55**(05), 103–109 (2018)
2. Zhang, Y., Yin, C.H.: Android malware outlier detection based on feature frequency. CAAI Trans. Intell. Syst. **13**(02), 168–173 (2018)
3. Wang, J., Zhang, Y.S., Chen, R.Y., et al.: Identification of user's role and discovery method of its malicious access behavior in web logs. Comput. Sci. **45**(10), 172–208 (2018)
4. Cui, Y., Song, W., Peng, Z.Y., et al.: Mining method of association rules based on differential privacy. Comput. Sci. **45**(06), 42–62 (2018)
5. Wang, J.Y., Liu, C., Fu, X.C., et al.: Crucial patterns mining with differential privacy over data streams. J. Softw. **30**(03), 158–176 (2019)
6. Zhang, B., Li, G.: Design of web anomaly data mining software based on improved clustering algorithm. Mod. Electron. Tech. **42**(08), 81–89 (2019)
7. Xiang, Z., Xiang, S.B.: Data mining based on fuzzy genetic algorithm. Control Eng. China **24**(05), 947–951 (2017)

8. Mei, Y., Xiong, T., Luo, S.B.: Research on NoSQL distributed big data mining method in complex attribute environment. Sci. Technol. Eng. **26**(09), 244–248 (2017)
9. Zhang, K.F., Liu, J.H., Zhang, J.F.: Local outlier mining algorithm for large scale high dimensional data set. Microelectron. Comput. **12**(03), 116–119 (2018)
10. Xu, L., Wang, J.X.: Data mining algorithm of abnormal network based on fuzzy neural network. Comput. Sci. **46**(04), 79–82 (2019)
11. Han, X., Du, S., Li, Z., et al.: Diagnosis of electric shock fault based on time-frequency singular value spectrum of leakage current and fuzzy clustering. Nongye Gongcheng Xuebao/Trans. Chin. Soc. Agric. Eng. **34**(04), 217–222 (2018)
12. Zhang, D., Yin, G., Jin, X., et al.: Two-stage and bi-direction feature selection method for EEG channel based on CSP and SFFS-SFBS. Dongnan Daxue Xuebao (Ziran Kexue Ban)/ J. Southeast Univ. (Nat. Sci. Edn.) **49**(01), 125–132 (2019)
13. Janus, T., Skomra, M., Marcin, D.: Individual security and network design with malicious nodes. Information (Switzerland) **9**(09), 214 (2018)
14. Cetinkaya, A., Ishii, H., Hayakawa, T.: A probabilistic characterization of random and malicious communication failures in multi-hop networked control. SIAM J. Control Optim. **56**(05), 3320–3350 (2017)
15. Preeti, P., Fawzia, G.F., Richard, S., et al.: Documenting attacks on health workers and facilities in armed conflicts. Bull. World Health Organ. **95**(01), 79–81 (2017)