Network Safety Policy Research for Analyzing Static and Dynamic Traffic Volume on the Basis of Data Mining

Jing Xu[*]

Ningbo Dahongying University, Ningbo 315175, China

Abstract: With popularization of network, higher requirement is proposed to intrusion detection system IDS for network safety consideration. The traditional electronic data processing is combined with safety auditing, which has become a necessary part of constituting integrated network safety technology at present, thus the methods as optimal matching mode and statistics, etc of intrusion detection system shall be adopted. This project shall respectively make comprehensive description to current situations of intrusion detection research via the aspects of intrusion detection research method (anomaly detection, misuse detection), intrusion detection system monitoring object (network based, host based), to comprehensively analyze the impact of intrusion detection system to system architecture. On this basis a network-based anomaly intrusion detection system NAIDS is designed to network anomaly intrusion, the association rules mining and frequent scenario mining are adopted to scan the intrusion characteristic, through static mining mode and dynamic mining mode, safety detection is conducted at single layer and domain layer, new type attack can be detected via improved NAIDS system. Next, NAIDS system performance shall be evaluated by aiming at various intrusion data. Generally speaking, the system performance can detect the rejection service attack and detection attack.

Keywords: Intrusion detection, data mining, anomaly detection, misuse detection, NAIDS system.

1. PRINCIPLE AND METHOD ANALYSIS OF DATA MINING IN INTRUSION DETECTION

NAIDS (network anomaly detection system) is the system prototype that can realize the proposed method in the available operation environment, the dynamic slide window is proposed at real time detection process to satisfy the real time requirement, classification engine is proposed to further reduce false report rate, decision tree algorithm is proposed to detect new type attack and improve application, and some hypothesis are proposed to above three main issues. The research method of NAIDS adopts data mining ideology and technology, the so called data mining is to mine the potential and unknown and valuable information from massive data. The data mining algorithm is adopted to extract safety related system characteristic attribute, and apply data mining technology to intrusion detection field, and create safety event classification model as per system characteristic attribute, to automatically identify safety event, and data mining based intrusion detection model can be established. This model includes a series of processes of safety event auditing data, such as data acquisition, pre-processing, characteristic variable selection, algorithm comparison, mining result processing and result visualization, etc, wherein, data mining is the key of whole process. This project deems that intrusion detection essentially belongs to data analysis process, data mining technology can mine the normal or intrusion behavior mode suspected association rule shall be acquired via line incremental mining algorithm, it is classified into unknown from a large number of auditing data. NAIDS system is composed of two operation mode stages as training and detection, during operation at detection status, current event, normal event or known attack type. In the anomaly event, if the attack type data once occurred during training stage, the attack type name can be further designated.

Data mining method is introduced into intrusion detection system to analyze the data, independent detection model is established through training, to relieve burden of detection system, and improve safety of detection system. The methods as association rule mining, frequent scenario mode mining, cluster mining, etc are mainly used for intrusion detection, in NAIDS, the usual data mining can be divided into: mining association rules, data class if canon, clustering analysis, mining sequential pattern. Also, data mining algorithm can also be divided into association analysis algorithm [1], data classification algorithm, clustering analysis algorithm [2], and sequential analysis algorithm [3, 4]. At present the association, sequence, classification and clustering type, etc are main data mining algorithm [5, 6] for intrusion detection, hereinafter these algorithms shall be briefly introduced.

1.2. Association Analysis Algorithm

The association rule form is $X \rightarrow Y_{[c,\ s],}$ herein $X \ \cup Y$ $X \cap Y = \Phi$, s is support degree of, c is confidence coefficient of the rule and is defined as SxUy/Sx, and database knowledge association rule is to describe the impact degree among each attribute.

1874-1290/14 2014 Bentham Open

1.3. Sequence Analysis Algorithm

The sequence analysis is used to mine the correlation among different data logs, its purpose is to seek out the big sequence in the transaction database satisfying minimum support degree designated by user, the algorithm can be divided into five steps as sequencing, big data item analysis, transformation, sequence modeling, sequence maximization, wherein, sequence modeling is the key of the algorithm.

1.4. Classification Algorithm

Data classification is to extract the characteristic attribute of data item in the database, the created classification model can map data item of database to any given class. Data classification processing is: acquire training data, centralized training data, analyze training data set, and create optimized classification model.

1.5. Clustering Algorithm

The analysis to data object set without knowing data rule is called clustering algorithm, the data object is divided into multiple classes or clusters, the object in the same cluster shall have very high similarity, the object of different clusters has very low similarity. The similarity mentioned herein is calculated subject to attribute value of description object, and it is usually measured by distance.

It is necessary to mine some attribute from auditing data to describe the data, some other shall only provide auxiliary information which means the important degree of attribute for describing data is different. The audit log of quantized data during mining process is divided into two sections by threshold of support degree and confidence degree, the problem of sharp boundary brought from such zoning shall impact performance of IDS. Therefore, Bridges adopted the combination of data mining algorithm (association rule mining algorithm and frequent scenario mining algorithm) and fuzzy logic to develop IDS. The attack behavior of multi level approximate mining to some application layer can be completed at one time connection, since its support degree is not higher than given threshold, it is very hard to detect such attack via association rule and frequent scenario rule. Nevertheless, if low support degree is adopted during mining process, related mode acquired shall be a large number of service types with high frequency, Lee and Stolfo [7] propose multi level approximate mining technology to resolve this problem. The labelled data can be established by simulating intrusion behavior, the shortcoming is that it can only simulate the attack behavior of known [8] type, therefore Eskin, etc proposed a non-monitoring anomaly detection algorithm, to adopt distance measure (k-center point algorithm) to cluster the data instance into cluster set, if the data is clustered, some small cluster shall be marked as anomaly instance.

2. OPERATION PRINCIPLE AND SYSTEM ARCHITECTURE OF MAIDS

MAIDS is the anomaly detection system by aiming at tcpdump data, the training data without attack type data in a given protected network environment is called pure data, which is used to establish network or used for user normal behavior profile archive (Normal behavior Prof 1e). The profile archive of network behavior mode is to establish an association database rule, to mine the pre-processing network connection log and seek out the suspected mode in the profile archive and acquire suspected association rule and classify them.

The work mode of MAIDS system is divided into two stages as training and detection, the training data of training stage shall adopt normal network connection log as main part mixed by some marked attack type data, the training stage is shown in Fig. (**1**), detection stage in Fig. (**2**), Fig. (**3**) is the operation process of system.

The training stage is composed of modelling period classification engine period. Before the system is and training classification engine period. Normal used to detect intrusion, the training stage shall only behavior profile archive is established at modelling be executed by one time. period; classifier is established at training classification engine period. Before the system is used to detect intrusion, the training stage shall only be executed by one time.

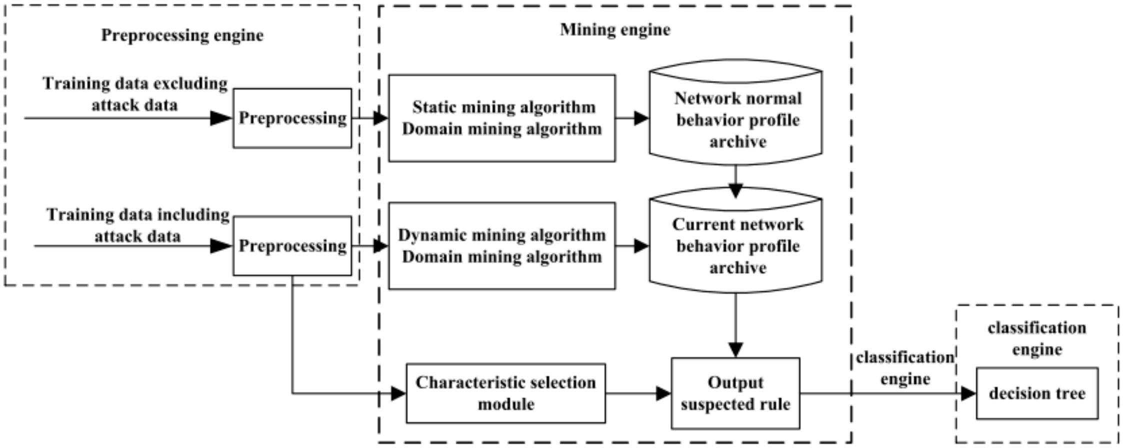

Preprocessing engine
Training data excluding attack data
Preprocessing
Training data including attack data
Preprocessing
Mining engine
Static mining algorithm Domain mining algorithm
Network normal behavior profile archive
Dynamic mining algorithm Domain mining algorithm
Current network behavior profile archive
Characteristic selection module
Output suspected rule
classification engine
classification engine
decision tree

Fig. (1). MAIDS training stage.

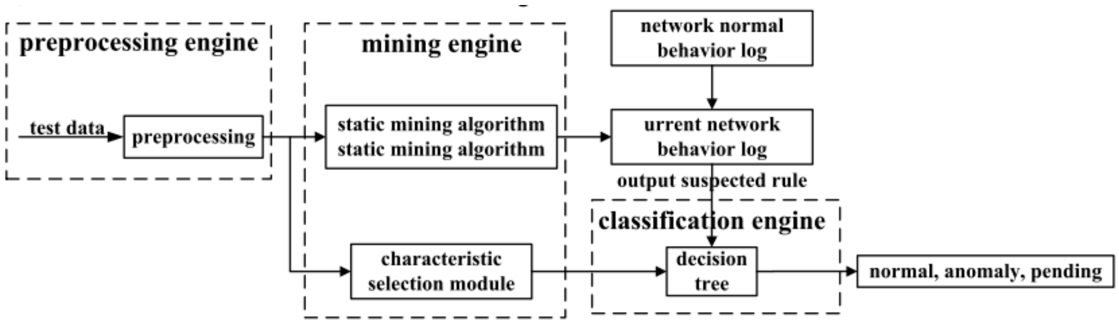


Fig. (2). Detection stage of MAIDS.

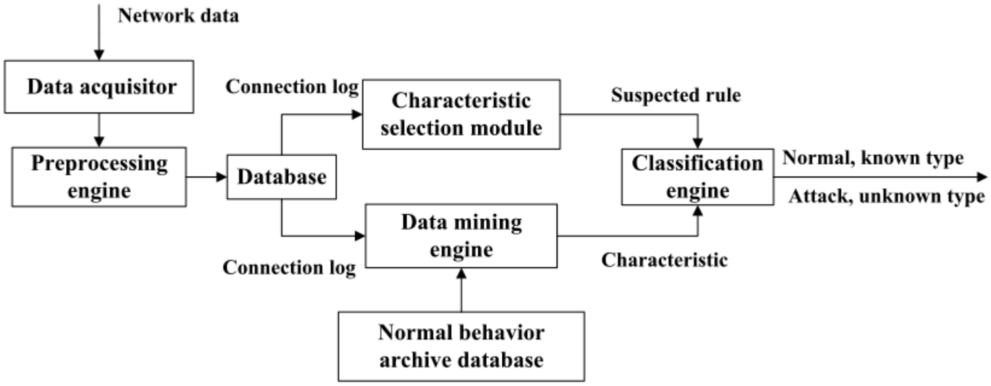


Fig. (3). NAIDS system operation process.

At detection stage, normal behavior profile archive has been acquired at training stage, and classification engine has been trained well, the real time analysis network connection log is adopted to seek out the rule set which is not contained in normal profile archive or has fairly great deviation with normal one. In classification engine of stage 1, the characteristic extracted by characteristic selection module is inputted to further divide suspected rule into normal, anomaly and unknown event. If such data type in anomaly event has occurred at classification and engine training period of training stage, it can be marked by the name of next attack type. For normal event they shall be filtered from suspected rule set without transmission to system safety staff. The classifier of unknown event cannot at first judge it as unknown type attack, in order to minimize missing report rate, it is deemed as unknown type attack and contained in the suspected rule set, and transmitted to system safety staff for analysis purpose.

NAIDS system is generally composed of three engines as pre-processing, mining and classification, in Fig. (**1**), Fig. (**2**), the position of each engine in system operation stage is marked by dotted line box, each function engine can all be presented in some extent at two stages: Update mechanism of normal behavior profile archive is shown in Fig. (**4**):

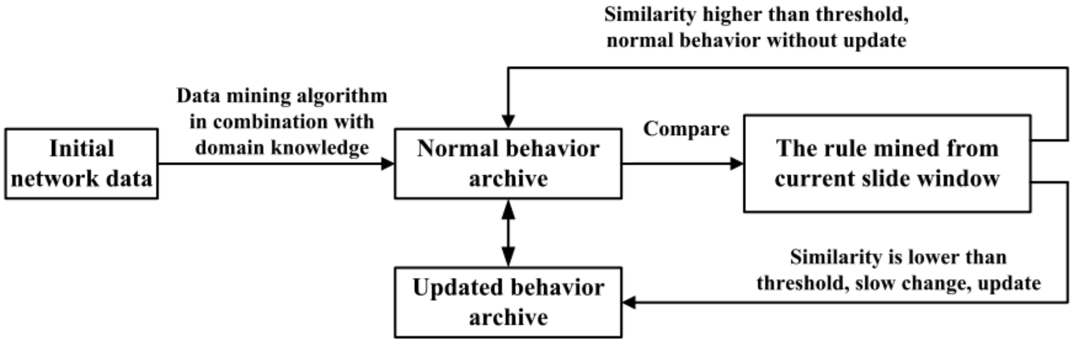

Similarity higher than threshold, normal behavior without update
Initial network data
Data mining algorithm in combination with domain knowledge
Normal behavior archive
Compare
The rule mined from current slide window
Updated behavior archive
Similarity is lower than threshold, slow change, update

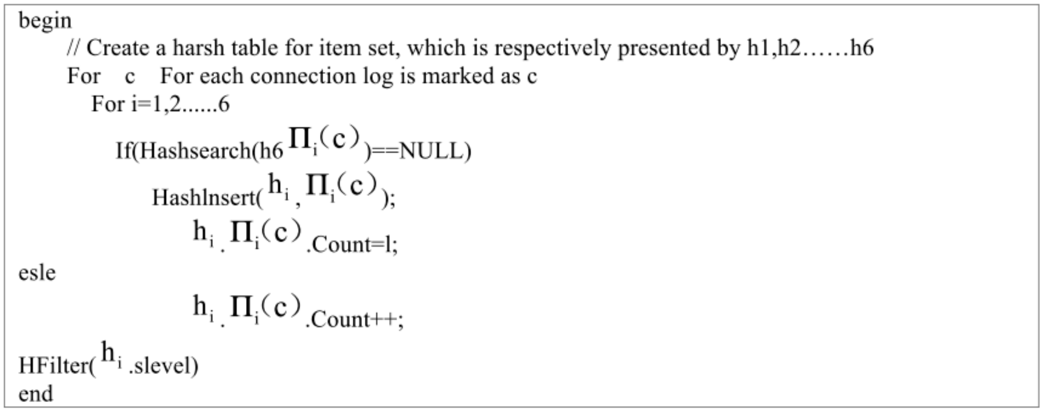
Fig. (4). Update mechanism of normal behavior profile archive.

The slide time window method is adopted to update benchmark behavior profile archive and resolve the auditing data of long time archive and mine the data. The support degree of certain frequent item set shall decrease with time past, while support degree of certain item set in the negative margin shall gradually increase. Discarding some frequent item set and containing new item set shall exclude some data log which is deemed as out of data from the slide window. The normal behavior profile archive self adaptability frame proposed for above behavior profile archive update time and update mechanism ideology is shown in Fig. (**4**).

3. NAIDS SYSTEM METHOD RESEARCH AND TEST DATA ANALYSIS

Hereinafter two mining modes and integrated realization of two levels mining NAIDS system shall be provided. Static mining and dynamic mining mode are two mining modes; single layer and domain layer mining are two mining layers. The detailed realization algorithm of core module is presented by false code, detailed test shall be done at DARPA 1998. DARPA 1999 intrusion detection evaluation data set.

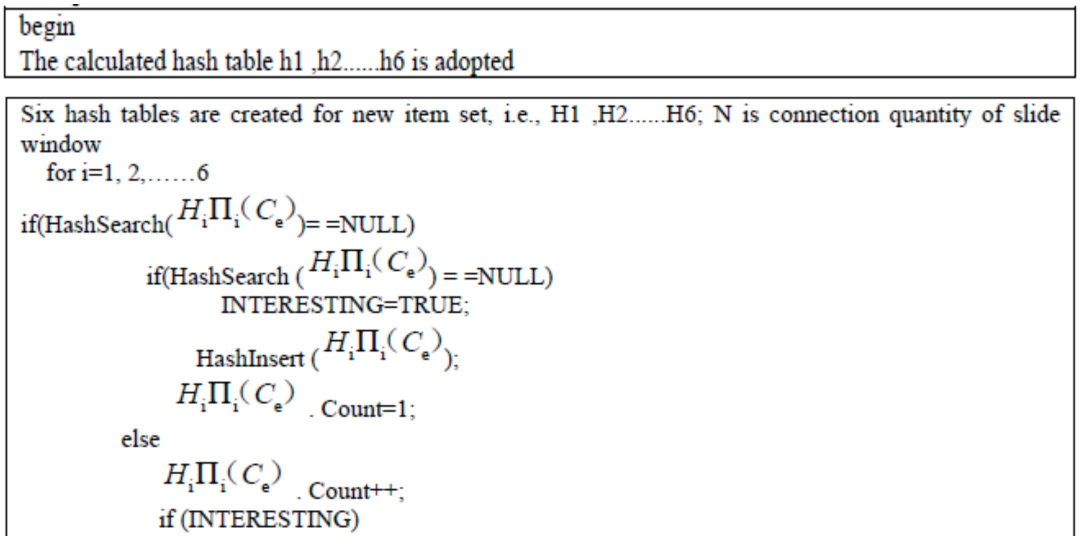
The static mining is used to seek out the rule set with its support degree greater than certain predefined threshold in the whole data set. At training stage the static mining process can be completed via offline, the pure data without containing intrusion data in some network environment is the data object with pertinence, the data is acquired by tcpdump, and pre-processed prior to mining. The result of static mining process is the normal behavior profile archive for delineating network or user normal behavior mode.



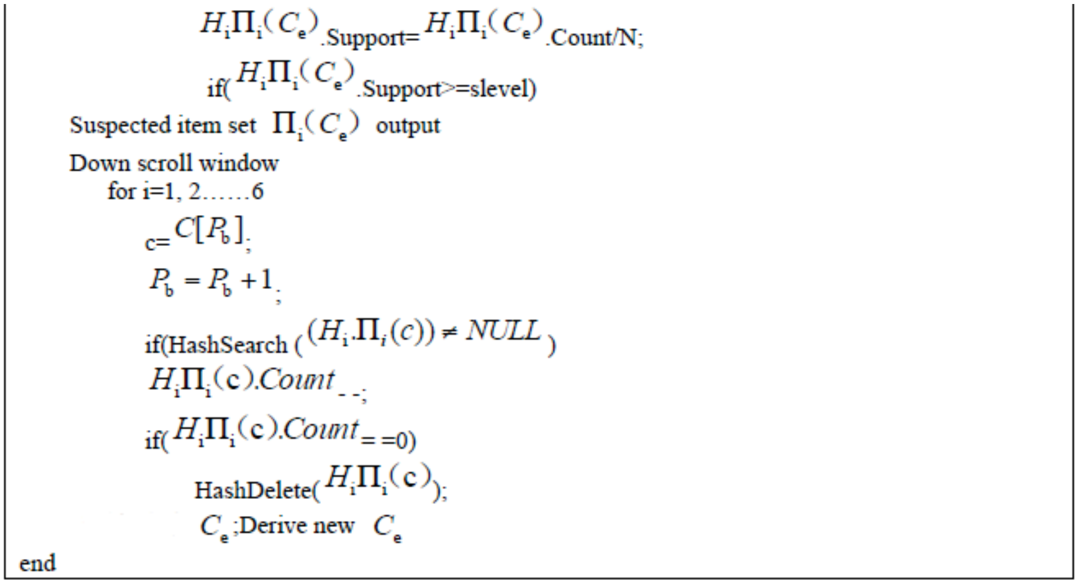
```
begin
      // Create a harsh table for item set, which is respectively presented by h1,h2……h6
      For    c    For each connection log is marked as c
         For i=1,2......6
               If(Hashsearch(h6 $\Pi_i(c)$)==NULL)
                    HashInsert($h_i$, $\Pi_i(c)$);
                         $h_i . \Pi_i(c)$.Count=l;
esle
                         $h_i . \Pi_i(c)$.Count++;
HFilter($h_i$ .slevel)
end
```

3.1. Static Mining Algorithm

Dynamic mining adopts slide window method to seek out suspected rule set via online incremental mining, next the detailed algorithm and false code description shall be analyzed. In order to seek out the suspected rule which is not consistent with normal behavior profile archive, In order to find out the suspected rule inconsistent with normal behavior profile archive, two operation stages of system shall be executed by dynamic mining process. During dynamic mining process it only needs to scan each connection log by one time, and the time window can also refer to volume window.



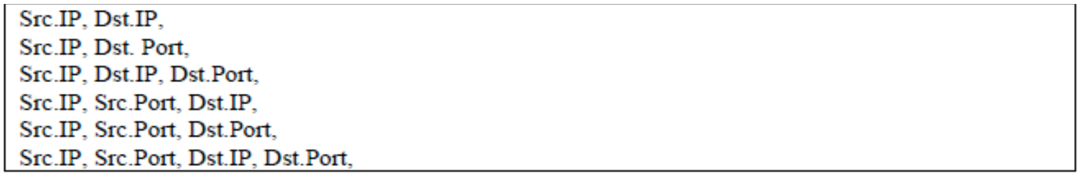
```
begin
The calculated hash table h1 ,h2......h6 is adopted

Six hash tables are created for new item set, i.e., H1 ,H2......H6; N is connection quantity of slide window
   for i=1, 2,……6
if(HashSearch($H_i \Pi_i(C_e)$)= =NULL)
               if(HashSearch ($H_i \Pi_i(C_e)$) = =NULL)
                    INTERESTING=TRUE;
                  HashInsert ($H_i \Pi_i(C_e)$);
             $H_i \Pi_i(C_e)$ . Count=1;
          else
             $H_i \Pi_i(C_e)$ . Count++;
             if (INTERESTING)
```



```
            $H_i\Pi_i(C_e)$.Support=$H_i\Pi_i(C_e)$.Count/N;
            if($H_i\Pi_i(C_e)$.Support>=slevel)
      Suspected item set $\Pi_i(C_e)$ output
      Down scroll window
         for i=1, 2……6
            c=$C[P_b]$;
            $P_b = P_b + 1$;
            if(HashSearch ($(H_i.\Pi_i(c)) \neq NULL$)
            $H_i\Pi_i(c).Count$--;
            if($H_i\Pi_i(c).Count$==0)
               HashDelete($H_i\Pi_i(c)$);
               $C_e$;Derive new $C_e$
end
```

3.2. Dynamic Mining Algorithm

The single layer mining is to mine the data of pre-processing log of database without considering domain knowledge, it is only to seek out the association rule which is greater than certain threshold via the angle of data mining, it is most difficult and most important to seek out high frequent item set, therefore single layer mining is to seek out high frequent item set in the database, next, two conceptions of item set and association rule shall be alternatively used.

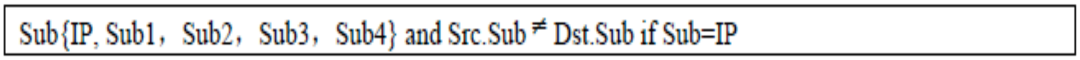


Src.IP, Dst.IP,
Src.IP, Dst. Port,
Src.IP, Dst.IP, Dst.Port,
Src.IP, Src.Port, Dst.IP,
Src.IP, Src.Port, Dst.Port,
Src.IP, Src.Port, Dst.IP, Dst.Port,

3.3. Single Layer Mining

The R (Ts, Src.IP, Src.Port, Dst.IP, Dst.Port, FLAG) mode can be used to create higher level abstract of IP related attribute. Usually for B type address, the first byte of IP address host domain shall determine the subnet of the host, the first two bytes of IP address shall determine the network number of the host.

Src.Sub, Dst.Sub
Src.Sub, Dst.Sub, Dst.Port
Src.Sub, Dst.Port
Src.Sub, Src.Port, Dst.Sub
Src.Sub, Src.Port, Dst.Sub, Dst.Port
Src.Sub, Src.Port, Dst.Port



Sub{IP, Sub1，Sub2，Sub3，Sub4} and Src.Sub $\neq$ Dst.Sub if Sub=IP

3.4. Domain Layer Mining

Table **1** is the detected attack quantity (PROBE. DOS, PASSWD and DIC), false report number, missing report number, when different support degrees are used in dynamic mining algorithm, it shall be used as rule quantity information inputted for decision tree.

Above test result shows that, decision tree plays an important role for reducing false report rate. Since attack type of test data once occurred in training data, decision tree after training shall have the knowledge of such attack type, and it is found that the decision tree misses four attack behaviors that should be detected, it is mainly due to overfitting determined by own limitation of decision tree. Naive Bayes: Naive [9, 10].

Table 1. DARPA1998 test data result.

Week		1st week					2nd week					Total
Workday		Mon	Tue	Wed	Thu	Fri	Mon	Tue	Wed	Thu	Fri	
#rules By MN	Sp=0.0 1	1045	1438	1353	964	1060	1044	1352	1098	1224	991	11569
	Sp=0.0 5	579	666	713	713	514	600	682	622	622	553	6264
	Sp=0.1	477	552	553	456	445	539	524	548	537	488	5119
	Sp=0.2	424	520	483	441	440	522	482	527	500	461	4800
#rules By DT	Sp=0.0 1	7	11	5	60	17	81	16	7	18	17	239
	Sp=0.0 5	7	11	5	46	17	81	13	7	17	16	220
	Sp=0.1	7	11	5	46	12	80	10	7	17	13	208
	Sp=0.2	7	10	5	46	10	80	9	6	14	12	199
#fp	Sp=0.0 1	0	0	0	5	0	0	0	0	0	0	5
	Sp=0.0 5	0	0	0	5	0	0	0	0	0	0	5
	Sp=0.1	0	0	0	5	0	0	0	0	0	0	5
	Sp=0.2	0	0	0	5	0	0	0	0	0	0	5
#det	Sp=0.0 1	4	5	4	7	7	10	8	5	10	11	71
	Sp=0.0 5	4	5	4	7	7	10	7	5	10	11	70
	Sp=0.1	4	5	4	7	7	10	6	5	10	11	69
	Sp=0.2	4	4	4	7	6	10	3	5	9	10	62
#fn	Sp=0.0 1	2	0	0	0	0	1	1	3	1	0	8
	Sp=0.0 5	2	0	0	0	0	1	2	3	1	0	9
	Sp=0.1	2	0	0	0	0	1	2	3	1	0	8
	Sp=0.2	2	1	0	0	0	1	2	3	1	1	10

Table 2. DARPA1999 test data result.

Week		1st week					2nd week					Total
Workday		Mon	Tue	Wed	Thu	Fri	Mon	Tue	Wed	Thu	Fri	
#rules By MN	Sp=0.0 1	1320	1137	1452	1416	2003	1615	1596	1549	1511	1346	14945
	Sp=0.0 5	866	750	1345	1263	1756	1245	1446	1375	1425	1156	12627
	Sp=0.1	422	415	664	641	666	533	634	613	625	632	5845
	Sp=0.2	333	368	595	595	617	388	530	600	525	496	5047
#rules By DT	Sp=0.0 1	21	4	15	11	10	39	17	7	19	10	153
	Sp=0.0 5	15	4	10	4	10	38	17	6	18	9	131
	Sp=0.1	9	4	6	3	5	37	16	5	13	8	106
	Sp=0.2	6	4	4	3	5	36	15	5	10	7	95
#fp	Sp=0.0 1	0	0	0	2	1	2	1	1	2	1	10
	Sp=0.0	0	0	0	2	1	2	1	1	2	1	10

Table 3. DARPA attack data information.

Year	Training data		Test data	
	Attack type	All attack quantity	Attack type	All attack quantity
DARPA1998	11	114	24	65
DARPA1999	7	18	27	106

Table 4. Training and test configuration scheme.

Data set	Training data	Test data
Data set 1	DARPA1998	DARPA1998
Data set 2	DARPA1999	DARPA1999
Data set 3	DARPA1998, DARPA1999	DARPA1999

Table 5. Attack cluster.

1	2	3	4
Neptune	Satan	Nmap	Smurf
Back	Portsweep	Ipsweep	Mscan
Dict.guest	Saint		
Pod.teardrop			
Mailbomb			
Udpstorm			
Snmpguess			
Snmpgetattack			

Table 6. Attack data distribution in DARPA98 training.

Cluster	1						2		3		4
Name	Netptune	Back	Dic t	Guest	Po d	Teardro p	Sata n	Portsweep	Nam p	Ipsweep	smurf
Attack	10	5	1	1	12	9	5	16	2	14	13

Table 7. Attack data distribution of DARPA98 test.

Cluster	1								
Name	Netptun e	Back	Dic t	Guest	Po d	Teardro p	Mailbo x	Udpstro m	Land
Attack	8	3	8	1	6	5	1	3	2

Cluster	**1**		**2**			**3**		**4**	
Name	snmpgue ss	snmpget	satan	Porswe ep	saint	nma p	ipsweep	smurf	Mscan
Attack	2	6	2	5	2	4	3	8	1

Bayes classifier bases one of following simplehypothesis: mutualconditi (on) on of each attribute value is independent. Bayes theorem is used to predict a possibility that an unknown sample belongs to each class, the class with biggest possibility is selected as final class of the sample. Detailed given instance E= XX ......X, herein E has m12 m attributes, Naive B(,)ayes is to classify E into P(C|E) with highest condition probability of K CK. According to independency hypothesis:

$$P\left(C_k|E\right)=\frac{P\left(E|C_k\right)}{P\left(E\right)} \\ =\frac{P\left(X_1|C_k\right)P\left(X_2|C_k\right)\mathrm{K}\ P\left(X_n|C_k\right)P\left(C_k\right)}{P\left(E\right)} \qquad (1)$$

For dispersed variable, $P\left(X|C_k\right)$ P(X|CK) is estimated as I per sample frequency in training data.

Table **2** is the attack type and attack instance information of DARPA data set. It is apparent that DARPA1998 training data has more attack types and attack instances than DARPA1999 training data, furthermore, difference of attack type and attack instance of DARPA1999 is far greater than that of DARPA1998 data set.

For detection of unknown type attack, the biggest advantage of anomaly detection method is to detect new attack type data, nevertheless at present, the performance of intrusion monitoring system for detecting new attack type is not ideal.

Tables **6** and **7** test results give the distribution situation of every cluster attack data of DARPA1998 training data and test data.

In this test, the false code of NAIDS system to core algorithm is given, a large number of tests on DARPA 1998. 1999 intrusion detection evaluation data set demonstrate the effectiveness and science of the method of this paper. In order to enable the system process massive network auditing data, slide window technology is firstly proposed in mining module.

4. GENERALIZATION

NAIDS is the first anomaly detection system based on data mining technology, association rule mining algorithm is adopted to establish system and user normal behavior profile archive under network environment, and seek out suspected mode of network traffic volume data, slide window technology based dynamic mining algorithm is proposed to realize online incremental mining, to in some extent ensure real time performance of NAIDS system, aiming at anomaly behavior inconsistent with normal behavior profile archive, the system can provide the capability of judging anomaly source by tracing back to auditing data, it can be in favor of system safety staffs' participation, normal and anomaly deviation of learning behavior profile archive can adopt monitoring type classification engine. The test data proves that it can greatly reduce false report rate of system, in order to enable NAIDS really have anomaly detection, decision tree is improved to seek out new attack type, absolute support degree conception is proposed in combination with domain issue, the test proves that it can greatly improve system performance.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the soft science study plan project of Ningbo (2011A1014).

REFERENCES

[1] R. Agrwaal and R.Sriknat, "Fast algorithm for mining association rules", in Proe. 20Int. Conf. Verye Date Large DatBaases, *Snatigao Chile,* pp. 487-499, Sep. 1994.

[2] J. R. Quinlna, "Induction of decision trees", *Machine Lemaing,* vol. 1, pp. 81-106, 1986.

[3] W. W. Cohen, "Fast effective rule induction", in Machine Lemaing: Pore of the12 international Conference, Lake Tahoe, Califonria,1995.

[4] R. Ng and J. Hna, "Efficient and effective setring method for spatial data mining", in Proe. 20Int. Conf Verye Large DatBaases, Snatigao Chile, Sep. 1994, pp. 144-155.

[5] T. Zhang, R. Rmakarishnna, and M. L. .Birch, "An effieient data clustering method for veyr large databases", in Proe.ACM-SIGMODInt.Conf Mnagaement of Data, Montreal, Canada, June 1996, pp. 103-144.

[6] R. Agarwal and R. Sriknat, "Mining Sequeniial Pattenr", in Proe. Int. Conf. Data Engineering, TaiPei Taiwan, May 1995, pp. 3-14.

[7] W. Lee and S. Stolfo, "Data mining approaches for intursion detection", in San Antonio. *Tx: Proe.7 USENIX Security Symposium*, 1998, pp. 79-94.

[8] E. Eskin, A. Arnold, and M. Prerau, "A Geometric framework for anomaly detection: detecting intrusions in unabled data", in Data mining Application, Kluwer, 2002.

[9] H. J. George and P. Langley, “Estimating continuous distributions in Bayesian classifiers”, in proceedings of Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995.

[10] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss”, *Machine Learning,* vol. 29, pp. 103-130, 1997.

Received: July 02, 2014 Revised: September 10, 2014 Accepted: November 15, 2014