Data flow network security strategies based on data mining

Bing Liu

College of Computer Science and Engineering, Changchun University of Technology, Jilin130012, China

Yang Zhao

Department of Electronic and Information Technology, Jiangmen Polytechnic, Jiangmen 529090,China

Yuanyuan Dang

College of Computer Science and Engineering, Changchun University of Technology, Jilin130012 ,China

**Corresponding author is Yuanyuan Dang, Email:34665065@qq.com*

Abstract

Based on the frequent pattern outlier algorithm, this paper establishes a mathematical model of data flow network security strategy based on data mining, and applies the improved outliers detection algorithm based on frequent pattern to the intrusion detection system. Experiments proved the effectiveness of the algorithm. Compared with the unimproved one, it has better detection accuracy, a shorter running time, so it is more suitable for the environment of network security detection.

Key words:DATA FLOW,NETWORK SECURITY,DATA MINING,INTRUSION DETECTION,FREQUENT PATTERN MINING

1. Introduction

With the development of information technology, the requirements for information security detection are higher and higher. Intrusion detection is an effective way to guarantee the network security.

The problem of network security intrusion detection has been studied by many people. In 2009, Mao Guojun[1] et al. proposed MaxFP-Tree, a new data

structure, based on the intrusion detection model of multidimensional data flow mining technology, and presented an efficient learning algorithm MaxFPinNDS. The results showed that the intrusion detection model is better. In 2010, Wu Feng [2] et al. established a data flow frequent pattern mining algorithm model by combining time exponential decay function with landmark window model. Experimental results show that compared with the similar algorithms, it has higher precision. In 2011, Sun Yanhua [3] et al established a vector mathematical model of the user behaviors on the basis of the CURE algorithm, and improved it, improving its detection rate of harmful behaviors. In 2012, Chen Wei [4] et al. established a data mining intrusion detection model by using the Apriori algorithm of correlation analysis. In 2013, Wang Hongxia [5] improved the security protection level of the system by combining the algorithm of association rules with sequence pattern mining algorithms. In 2014, Zhu Lin [6], et al. established an IDS network security defense model by using sliding window and data flow clustering. The experiment showed that the model can meet the requirements of detection. In 2015, Wang Ninan [7] established a mathematical model of data mining network security by using Rough Set as a tool, and provided a reference for the application of Rough Set data mining technology.

In this paper, the mathematical model of data flow network security strategy based on data mining is proposed by using the frequent pattern outlier algorithm, and the security strategy of the data stream network has been improved. Compared with the former one, the improved frequent pattern has better detection accuracy, a shorter running time, so it is more suitable for the environment of network security detection.

2. Data mining technology in intrusion detection

2.1. Related concepts of intrusion detection

At present, the most widely used security model is adaptive network security model (P2DR) [8], as shown in Figure 1, the P2DR model is composed of 4 parts: security strategy, detection, protection and response.

In 1987, D.E.Denning designed the abstract model of intrusion detection system [9], as shown in Figure 2, the model includes 3 parts: activity record, event generator and rule set.

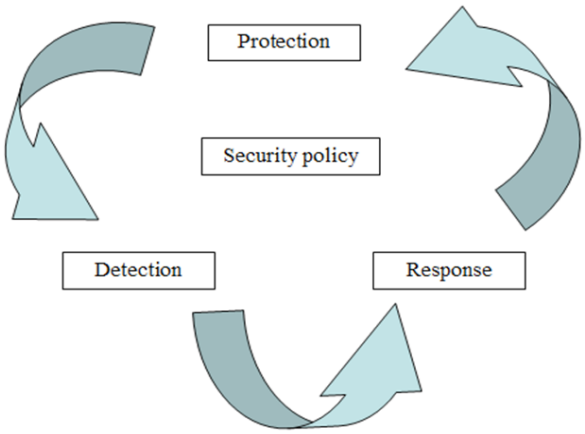

Figure1. P2DR model diagram

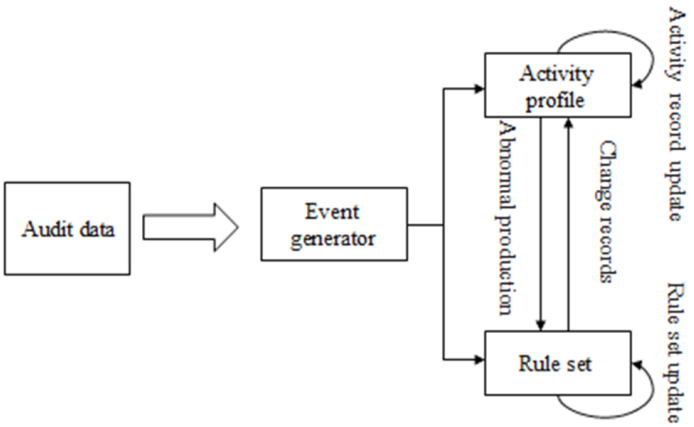

Figure 2. The abstract model of intrusion detection system

2.2 Overview of data mining

Data mining is to extract or «mine» knowledge from a large amount of data [10], the work flow is shown in Figure 3.

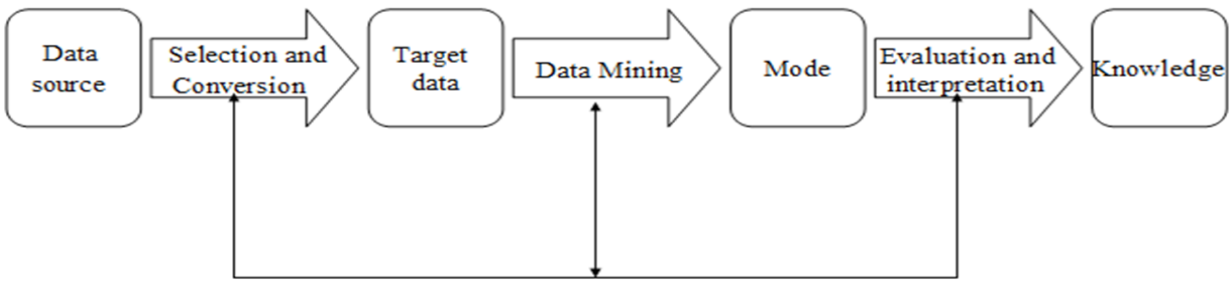

Figure 3. The steps of data mining

3. Data flow detection technology based on frequent pattern mining

3.1. Problem description based on frequent pattern mining

Because the frequent set stems from the association rules, the related description of the association rules is given [11].

Assuming $I=\{i_1,i_2,\cdots,i_m\}$ to be a binary set, each of which is called a term. Assuming D to be a set of transaction T, then T is also a set of terms, satisfying $T\subseteq I$. Assuming X to be the set of term in the set of I, if $X\subseteq T$, then it can be called X is contained in T.

If $X\Rightarrow Y$ an association rule, in which, $X\subset I$ $Y\subset I$, $X\cap Y=\Phi$,then the support of Rule $X\Rightarrow Y$ in the database D is the ratio of all transaction number of D in the database D, which is denoted by $S(X\Rightarrow Y)$, then

$$S(X\Rightarrow Y)=\left|\{T:\mathrm{X}\cap Y\subseteq T,T\in D\}\right|/|D| \quad (1)$$

The credibility of Rule $X\Rightarrow Y$ is the ratio of transaction number of X contained in Φ, denoted by $C(X\Rightarrow Y)$, then

$$C(X\Rightarrow Y)=\left|\{T:\mathrm{X}\cap Y\subseteq T,T\in D\}\right|/\left|T:\mathrm{X}\subseteq T,T\in D\right| \quad (2)$$

3.2. Introduction to the frequent pattern algorithm for outlier based on data flow

FindFPOF algorithm is the most typical algorithm in the frequent outlier mining algorithm [12]. In the FindFPOF algorithm, the outlier measure factor in the FPOF algorithm is used to calculate all the data of FPOF, and then all the data are sorted according to the value; the first n numbers are regarded as the outliers.

Assuming $D=\{t_1,t_2,\cdots,t_n\}$ is the data set of n audit data, FPS (D, minisupport) is a frequent pattern sets for a given support threshold. The FPOF of each audit data t can be defined as:

$$FPOF=\frac{\sum\mathrm{support}(P)}{\|FPS(D,\mathrm{minisupport})\|} \quad (3)$$

Wherein, $\|FPS(D,\mathrm{minisupport})\|$ refers to the number of each frequent pattern, $\sum\mathrm{support}(P)$ indicates the support of the frequent pattern P in the transaction t, P is the frequent pattern.

4. Application of outlier algorithm based on frequent pattern in intrusion detection

4.1. Introduction to the intrusion detection model based on frequent pattern outlier algorithm (FindFPOF)

Outlier mining can find out most abnormal data deviated, the main purpose of intrusion detection system is to find out the data deviated, so as to detect the intrusion behavior.

Intrusion detection system model is divided into the following four parts: data packet capture module,data flow mining module,data management module and data detection module, as shown in Figure 4.

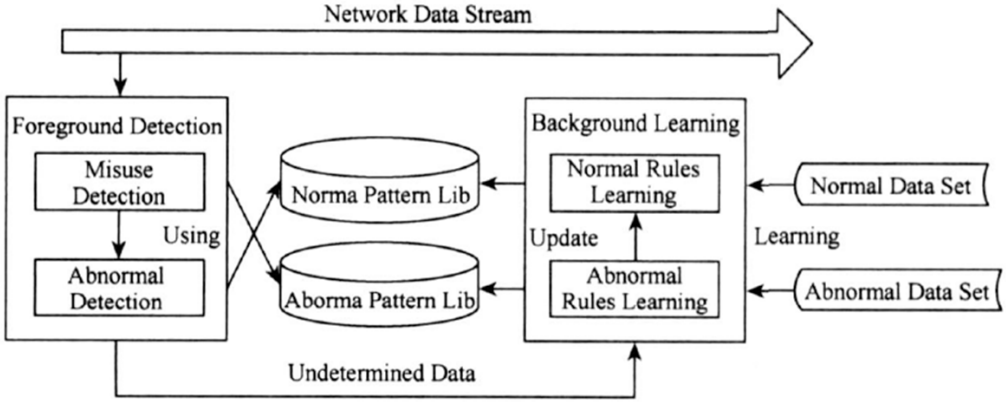

Figure 4. Design of intrusion detection model

4.2. The improved outlier detection algorithm based on frequent pattern (NFPOF algorithm)

4.2.1. Concept definition

Assuming some network audit data set to be D, $D=\{D_1,D_2,\cdots,D_i,\cdots,D_n\}$,$1\le i\le n$, in which, D_i indicates a datum in set D, $I=\{I_1,I_2,\cdots,I_n\}$ is the set of all terms.

Definition 1 (Pattern X)Pattern X is assumed to have a set of terms X, satisfying $X\subseteq I$, then X is called to be the pattern of data set, if Data D_i can satisfy X in each term of D_i, then Data D_i satisfies Pattern X.

Definition 2 (Support of Patter X)

$$\mathrm{supp}(X)=\frac{\|X\|}{\|D\|} \quad (4)$$

Wherein, $\|X\|$ is the occurrence frequency of the pattern X contained in all data, $\|D\|$ is the total number of data contained in the whole data sets.

Definition 3 (Frequent pattern P), Pattern P satisfies $\text{supp}(P) \geq \min\text{sup}$, then P is called the frequent pattern, in which, $\min\text{sup}$ is the minimum support.

Definition 4 (Frequent weight of Pattern P)

$$W(P) = \text{supp}(P)\left(\frac{|P|}{k} + 1\right) \tag{5}$$

Wherein, $W(P)$ is the frequent weight of Pattern P, $|P|$ is the number of attributes contained in Pattern P, k is the data space dimension.

Definition 5 (new frequent pattern outlier factor)

$$NFPOF = \frac{\sum W(P)}{\left\|\{p|\sup(p) \geq \min\sup\}\right\|} \tag{6}$$

Wherein, P is the pattern contained in $W(P)$, $\left\|\{p|\sup(p) \geq \min\sup\}\right\|$ indicates the number of frequent patterns contained in D.

Definition 6 (abnormal attribute), assuming the attribute to be A ,its value to be a_i, the value range to be $\{a_1, a_2, \cdots, a_i, \cdots, a_n\}$, definition S_i to be a set, which contains all the frequent pattern P_j of a_i, $Sum(a_i)$ records the total weight of frequent pattern P_j contained in S_i. In which, a_m is the maximum of $Sum(a_m)$.

Then,

$$Nor(a_i) = \frac{Sum(a_i)}{Sum(a_m)} \tag{7}$$

If $Nor(a_i) \leq \min Nor$,in which, $\min Nor$ is the minimum threshold value set by the user, if the credibility of abnormal attribute is lower than the minimum threshold, the attribute is called the abnormal attribute.

4.2.2. Algorithm process

The algorithm is divided into three steps: 1 the data are pre-processed; 2 the frequent pattern is found and the weight of frequent pattern is recorded; 3 the weighted frequent pattern outlier factors of the data are calculated, the outlier and the abnormal attribute are detected. The algorithm process is shown in Figure 5:

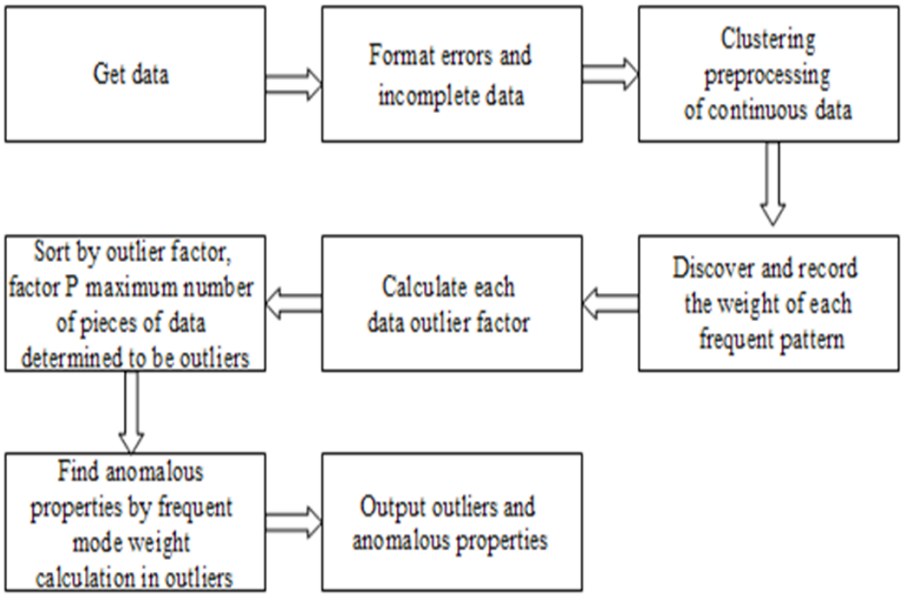

Figure 5. NFPOF algorithm flow chart

5. Experimental results and analysis

5.1. Preparation of experimental data

Experimental data used are the KDDcup99 data set. Table 1 presents the specific distribution of attack data in the primary data set:

Table 1. Attack classification of test datasets

Attack Types	Test data set	
	Amount	Percentage/%
Normal	96274	20.40
Probe(1)	3107	0.66
DOS(2)	371468	78.71
U2R(3)	47	0.01
R2L(4)	1020	0.22

In the data set, there are 22 attribute values, and all attributes are specifically shown in Table 2:

Table 2. The attribute of data set

No.	Feature	Description	Type
1	duration	The duration of the connection	C
2	protocol_type	Protocol Type	D
3	service	Types of host network services	D
4	flag	Connection Status	D
5	src_bytes	Count byte of the source port to the destination port	C
6	dst_bytes	Count byte of the destination port to the source port	C
7	land	A connection port is the same as 1, otherwise 0	D
8	wrong fragment	The number of slices in a connection error	C
9	urgent	The number of package of emergency in a connections	C
10	hot	Number of "hot" indicator	C
11	num_failed_logins	Number of failed login	C
12	logged_in	Successful login is 1, otherwise 0	D
13	num_compromised	The number of conditions to meet the attack	C
14	root_shell	Gain superuser shell is 1, otherwise 0	D
15	su_attempted	Get "suroot" is 1, otherwise 0	D
16	num_root	Number of root access	C
17	num_file_creations	Number of file creation operation	C
18	num_shells	Number of shell prompt	C
19	num_access_files	Number of the access control file operations	C
20	num_outbound_cmds	Number of ftp session band command	C
21	is_hot_login	Login belong to "hot" list is 1, otherwise 0	D
22	is_guest_login	Login user "guest" is 1, otherwise 0	D

C: Continuous,D: Discrete

Wherein, all the attributes can be divided into two categories, the first category is the attributes from No.1 to 9, they are the basic attribute of the network connection, No.10 to 22 are the the attribute of net-

work connection based on the content. In which, the specific value of the flag value of Attribute 4 is shown in Table 3:

Table 3. The values of flag

Flag Value	Description
SF	TCP session is normally completed
REJ	One issue SYN, RST party giving as a response
S1	One issue SYN, the other party do not answer
S2	After connection establishment, there is no further data exchange
S3	After connection establishment, initiating close connection
S4	Received SYNACK, but not receive the corresponding SYN
RSTOSn	When the connection status is n, initiator is reset
RSTRSn	When the connection status is n, the recipient is reset
SS	Received all been semi-closed network connection
SH	In the absence of receipt of SYNACK, all connections are closed in the state 0
SHR	In the absence of receipt of the initial SYN, all in a state of connection 4 is closed
OOS1	The initial SYN and SYNACK not exactly match
OOS2	Depending on the sequence number, retransmit the initial SYN

5.2. Experiment process

In this paper, the optimized PSO-NR algorithm [13] is used; the characteristic number is selected so as to conduct its discrete processing [14]. In order to test the effectiveness of the algorithm, the three algorithms of FindFPOF, Evolutionary Outlier Search and NFPOF are compared for their accuracy and running time.

The detection rate is the ratio of the number of real detected outliers and the number of outliers in the data set. Namely,

$$Dr = \frac{O}{N} \times 100\% \quad (8)$$

Wherein, Dr is the detection rate, O is the existing number of outliers, N is total outliers for event base.

The false drop rate is the ratio of the number of error detection and the total amount of the non-abnormal data of the data set, namely,

$$Edr = \frac{X}{N} \times 100\% \quad (9)$$

Wherein, Edr is the false alarm rate, X is the number of events of the false alarm, N is total amount of outliers in the event base.

Before the algorithm experiment, it is needed to the train the parameters. First, minsup is trained; the training set is trained by using the prepared 5 sets of parameters. The improved frequent weighted factor algorithm is used for the training; the results of minsup value are shown in Table 4:

Table 4. Training of minsup's value

minsup	KDDcup99	
	Dr	Edr
0.35	95.4	2.94
0.55	93.2	3.24
0.60	96.5	3.73
0.75	54.7	4.18
0.85	39.2	9.35

Dr: Detection rate, Edr: Error detection rate

It can be known from Table 4, when the minsup is not sensitive to the set value and it has good stability, its false drop rate and the detection rate are better,

NFPOF algorithm also has good stability.

According to the FindFPOF experimental algorithm, the minsup value is fixed, the minNor value is trained, the training results are shown in Table 5:

Table 5. Training of minNor value

minNor	KDDcup99	
	Dr	Edr
0.15	19.6	0.12
0.30	58.9	2.24
0.55	88.5	3.78
0.70	92.4	13.22

Dr: Detection rate, Edr: Error detection rate

Wherein, $\min sup = 0.40, l = 5$, $\min Nor = 0.6$, data sets are the subset data of network intrusion detection dataset KDD CUP - 99.

The abnormal attribute detection of the algorithm and false drop cases are shown in Table 6:

Table 6. Abnormal properties test results

Data set	Attribute detection	
	Dr	Edr
DOS	88.3	4.4
PROBE	86.2	3.6
U2R	82.1	2.9
R2L	87.7	4.1

Dr: Detection rate, Edr: Error detection rate

The comparison of the false drop rate and the detection rate of NFPOF algorithm, FindFPOF algorithm, EOS algorithm is shown in Table 7:

Table 7. The FindFPOF algorithm ,the EOS algorithm and the NFPOF algorithm differences in the control table

Data set	FindFPOF		EOS		NFPOF	
	Dr	Edr	Dr	Edr	Dr	Edr
DOS	93.2	4.79	95.2	4.14	95.8	3.19
PROBE	92.5	2.28	95.6	3.12	95.3	3.14
U2R	91.5	2.59	94.8	1.93	94.3	2.03
R2L	93.4	4.44	94.3	3.22	95.0	3.11

Dr: Detection rate, Edr: Error detection rate,EOS: Evolutionary Outlier Search

As shown in Figures 6 and 7, the detection rate of the three methods are over 90%, the accuracy of NFPOF algorithm is over 90%, slightly higher than the former algorithm, its running time is the least, so it is obviously better than the EOS algorithm, and it can meet the actual use demand. The experimental results show that the NFPOF algorithm can more accurately detect abnormal attributes in guaranteeing higher accuracy and a shorter operation time.

Through the calculation of the weighted frequent outlier factor of the security data, the corresponding outlier attributes are obtained, if there is good effect, and then it is considered to be feasible. The accuracy calculation formula is:

$$A = \frac{P}{Q} \tag{10}$$

In which, A is the accuracy, P is the correct outliers, Q is the outliers detected.

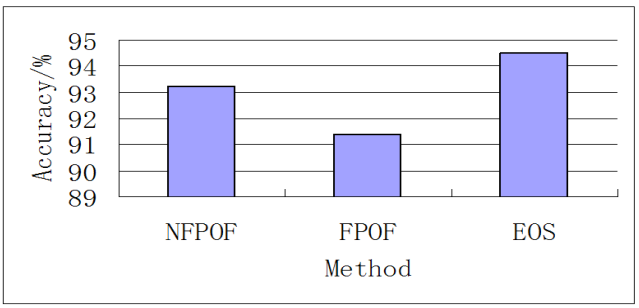

Figure 6. The accuracy of algorithms comparison

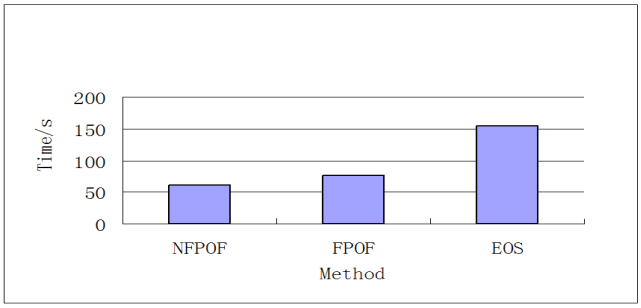

Figure 7. The time of algorithms comparison

Experiments showed that under the condition of complex space-time, NFPOF algorithm can effectively detect the abnormal attribute of the data, and then conduct intrusion detection, and its effect is better than that of FindFPOF algorithm and that of EOS algorithm.

6. Conclusions

This paper studied the outlier detection algorithm based on frequent pattern and the architecture of the intrusion detection system, and the improved outlier detection algorithm based on frequent pattern was applied to intrusion detection system. The experimental results proved the effectiveness of the proposed algorithm, compared with the previous one, its detection accuracy is improved, and the running time is reduced, so it is more suitable for the environment of network security detection.

References

1. Mao Guojun,Zong Dongjun(2009)Intrusion Detection Model and Algorithm based on Multi dimensional Data Stream Mining Technology. *Journal of Computer Research and Development*, 46(4),p.p.602-609.
2. WU Feng, ZHONG Yan,WU Quan-Yuan(2010) Data stream frequent pattern mining based on time decay model. *Acta Automatica Sinica*, 36(5),p.p.77-86.
3. SUN Yan-hua,LI Jie,LI Jian(2011)Network user behavior analysis based on CURE algorithm. *Computer Technology and Development*, 21(9),p.p.16-26.
4. XU Gang,LI Duan(2012)Research on Intrusion Detection System Based on Data Mining Technology. *Microprocessors*, 9(5),p.p.202-211.
5. NING BIN(2008)Research on Intrusion Detection System Based on Data Mining Technology. *Microcomputer Information*, 27(6),p.p.27-39.
6. ZHU Lin,ZHU Canshi(2014)Application of sliding window data stream clustering algorithm in IDS.*Computer Engineering and Applications*,50(1),p.p.88-99.
7. WANG ni-nan(2015)Research on Application of data mining technology in network security based on Set Rough. *Electronic Technology & Software Engineering*, 37(7), p.p.102-109.
8. YANG Hong-yu,ZHU Dan,XIE Feng(2009) Overview of Intrusion Detection Research.*Journal of University of Electronic Science and Technology* ,38(5),p.p.587-595.
9. Denning D.E(1987)Intrusion-Detection Model.*IEEE Transaction on Software Engineering*,13(2),p.p.222-232.
10. FraserSJ, MikulaPA, LeeMF, DicksonBL, KinnerslyE(2006)Data Mining - Ordered Vector Quantisation and Examples of Its Application to Mine Geotechnical Data Sets.*Proc. Conf. on Mining Geology* , Sydney,Australia,p.p.259-268.
11. HE Jun, LIU Hong-Yan, DU Xiao-Yong(2007) Mining Association Rules.*Journal of Software*, 18(11),p.p.2752-2765.
12. HE Zhenyou, XU Xiaofei, HUNG Zhexue(2005) FP-Outlier: Frequent Pattern based Outlier Detection,*Computer Science and Information System*, 2(1),p.p.103-118.
13. CHEN Shitao, CHEN Guolong, GUOWenzhong(2010)Feature Selection of the Intrusion Detection Data Based on Particle Swarm Optimization and Neighborhood Reduction. *Journal of Computer Research and Development*,47(7),p.p.1261-1267.
14. SANG Yu,YAN De-qin,LIU Lei(2008)Continuous Attribute Discretization Imp-Chi2 Algorithm. *Computer Engineering*,34(17),p.p.39-41.