Association rule mining of network security monitoring data based on time series

XUE Fang[1], LIU Rong[2]

{aaa21152@163.com[1], wz26677@163.com [2]}

[1]Information Technology Center, Jimei University, XiaMen 361021, China;

[2]School of Information Engineering, Changsha Medical University, Changsha 410219, China

Abstract: The traditional network security monitoring number association rule mining technology has low mining accuracy, so a time series based network security monitoring data association rule mining technology is designed. The preprocessing of time series to construct the corresponding time series frequency set, using SWFI - tree structure data storage model is set up, get after filtering and reorder the transaction data set, data sets of will be clean and remove invalid data and the remaining data formatting, finally USES the particle swarm optimization (pso) algorithm with limited data flow, recursive calculation of particle movement, build sparse list, complete monitoring data mining of association rules. The designed mining technology was used in the experiment with the traditional technology, and the experimental results showed that the designed mining technology was 23.22% more accurate than the traditional technology.

Key words: Time series; Network security monitoring; Association rule mining; Data cleaning;

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2295724

1 Introduction

With the rapid development of computer technology and the rapid popularization of the Internet, the role of computers in daily life and work is becoming more and more important. At the same time, the threats to network security are also increasing. Network viruses, Dos attacks, etc. are commonly used. Anti virus software, firewall, intrusion detection system and other network security defense technologies achieve passive security protection, which can not meet the requirements of network security. Therefore, real-time security monitoring of network security is needed [1-2], and association rules mining between security indicators Technology can implement a quantitative evaluation of the network status and monitor the network status in real time. It is a novel and superior network security technology. However, the traditional network security monitoring data association rule mining technology has low mining accuracy. Therefore, a time series-based network security monitoring data association rule mining technology is designed.

2 Association Rule Mining

Association rule mining is an important research direction of data mining. It excavates some interesting rules from a large amount of data. From these association rules, you can understand the association relationship between events and guide people Some actual meaningful behavior. This article is based on time series to mine the association rules of network security monitoring data. The relevant mining flowchart is shown below:

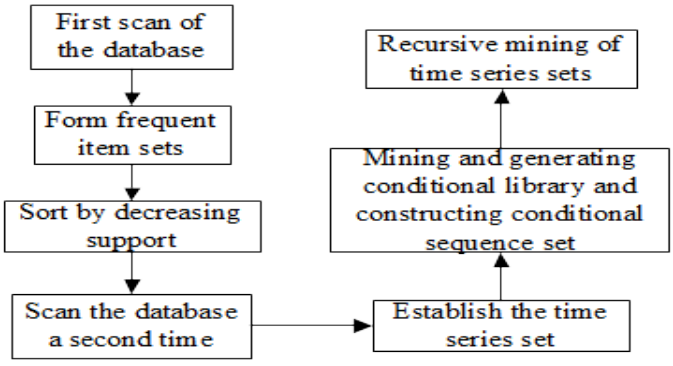

Fig. 1 Time series association rule mining flowchart

The time series algorithm belongs to depth-first search. Recursive search for short

patterns instead of long frequent pattern mining requires only two database scans. The first scan of the database produces frequent itemsets, and the second scan of the database establishes a global time series. The analysis from the mining process is divided into two steps [3-4]: First, compress the frequent sets in the database, retain the associated information between the data, and store them in the time series set; Second, because the time series contains all frequency sets, data mining only needs to work on the time series set, and separate it into the condition pattern database, respectively establish the condition time series set, and then carry out recursive mining.

2.1 Time series preprocessing

To apply time series to the mining of association rules between security indicators in network communication, we must first construct a corresponding time series set. However, since any monitoring data packet contains a large amount of network information, even if the time series set is fast and directly calculated, Processing is also difficult, so it is necessary to preprocess the monitoring data [5-6]. Information entropy is a better method of preprocessing. The detection data is packaged and the information is selected as a set of random events, as follows:

$$X=n_i(i=1,2,...,N) \quad (1)$$

The above formula can represent the number of times the geology in the information packet occurs at different events, and the information entropy $H(X)$ can be obtained as:

$$H(X)=-\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\lg_2\left(\frac{n_i}{S}\right) \quad (2)$$

In the above formula, S represents the total number of times an address appears, which can be expressed as:

$$S=\sum_{i=1}^{N}n_i \quad (3)$$

N is the total number of random events, n_i Represents a single event in a collection of events.It is a description of network uncertainty, which can intuitively and accurately reflect the distribution of system status. By preprocessing through information entropy, network security indicators can be expressed by the change in entropy of IP source and destination

addresses and port numbers. In network communication, the information entropy is smaller where the data is more concentrated, and vice versa. The preprocessed data information is stored in the database. When index mining is performed, the database is scanned once to obtain the set of frequency sets and the corresponding minimum support. The frequency set is arranged in descending order to obtain set L. The entire process is shown in the following figure:

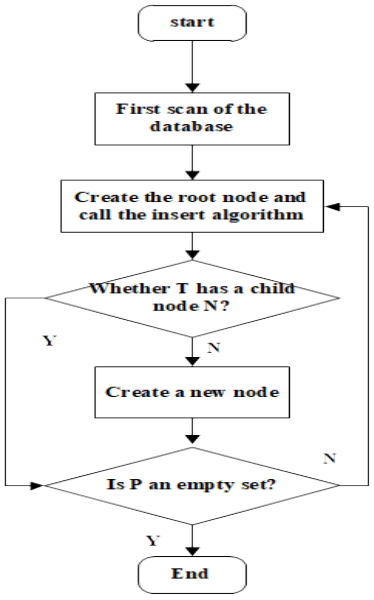

Fig. 2. Flow of frequency set generation

Where P is the information entropy of the packet information and P is the relevant network information. The FP-tree frequency set insertion algorithm is called. If the time series N have the same item name, the count of N is increased by 1, otherwise a new time node N is created. And set the count to 1, while linking to the parent node T, the traversal node is linked to the node with the same time. If P is not the empty set, recursively call the generative insertion algorithm. This completes the preprocessing of the time series**.**

2.2 Building a data storage model

After the time series is preprocessed, the time series stream segment in the basic serial port is converted into a transaction set. In order to run the association rule mining algorithm from this transaction set, a data storage model needs to be established. Based on the characteristics of the association rules to be mined and the characteristics of the data set [7], this

paper uses the SWFI-tree model, which can maintain two trees at the same time, divided into global SWFI-tree and local SWFI-tree. The structure is shown below:

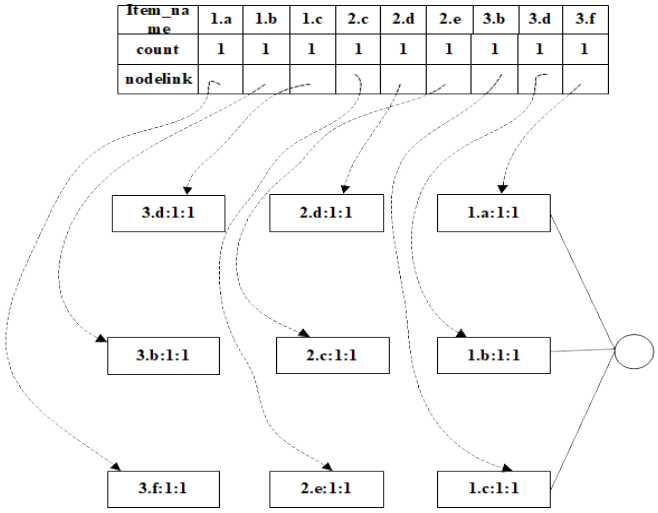

Fig.3 SWFI-tree model structure

With the established SWFI-tree model, it is not necessary to scan the data set multiple times. It is assumed that the original data set is shown in the following table:

Table 1 Raw data set

Transaction ID	Element items in a transaction	Minimum support
001	h,r,s,f,w	3
002	q,e,r,g,c,s,f,v,s	5
003	z	4
004	r,d,f,g,h	7
005	y,d,e,r,t,h	6
006	y,z,x,e,q,s,t,m	5

Using the SWFI-tree model only needs to scan the data set twice. The first scan counts the support of each item. The second scan sorts each transaction by the support count of the item, removing infrequent items, such as the table 2 shows:

Table 2 Scanned transaction data set

Transaction ID	Element items in a transaction	Filtering and reordering of transactions
001	h,r,s,f,w	h,s,f,w
002	q,e,r,g,c,s,f,v,s	r,v,s,q
003	z	z
004	r,d,f,g,h	r,h,f
005	y,d,e,r,t,h	y,r,t
006	y,z,x,e,q,s,t,m	z,x,y,s,t

Using the filtered and reordered transaction data sets obtained after scanning, a frequent pattern tree is generated to mine frequent patterns. Complete the establishment of the data storage model.

2.3 Cleaning monitoring data

The data used in this paper are derived from the real data collected during the network security monitoring and maintenance process. The data comes from the Internet central control device, and all the collected fault information and network operation status information are transmitted to the ground in a timely manner. The data flow is shown in the following figure:

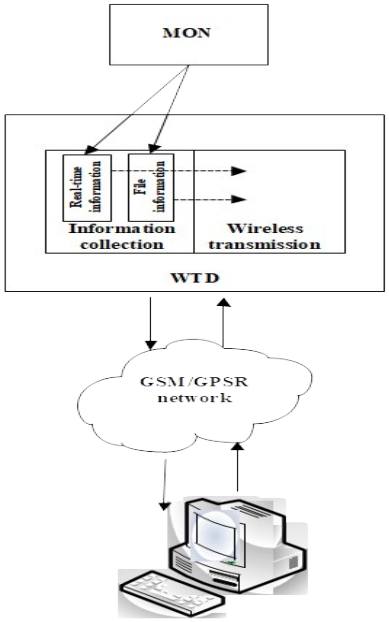

Fig. 4 Monitoring transmission data flow diagram

Under normal operating conditions, WTD sends real-time status data and screen display fault information collected by the network information collection and processing module every 60 seconds to the ground server. This real-time data does not require a ground response. When the main fault data is sent, the sent main fault record information is immediately sent to the ground, waiting for the response information from the ground, and the waiting time is generally set to seconds [8-9]. If it times out, the ground will be considered to have after receiving this major fault message, this major fault message will be resent. When a response is received, the data is considered to be sent successfully and the next data is sent.

Because network security generates a large amount of data during the monitoring process, it is difficult for rabbits to have some incomplete and dirty data. In addition, duplicate data, erroneous data, and garbled data are also important factors that reduce data quality. Therefore, in order to obtain more accurate mining results, this requires cleaning the original data, which can improve the accuracy of data mining. And effectiveness, on the other hand can also save the time required for mining. Therefore, two aspects have to be done in data cleaning: one is to remove invalid data; the other is data formatting. For the fault data set, mainly clean the fields of number type, null value and measurement type. During the cleaning

process, export the table data to a CSV format file in Oracle, upload the data file to HDFS using Hadoop commands, create a data table in Hive, and use HQL statements to clean the data. This completes the cleaning of the data.

2.4 Complete data association rule mining

After completing the steps of time series preprocessing, building a data storage model, and cleaning monitoring data, the time series data is symbolized and transformed into a transaction set. Because time series are fluid, particle swarms are used to limit the data flow. The particle swarm approach will have new data coming in and old data slipping out. Therefore, when mining association rules, it is necessary to be able to support incremental mining of increased data, and also to remove old data. The data set is dynamically changed, and the mining association rules are also dynamically changed. It is an association rule for multivariate time series data, and the amount of data is relatively large, so it is necessary to perform recursive calculations on the motion of ions:

$$\overline{V}_i^{k+1} = \overline{V}_i^k + c_1 * r_1(\overline{P}_i^k - \overline{X}_i^k) \quad (4)$$

In the above formula, the particle number is $i = 1, 2, ..., m$, k is the number of iterations, $\overline{V}_i^{k+1}$ is the particle speed, $\overline{V}_i^k$ is the speed of the particle immediately before, c_1 and r_1 are learning factors, and $\overline{P}_i^k$ and $\overline{X}_i^k$ are motion vectors. among them.

$$\overline{X}_i^{k+1} = \overline{X}_i^k + \overline{V}_i^{k+1} \quad (5)$$

The above two formulas will work together on the particle's motion position in the next step. Taking two-dimensional space as an example, the process from the initial position to the new position is shown in the following figure:

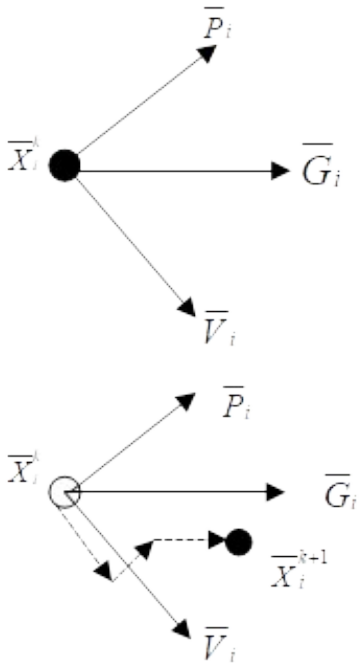

Fig. 5 Particle movement process diagram

Use particle traversal to obtain the support of the item set represented by the particle. Combining the characteristics of the data source and the basic algorithm, the linked list data structure is applied to the dynamic storage structure. Create a linked list header for each item in the full database, scan the set of items in order, and add the items in each item set to the end of the corresponding linked list [10-13]. For example, if the nth item set in the database is (50, 180, 17), then add the nth item set to the end of the linked list 50, the end of the linked list 108, and so on, and finally create 200 linked lists. That is, a sparse linked list, the process is as follows

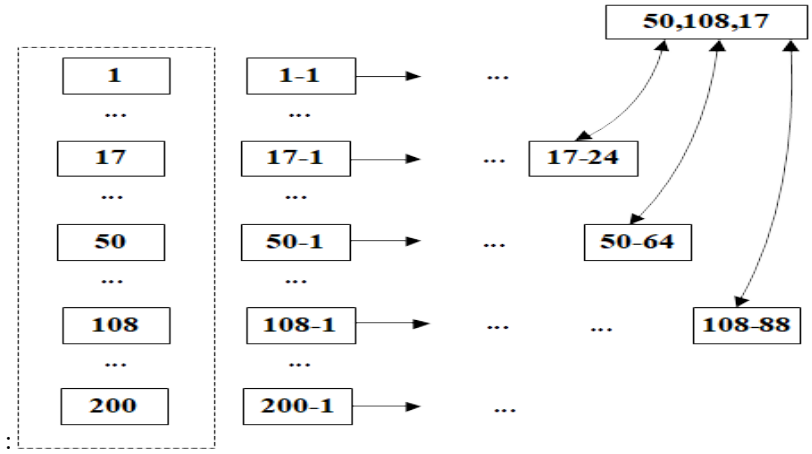

Fig. 6 Sparse linked list construction process

Take the nth item set (50, 180, 17) and 200 linked list headers in the linked list as an example, and start the search from the linked list 50. Suppose that the 64th item set found contains the 50 item. Similarly, linked lists 108 and 17 are Corresponds to 88,24, that is, the items before the 88th item set do not contain the item corresponding to the particle, and the 88th item set is directly found. If it contains (50, 108, 17), the particle support is increased by 1, otherwise the search continues, and the linked list (50, 108, 17) finds the next data, assuming corresponding 121, 90, 65, respectively, and directly find the 121st item set If it contains (50, 108, 17), the particle support is increased by 1, otherwise, the search continues. The linked list (50, 108, 17) finds the next data respectively, assuming that it corresponds to 121, 184, 121, and directly searches for the 184th Itemsets. If it contains (50, 108, 17), the particle support is increased by 1. This completes the mining of association rules for network security monitoring data based on time series.

3 Experiment

In order to verify the effectiveness of the association rule mining technology of network security monitoring data designed in this paper, it is necessary to design experiments to compare with the traditional association rule mining technology.

3.1 Experimental design

The experimental network structure is shown in the figure:

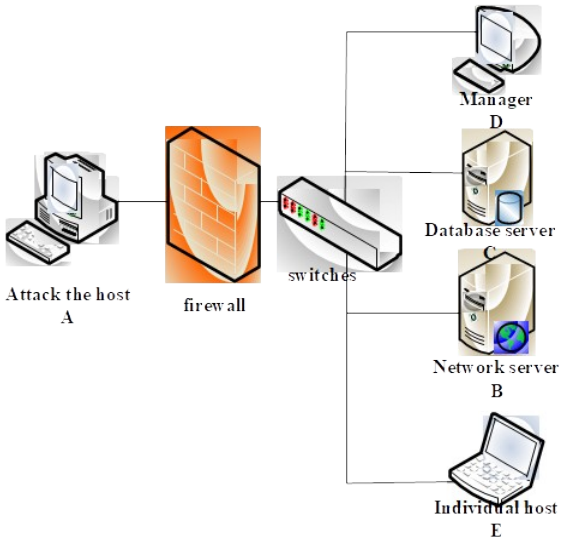

Fig. 7 Experimental network topology diagram

In the figure above, there are 5 nodes A, B, C, D, and E. Node A represents the intruder from the external network. The four nodes on the internal network inside the firewall are mainly responsible for providing network services, database services, server protection management, and storage. Important documents. The intranet access to the external network or communication between the intranet does not need to go through the firewall, but when the extranet nodes access the intranet nodes, they need to go through the firewall and can only access the network server. The data mining goal of this article is to analyze the network security monitoring data, and to attack the nodes of the internal network through some actual vulnerabilities in the internal network. The vulnerability settings of the internal network nodes are shown in Table 3:

Table3 Intranet node vulnerability settings

node	Hole number	Server or software	Elevated privileges
Network server B	Mc10-062	WWW	U-A
Network server B	Mc11-002	WWW	U-A
Database server C	Mc11-012	Oracle	O-A

Database server C	Mc10-065	Oracle	O-U
Manager D	Mc10-045	Windows	O-A
Individual host E	Mc10-006	Office	O-A

Under the above conditions, all the data in the network security monitoring database are known conditions, and the mining slope K value is estimated through experiments. The closer the K value is to 1.85, the more accurate the association rule mining is.

3.2 Analysis of results

Under the above experimental conditions, the method of this paper and the traditional method are used for data mining experiments, and the K value obtained by the two methods is compared with the set value, and the accuracy of data mining is obtained. The experimental results are shown in the following table:

Table4 Comparison of experimental results

Number of training sets	Number of test sets	K for traditional method	K for Method of this paper
50	250	2.81	2.13
100	200	2.47	2.01
150	150	2.15	1.89
200	100	1.69	1.84
250	50	1.20	1.79

Through experiments on 300 data sets, in 5 experiments, the number of different training and test sets was set, the standard K value was 1.85, and through calculation, the average accuracy of traditional technology was 67.84%. It is 91.06%, which shows that the technique in this paper has certain effectiveness in improving the accuracy of mining association rules.

4Concluding remarks

The traditional network security monitoring data association rule mining technology has low mining accuracy. Therefore, a time series-based network security monitoring data association rule mining technology is designed. Preprocess the time series, use the SWFI-tree structure to build a data storage model, clean the obtained data set, and finally use particle

swarm algorithm to recursively calculate the particle motion to complete the mining of the association rules of the monitoring data. The designed mining technology is used to conduct experiments with traditional technology. The experimental results show that the designed technology mining accuracy is 23.22% higher than the traditional technology.

5 Fund projects

Research on Energy and Delay Optimization of Delay Sensitive Wireless Sensor Networks, Project No .: Xiangjiao Tong [2018] 519

References

[1] AI Chu-han,JIANG Di,WU Jian-de: Research on Patent Text Data Mining Based on Topic Model and Association Rules. Journal of North University of China （Natural Science Edition), 40(06):524-530(2019).

[2] Wang Zhuofang,Zhao Huijun,Li Cong,etal: MINING ASSOCIATION RULES FOR MULTI-CLASS DIFFERENCE DATA OF WEB SERVICES BASED ON INTEREST MEASURE FUNCTION. Computer Applications and Software, ,36(12):60-65+105(2019).

[3] Sun Hong,Li Cunjin: Improvement of Data Mining Method Combining Genetic Algorithm and Association Rules. Journal of Data Acquisition & Processing, 34(05):863-871(2019).

[4] Liang Fan,Zhao Li: BIG DATA MINING AND ANALYSIS ALGORITHM BASED ON NEUTROSOPHIC FUZZY ASSOCIATION RULE GENERATION. Computer Applications and Software,36(10):285-292+298(2019).

[5] ZHANG Chao: Analysis of Two Optimization Strategies for Apriori Algorithm of Data Mining Algorithm Based on Association Rules. Journal of Shaoguan University,40(09):16-20(2019).

[6] Zhang Wumei,Dong Qiong: Fuzzy association rules mining method based on GSO optimization MF in uncertainty data. Application Research of Computers,36(08):2284-2288(2019).

[7] XU Yi,YUAN Bao-ping,XIA Yi-wei: Analysis of On-line Monitoring of Network Security in Power Grid Dispatching Data Network . Telecom Power Technologies, 36(11):178-179(2019).

[8] Bao Lijun: Application of Network Security Situational Awareness Platform Based on Big Data in the Field of Private Network. Journal of Information Securyity Research,5(02):168-175(2019).

[9] HUANG Kun,YANG Xudong,XU Ke: Network Security Situation Analysis of Large Power Grid Enterprises Based on Multi-Dimensional Heterogeneous Model. Electric Power Information and Communication Technology, 17(01):72-77(2019).

[10] WU Li-ming: Research on Network Security Situation Perception Technology and Evaluation Method Based on Big Data Analysis. Electronic Component and Information Technology,(09):49-51+101(2018).

[11] Mark Eshwar Lokanan: Data mining for statistical analysis of money laundering transactions. Journal of Money Laundering Control, 22(4):753-763(2019).

[12] Shou-Chieh Chao, Kate Ching-Ju Lin, Ming-Syan Chen: Flow Classification for Software-Defined Data Centers Using Stream Mining. IEEE Transactions on Services Computing, 12(1):105-116(2019).

[13] Jisheng Xia, Jinne Li, Pinliang Dong, et al: An ArcGIS add-in for spatiotemporal data mining in climate data. Earth Science Informatics, 13(4):1-6(2019).