Realization of Intrusion Detection System based on the Improved Data Mining Technology

ZHAO Yan jun
College of Science
Hebei United University
Tangshan,China
zhaoyanjun@heuu.edu.cn

WEI Ming jun
College of Information Engineering
Hebei United University
Tangshan,China
weimj@heuu.edu.cn

WANG Jing
Tangshan College
Tangshan, China
WJ-0315@163.com

***Abstract*—On the basis of further analyzing the operational mechanism of the existing intrusion detection system model, in allusion to the existing problem the powerless 、high false negative rate、low detection efficiency and the lack of the rule base automatic extension mechanism to unknown aggressive behavior for existing detection mechanisms, Combining the relevant knowledge of data mining technology, then to design one improved network intrusion detection system model based on data mining, combined misuse detection and anomaly detection. In the model, we select the K-means algorithm in clustering analysis and the Apriori algorithm in association rule mining and improve it. Applying the improved K-means algorithm to achieve normal behavior classes and data separation module, then utilizing the improved Apriori algorithm to achieve automatic extension of the rule base. Finally, by the experiment to verify the function of the two algorithms.**

***Index Terms*—data mining, intrusion detection, improved, K-means algorithm, Apriori algorithm.**

I. INTRODUCTION

With the growing importance of network security issues in people's lives, as the new generation of network security intrusion detection technology in the network security also plays one more and more important role. At the same time, the problem in the existing intrusion detection systems has become increasingly prominent. The most existing intrusion detection adopt the method of pattern matching to detect attacks, the data packet need to be detected match with the data in the rule base. The high detection rate for the known attack behavior and the low false alarm rate, But for the variant of the known attack or the unknown attacks can not be detected. Moreover, the pattern matching should establish the detection rule and pattern base of the known attack; the security experts need update and maintenance continuously for the rule base. Otherwise, this will leads to the high rate to the false negative rate. Therefore, improving the detection rate for the existing intrusion detection system and reducing the false negative rate will have a very important practical significance.

According to using different methods to test, Traditional intrusion detection model can be divided into two types; they are intrusion detection model based on misuse and anomaly-based intrusion detection model[1-2]. Misuse-based intrusion detection model needs to establish a rule that base on known attacks, then need to update constantly the knowledge base, to track the development of the attack technology, and detect timely new attack intrusion. Therefore, the effect on misuse detection model detection good or bad largely depends on updates timely of the pattern library. Thus, as can be seen from above, the misuse detection attack only protects the one that has been found, and it does not have the ability to perceive the unknown attacks. Moreover, the anomaly-based intrusion detection model based on the statistical data, to establish a pattern library for normal behavior, once it found a certain behavior that beyond the scope of normal behavior then will be treated as invasion and react accordingly[3-5]. The benefit of such detection model is that has a natural good perception for the unknown attacks. However, the activity behavior of system is changing. Therefore, the normal behavior patterns library need to constantly adjust and difficult to calculate. Comparing these two detection models can be found the conclusion, the anomalous model difficult to perform quantitative analysis and hard to achieve, but the misuse model in accordance with pre-defined rules, to do pattern matching between the detected data and the data in the rule base is relatively simple to realize. Hence, currently, most of the intrusion detection system is based on misuse intrusion detection model, and based on the anomaly detection intrusion detection system and a combination of both is still relatively little, and the more are still in the research stage.

In order to improve the performance of the existing intrusion detection systems, this paper will combine the misuse detection and anomaly detection to construct a hybrid intrusion detection model based on misuse detection and anomaly detection. The most important feature of data mining is that, it can find the knowledge and laws the people unknown from complicated data, and analytical process has the advantage of automation and speediness. This model adopt the data mining techniques[6] to achieve the improvement of the intrusion detection system raised above, Utilizing the clustering analysis algorithm of data mining for network data to conduct processing and establishing the normal behavior class, so as to exclude the most normal data. And utilize the association rules algorithm to realize the automatic extension mechanism of the rule set. Finally, by using the experimental data for the two improved algorithms to verify function.

II. BUILDING THE NETWORK INTRUSION DETECTION SYSTEM MODEL BASED ON IMPROVED DATA MINING TECHNOLOGY

A. *System components*

To analyze the composition of the data from the network, the network data consists of three parts: the normal behavior, aggressive behavior and unknown behavior. And in the three data, the normal behavior data occupies most of the data in the entire network[6]. If adopt only the misuse detection model to conduct pattern matching, the intrusion detection system will take a lot of time in the detection of the normal behavior data and will inevitably result in the waste of system resources. In addition, with accelerating speed of the network, the system needs to handle the network data will greatly increase. When the system capacity is exceeded, the system will lose a portion of the data packet, and may not detect the attacks, then result in a system with the problems of low detection and high omission rate. In order to improve this situation, we consider shunting the network data before performing pattern matching. First through the certain kind of processing to filter off most normal behavior data, and then pass the remaining data to function module of performing pattern matching, in this way can improve the performance of the system to a large extent. At the same time, saving the data that pattern matching module not success to discover the new rules.

The most important feature of the data mining is that it can be found the knowledge and laws from complicated data that people unknown, and the advantages of analytical process is of automation and rapid. Therefore, data mining techniques can be used to achieve the improvements of intrusion detection above.

Cluster analysis algorithm in data mining[7-11] in accordance with certain rules which divide the data into different classes and each class has one cluster center vector. Therefore, taking advantage of the algorithm can process the network data, to establish the normal behavior class, so as to exclude most of the normal data.

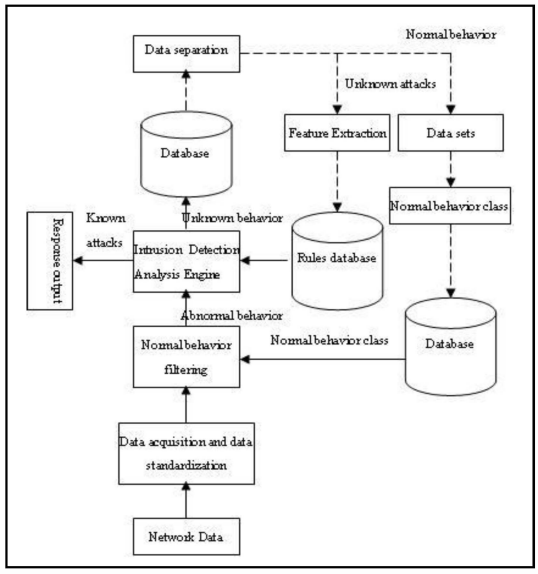

Figure 1. Intrusion detection system module based on data mining

Association rules algorithm in data mining from large amounts of the data to find the links that exist among the various attributes, digging out the relevant rules and using these rules to build a relevant rule base, then to provide a basis for intrusion detection. Therefore, we can use this method to achieve automatic extension mechanism of the rule set.

Based on the above design ideas, this paper design a network intrusion detection system model based on data mining, as shown in Fig. 1:

In Fig. 1, the solid line represents the real-time portion, and the dotted line represents the non-real-time portion.

The basic functions of the various modules as follows:

1) *The acquisition and standardization of data*

This module is mainly responsible for collecting network data and standardizing data, to make preparations for attack detection.

2) *The normal behavior class*

This module is based on k-means clustering mining algorithm to improve the new algorithm formed to clustering analyze the normal behavior data, then to establish the normal behavior class and filter modules provide the basis for normal behavior.

3) *Filtering the normal behavior*

This module is use the normal behavior class is generated by the cluster analysis module to distinguish between normal behavior and abnormal behavior. If it is the normal behavior, then discarded; if it is abnormal behavior, and then pass it to the intrusion detection analysis engine for further judgment.

4) *The intrusion detection analysis engine*

This module make further analysis for the abnormal behavior from normal behavior filtered module, and mainly take pattern matching method to match with the known attack in rule database, if it is the attack then response output; if not, it is unknown behavior then save in database for further processing.

5) *The data separation*

Because the unknown behavioral data include unknown attacks and normal behavior, the main function module is to take advantage the improved k-means algorithm, to make the

normal data and unknown attacks separated, to make preparation for feature extraction and normal behavior class module.

6) *The rule generation*

This module make feature extraction for unknown attack data stored in the database, to establish new attack rule, and to add to the rules in the database automatically, to realize updated automatically of the rules database. The association rule mining algorithm in data mining can dig out the relationship among the different attributes in the record, and it is suitable for the features extraction, this paper intends to adopt the association rule mining Apriori algorithm to achieve the module.

7) *The rules database*

The rules database for storing the rules that the intrusion detection analysis engine to make pattern match.

8) *The data set*

The beginning is initial data sets KDD CUP1999.

B. System workflow

The above running of the intrusion detection system model based on data mining can be divided into the following steps:

1) To establish the normal behavior class

- Using intrusion detection system based on misuse, such as snort to collect the network normal behavior data as the pre-training data.
- Using the cluster analysis algorithm in data mining for processing the collected data, clustering into a few categories, forming the normal behavior class, then is stored in the database.

2) The intrusion detection

- Using the network sniffer to collect network data packet.
- Decoding the data packet, and the data field is stored in which the corresponding data structure.
- Standardizing the data to prepare for filtering the normal behavior.
- Comparing the data with the normal behavior class, if it belongs to certain type and indicates that is the normal data ,then lost it; If not, it is forwarded to the detection engine pattern matching to do the further analysis.
- If the pattern matching is successful so it is the attack, then the corresponding module may take the settings measures; If not, then the data may be included a new attack, then the data stored as the data set of generating the new rules.

3) Adding the new rules

Using the cluster analysis algorithm to cluster the stored data and rule out a small part of the normal data. Then, using association rules algorithm in data mining to make associated analysis to find the new rule and add it to the rule base.

C. The designation of main module

1) The designation of normal behavior class module

The role of this module is to collect data as training data, through adopt improved clustering analysis algorithm K-means and group the training data into several categories, and then extract the characteristics of the normal behavior of formation as a sign of class， and then form normal behavior class model, and take it as the basis for normal behavior filter module.

2) The designation of normal behavior filter module

This module is preprocessor as network data into the detection analysis engine. It uses the normal behavior class mode that generated by normal behavior class module ， to filter the data packet that is about to enter detection analysis engine ,then filter out the data which consistent with the normal behavior, which can greatly reduce the workload of the analysis engine.

3) The designation of data separation module

This module processing data that analysis engine has processed. This part contains unknown attack data and normal data. The data of the two parts should be separated. The rule generation module will be available to unknown attack data. Normal data saved for updating the normal behavior class. The algorithm is still the improved K-means algorithm.

As the substantial between the functionality of this module and the normal behavior class module is the same. Only the processing of data sources is different, the process and the process will not repeat. We just need change the data to logging and normal data is saved to the data set. The attack data is transmitted to the rule generation module.

4) The designation of rules generation module

This module adopts an improved Apriori algorithm ,which associated rule mining for unknown attack data from the data separation module, then represent the found unknown intrusion mode as rules, finally save it to a rule base.

III. THE IMPROVED K-MEANS ALGORITHM

A. The introduction of the K-means algorithm

K-means algorithm[7-11] is one hard clustering algorithm, and the representative of typical clustering method based on clustering criterion function. At the same time, adopting the method of the function get the extremum to get the adjusted rules of conducting the iterative computation. The ultimate goal of the algorithm is based on the input of the number of clustering k and the data objects are divided into k classes.

B. The shortage of K-means algorithm

The thought of K-means algorithm is simple, the computational complexity is small, it can meet the real-time requirement for intrusion detection and relatively simple to achieve. However, the algorithm itself there is some urgent problems to solve:

- It is not able to independently determine the number of clustering and need to input determined value of k in advance. Before the K-means algorithm to clustering, it must be determined the number of clustering k in advance, meanwhile, selecting the same number of data object as initial clustering center randomly. This will result in the division of class is not very accurate, and it is also very difficult to determine the number of clustering independently. Sometimes, it requires the

combination of the priori knowledge of related field as a reference. At the same time, the process of network intrusion detection is timely. So we may not have one way to know the number of clustering k in advance, this will not be able to select k data objects as the initial clustering center.

- The class by K-means algorithm to clustering can not determine which class is normal and which is the exception. But, it needs to pass the normal behavior class to exclude normal packet during the detection process, to reduce the workload of detection engine.

C. *To improve K-means algorithm*

In allusion to the shortcomings of K-means algorithm to make some improvements. We introduce the clustering guide function $f(x_i)$ for the improved K-means algorithm. This function can help the clustering toward direction of the high point density.

To define the Manhattan distance.

$$d(x_i,x_j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{is}-x_{js}| \quad (1)$$

In the formula, $x_i=(x_{i1},x_{i2},\cdots,x_{is})$ and $y_j=(y_{j1},y_{j2},\cdots,y_{js})$ are the s-dimensional data objects.

To define the calculation of r in clustering guide function.

$$r=\frac{1}{m}*\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}d(x_i,y_j) \quad (2)$$

Thereinto, m is the positive number and can be adjusted, the usual circumstance to 2.

To define the clustering guide function.

$$f(x_i)=\{p \mid Dist(p,x_i)\prec r,\ p\in X, x_i\in X\} \quad (3)$$

In the formula, X is the data object set, $Dist(p,x_i)$ is the distance from data object p to data object x_i, r represents the distance radius, then using the Manhattan distance to calculate.

Each object in the dataset can be calculated according to the calculation method above to get the corresponding clustering guide function value, the value will be used as the basis for selecting the clustering object.

The improved K-means algorithm to find the object of possessing the biggest clustering guide function value within a small range, that is the object of the highest point density as the representatives of the class, Each class will be composed of the object of representative point and the object belonging to the representative point.

The entire algorithm of the improved K-means algorithm is described as follows:

Input: $X=\{x_1,x_2,\cdots,x_n\}$ is the data set that contain the number of object is n.

Output: k classes

The procedure:

According to the data object input to calculate r and the clustering guide function of every object. Regarding every object as one class and forming the n classes, only one representative point in each class.

Do

For the representative point x_i in each class to find the object x_j, $i\neq j$, the distance from x_j to x_i less then r, and $f(x_j)$ is the biggest of the object less then r. Comparing $f(x_i)$ and $f(x_j)$,if $f(x_i)\prec f(x_j)$, then x_j is classified into class x_i.Otherwise, the class x_i invariant.

Until for every $i=1,2,\cdots,n$, the class x_i will no longer change.

IV. THE IMPROVED APRIORI ALGORITHM

A. *The introduction of the Apriori algorithm*

Apriori is a classic association rules algorithm proposed by R.Agrawal in data mining. For the database, the Apriori take multiple scan to generate frequent itemset[12]. The particular way is: after the first scan only to generate candidate itemsets with the width is 1, then comparing the support degree and the minimum support degree of each item, then adding the 1-order candidate itemset which less than the minimum support degree to the frequent itemset, at this moment, only forming the initial frequent itemset contains only 1-order item. Hereafter, every scan must be based on the previous frequent itemsets and construct the candidate itemsets, then scan the database item from the candidate set to determine the frequent itemsets. Repeat the process until no new frequent itemsets to emerge.

B. *The shortage of Apriori algorithm*

Apriori algorithm is one common data mining algorithms, and does not support the multi-attribute problems of intrusion detection. There is no mutual relationship among the item of every affair in Apriori algorithm, the existence of one item will not affect the other item, and any item portfolio can. However, for the data of intrusion detection database is not so. If utilizing the data in this database to conduct the association rule, in the middle of the process will produce the invalid candidate item, and need to scan the database to calculate its support degree, it would greatly reduce the efficiency of the algorithm.

C. *To improve Apriori algorithm*

Based on knowledge of the intrusion detection field, we know the property that the different support degree is 0 in the same feature. For the Apriori algorithm to do the following improvements:

The Apriori algorithm cand-gen ( ) subroutine to be changed is:

Input: frequent itemsets Lk-1

Output: Initial candidate itemsets Ck

```
Begin
for i=1 to Lk-1 the number of itemsets
  for j=1 to Lk-1 the number of itemsets
{
  if li and lj only have one different item, and the two items do not belong to the same feature.
    li∪lj∈Ck;
}
End
```

V. AUTHENTICATION ALGORITHM FUNCTION

A. *The verification experiment of K-means algorithm performance*

The experimental data set: adopting the authoritative test data in the intrusion detection field, "kddcup.data_10.percent" 10% data set in KDD cup99[13].

The experimental data set mainly includes: DoS、U2R、R2L and Probe, the four categories attack data. The following is contraposing the four category attack data to detect the algorithm performance.

In order to meet the requirement of the two hypothetical in detection algorithm, each experiment selects 2000 records from the corresponding data set for the experiment. Thereinto, 1966 normal data and 34 intrusional data. The normal data take the proportion of 98.30% in all data. To meet the need for assumption that the number of normal data much larger than the number of the intrusion data in detection algorithm. Setting each data on the K-means and improved K-means to conduct experiment respectively, then making the comparison of the algorithm performance. The higher detection rate and the lower false detection rate, the better the detection capabilities of the algorithm.

The following experiment will adopt the two different angles for the detection performance of the algorithm. One way is adopt the single attack dataset, that is all the dataset attack data belong to one single category. The other way is to adopt the mixed attack detection, that is the dataset attack data is one mix of various attack data and the category is richer.

1) *The algorithm for the detection performance of the single attack*

K-means algorithm needs to specify the number of clustering in advance, and the different number of clustering has a great impact on the clustering result of the algorithm, so it needs for repeated experiments. By adjusting the numbers of clustering to get the clustering performance of K-means algorithm on various attack data set, as shown in Table 1.

TABLE I. PERFORMANCE OF K-MEANS ALGORITHM ON SINGLE ATTACK

Data set	The number of clustering	The false alarm rate %	The detection rate %
DOS	15	0.864	100
R2L	11	2.339	91.176
Probe	20	6.256	100
U2R	15	4.629	8.823

TABLE II. PERFORMANCE OF IMPROVED K-MEANS ALGORITHM ON SINGLE ATTACK

Data set	The number of clustering	The false alarm rate %	The detection rate %
DOS	35	0.712	100
R2L	35	4.425	100
Probe	50	0.966	94.118
U2R	15	0.560	5.882

TABLE III. PERFORMANCE OF K-MEANS ALGORITHM ON COMPREHENSIVE ATTACK

The test sequence number	The number of clustering	The false alarm rate %	The detection rate %
1	36	1.424	20.588
2	40	0.102	67.647
3	44	2.747	88.235
4	50	5.239	91.177
5	55	6.205	91.177

TABLE IV. PERFORMANCE OF IMPROVED K-MEANS ALGORITHM COMPREHENSIVE ON ATTACK

The test sequence number	The number of clustering	The false alarm rate %	The detection rate %
1	44	1.984	91.177

Modifying K-means algorithm does not require specifying the number of clustering in advance, it is based on the specific independent determine the number of clustering in data set. Final to get the clustering performance of the improved K-means algorithm on the various attack.

From the experimental results to analyze: It can be seen from Table 1 and Table 2, comparing to the K-means algorithm, the single attack performance of the improved the K-means algorithm has the lower false alarm rate and the higher detection rate.

2) *The algorithm for detection performance on comprehensive attack*

The clustering performance of the K-means algorithm is shown in table 3.

The performance of improved algorithm is shown in table 4.

From the experimental results to analyze: It can be seen from Table 3 and Table 4, comparing to the K-means algorithm, the comprehensive attack performance of the improved the K-means algorithm has the lower false alarm rate and the higher detection rate.

In summary, the improved K-means algorithm can find the appropriate number of clustering that according to the data characteristics under the premise of not need enter the number of clustering; it shows the lower false alarm rate and higher detection rate on the single attack and the comprehensive attack. Therefore, the improved K-means algorithm is feasible.

B. *The verification experiment of Apriori algorithm performance*

Utilizing the improved Apriori algorithm to realize the capacity of the rule automatically expands. First, mining the frequent itemset in the data set to form the association rule and stored in the rule base.

Figure 2. Generation of the rules

This test output rule adopts the format of Snort rule, therefore, selecting duration，protocol_type，service，flag，src_bytes and dst_bytes these fields related with the Snort rule.

The program interface of realizing the performance of the rule automatically expands, as shown in Fig.2:

According to the figure, the process of generation rule is: when setting minimum support degree is 0.2, then click on the button of "frequent set to generate", in the below list box will display the frequent set that satisfy minimum support degree of the order is the highest. As shown in the figure, it generates the two frequent set of the support degree is more than 0.2. Then, setting minimum confidence degree is 0.2, then click on the button of "the strong rule to generate", in the below list box will display the rule that satisfy minimum confidence degree. As shown in the figure, the frequent set generated just is the strong rule. Finally, clicking the button of "outputting the Snort rule", The program will write the strong rule just generated according tu the Snort rule format into the rules.txt.

The conclusion: by the rule generation program for the unknown attack data to mining association rules, it can achieve the expansion of the intrusion detection system rule base.

VI. CONCLUSION

This article designs one intrusion detection system model which based on data mining, combining misuse detection with anomaly detection, carried out a detailed introduction for the work processes of associated modules and the work steps. In allusion to the function of the important modules to realize in the model, this paper selects the appropriate algorithm in data mining algorithms. Utilizing the improved K-means algorithm to achieve the normal behavior classes and data separation module, then with the improved Apriori algorithm to achieve automatically extensions of the rule base. Finally, by the experiment to verify the functions of two algorithms.

Therefore, the research of the intrusion detection system model based on improved data mining algorithms has improved the detection rate and reduced false negative rate, to improve the detection performance of all the detection system.

REFERENCES

[1] Wang Yan,Xiao Wei-min. Research on the Application of Data Mining Technique in Intrusion Detection System. Anhui:Anhui University,2010.

[2] Verwoerd T, Hunt R. Intrusion detection techniques and approaches. Computer Communications, 2002, 25(15), 1356-1365.

[3] Yu Kun. Designs and Implementation of Abnormal Intrusion Detection Based on the High Frequency Statistics. Beijing :Beijing University of Posts and Telecommunications,2006.

[4] Cai Jian. The Research and Implementation of Intrusion Detection System Based on Artificial Neural Network. Guizhou :Guizhou University,2005..

[5] Wu Tao,Wang Xin-fang. Research of Intrusion Detection Techniques Based on Data Mining]. Xi'an:Xi'an University of Technology,2010.

[6] Chen Yu-hui,Fu Ming. Study on Intrusion Detection Technology Base on Data Mining. Changsha:ChangshaUniversity of Technology,2010.

[7] LI Yang. K-meansApplication of K-means Clustering Algorithm in Intrusion Detection. Computer Engineering,2007,33(14),154-156.

[8] ZHANG Jian-ping. Application in Cluster’s Analysis Is Analyzed in Children Development Period. Application Research of Computers,2007,24(5),166-168.

[9] Chinrungrueng.C, Sequin, C.H. Optimal adaptive k-means algorithm with dynamic adjustment of learning rate, Columbia University, 2002.

[10] Yan Xiao-guang. Application to Cluster Algorithm in Anomaly Detection of Network Intrusion]. Applications of The Computer Systems,2005,10,34-37.

[11] MA Xiao-chun, GAO Xiang, GAO De-yuan. The Application of Cluster Analysis to Intrusion Detection System. Microelectronics & Computer,2005,22(4),134-136.

[12] Ying Sun. Using Data Mining Technology Solve Intrusion Detection of Network. Computer Knowledge and Technology, 2010, 6463-6464.

[13] Zhang Xinyou,Jia Lei. The Research of Data Mining Datasets KDD CUP99.Computer Engineering and Design,2010,31(22):4809-4812.