**Yunfeng DONG[1], Bei QI[1], Wieyue ZHU[1], Wushi GAO[1]**

Shandong Polytechnic University (1)

# A New Intrusion Detection Model Based on Data Mining and Neural Network

***Abstract***. *Today, we often apply the intrusion detection to aid the firewall to maintain the network security. But now network intrusion detection have problem of higher false alarm rate, we apply the data warehouse and the data mining in intrusion detection and the technology of network traffic monitoring and analysis. After network data is processed by data mining, we will get the certain data and the uncertain data. Then we process the data by the BP neural network, which based on the genetic algorithm, again. Finally, we propose a new model of intrusion detection based on the data warehouse, the data mining and the BP neural network. The experimental result indicates this model can find effectively many kinds behavior of network intrusion and have higher intelligence and environment accommodation.*

***Streszczenie:*** *Obecnie, w celu utrzymania bezpieczeństwa sieci, stosuje się wykrywanie ataków przy pomocy zapory ogniowej, co często powoduje za wysoki poziom fałszywych ataków. W proponowanym rozwiązaniu proponuje się wykorzystanie magazynowania i pozyskiwania danych oraz analizę monitoringu ruchu sieci. Przetwarzanie danych polegało dotychczas na ustaleniu danych pewnych i niepewnych; obecnie proponujemy wykorzystanie genetycznego algorytmu sieci neuronowych BP. Ostatecznie, wprowadzono nowy model detekcji ataków bazujący na magazynowaniu i pozyskiwaniu danych oraz neuronowych sieciach BP. Badania eksperymentalne wykazują, że zaprezentowany model pozwala na znalezienie wielu rodzajów zachowań ataków sieci, jest bardziej inteligentny, zapewnia wyższy standard obsługi środowiska.* ***Nowy model detekcji ataków oparty o pozyskiwanie danych i sieć neuronową*** .

**Keywords:** Intrusion Detection; Data Mining; BP Neural Network; Genetic Algorithm.
**Słowa kluczowe:** Detekcja ataków, Pozyskiwanie danych, Sieć neuronowa BP, Algorytm genetyczny

## Introduction

In the last few years, the specifications of network are growing rapidly, and the security of network is threaten from double layers-inside and outside network. To ensure the normal operation of the network, we adopt the technology of firewall to defense against various network attacks, but even the best firewall will become "Maginot line of defense" when it is bypassed. Therefore, most of the network apply the intrusion detection technology to aid the firewall. Based on researching into a large number of network intrusion detection technology and the features that network's intrusion data is a small fraction of the all data transmitted by network and that intrusion data and normal data can be distinguished, we establish intelligent intrusion detection model based on data warehouse and data mining, combining with network traffic monitoring and abnormal traffic analysis techniques.

## The overall structure of the intrusion detection model

The model mainly consists of four modules: data acquisition and classification module, uncertain data's abnormal traffic anomaly detection module, uncertain data's abnormal traffic analysis module and the neural network decision module, the overall framework shown in figure 1. In data collection and classification module, we use the clustering algorithm to classify network data into certain data and uncertain data, and use the certain data to train the fourth module of neural networks, and uncertain data to analyze network traffic; In uncertain data's abnormal traffic anomaly detection module, we detect network packet traffic and start abnormal data analysis module when finding the abnormal traffic; abnormal traffic analysis module uses data mining algorithms to get the abnormal network traffic's frequent sequence, establishes simplified and efficient sequence rules, judges the uncertain data's traffic anomaly, and provides traffic data for reference to the neural network decision module; the neural network decision module learns by the certain data, judges uncertain data according to traffic data and the rules that has learned, establishes an intelligent intrusion detection model.

## Improved clustering algorithm analyzing network data

Now more and more researches apply data mining techniques to intrusion detection, and various data mining algorithms, such as association rules mining, frequent episodes rules mining, classification algorithm, clustering algorithm [2][3] have been raised. In our model, we have chosen clustering algorithms. Clustering algorithm is a "non-supervised learning algorithm", which frees itself from the training data, is very suitable for the first part of our model — the network data classification. The clustering algorithm we used is an improved clustering algorithm, combining K-MEANS algorithm

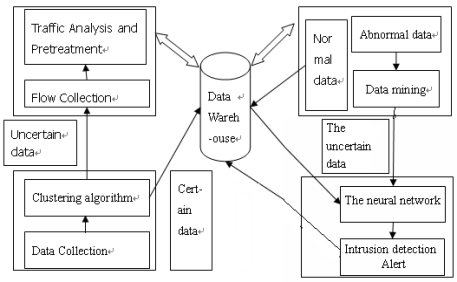


Fig.1. Frame of the model

K-MEANS algorithm accepts input k, then divides N of data objects into k of cluster in order to make the clustering obtained: the objects of the same cluster are of high similarity; but the objects of different cluster are of smaller similarity. The working process is explained as follows: first, to choose k of objects as initial cluster centers random from the n of data objects; for the remaining others, to assign them to their most similar (represented by cluster center) clustering according to their similarity with these cluster centers (distance); second, to calculate each cluster centers of the new cluster; to repeat this process until the criteria measure function begin convergence. Mean square error function is generally used as a standard measure function.

K of cluster has the following characteristics: the cluster itself as compact as possible, but as separated as possible between each cluster.

K-MEANS algorithm flow is as follows:

K of K-MEANS algorithm is the number of clusters that average value of objects input.

(0) inputting the number of clusters to k, and then a large value is assigned to MIN;

(1) all of objects' Data measure value is standardized;

(2) choosing K object random as initial cluster centers;

(3) repeat；//3-6

(4) calculating the Euclidean distance of remaining each object to cluster center;

(5) finding the shortest distance of cluster center, and distributing the object to the cluste whose core is this cluster center according to the distance of cluster centers object,

(6) until the distribution of objects doesn't change

(7) calculating the criterion function of current cluster's situation $E = \sum_{i=1}^{k} \sum_{x \in c_i} \left| x - m_i \right|^2$, if it's convergence value >MIN, then MIN=VALUE；

(8) switching to (2) and continuing, until finding the value of the least MIN by the function $E = \sum_{i=1}^{k} \sum_{x \in c_i} \left| x - m_i \right|^2$.

At last the K of cluster is result when convergence value is the minimum MIN. After the calculation of the above algorithm, the set D including Cl, C2, …, Ci of clusters, which i=l…k. and the objects collection of each cluster C1={obi,…}, …, Ci={obm…}; Which i, …, m∈1～N are got. The two conditions—the radius E and the minimum MinPts needed in DBSCAN are calculated. With the two data, DBSCAN (E, MinPts) algorithm is used to analyze abnormal record collection in the k of clusters again to get more precise results.

The process of calculating the radius E is: first calculating the average distances of every number of objects in each set of D, more than two clusters, then summing these distances. The distances' sum divided by the number of clusters in set D is the radius E. In this process, the cluster average distance is the quotient that the sum of objects' Euclidean distance is divided by the number of combinations in clusters' objects. MinPts is the proportion that the total number of objects whose adjacent objects distances is smaller than the radius E in cluster acounts for total number of all objects in collection D.

**Anomaly network traffic analysis**

First: Flow Collection

The main work of collecting network traffic and pretreatment is to collect abnormal traffic data, do traffic statistics according to port, record information of port traffic, and store extracted information into database. Based on the data warehouse, we build a database of network traffic analysis, which is a secondary database. First-level database is used to receive the raw data, and used when the very detailed and specific analysis of network information traffic is needed. Second-level database is used to store aggregate information of first-level database, providing raw data for network management, network feedback, network control, network planning, etc.

First-level database contains the following attributes of network traffic:

(1) Header fields of packets' IP layer, transport layer, application layer or link layer, such as: purpose of IP address, destination port numbers, ToS field, and so on.

(2) One or more of packets' statistical properties, such as: MPLS, label quantity, and so on

(3) One or more fields coming from network switching and routing equipment due to packet processing, such as: hop IP address, output interface, and so on.

After firsr-level database collecting raw data of network traffic, second-level database extracts data from first-level database, then cleans and converses, to ensure uniform data format.

Finally, loading data, and aggregating data of first-level database according to minutes, hours and days to get types of the database table structure of second-level database and network traffic data：

Table type1: Aggregating data according to sources IP address, source port number, destination IP address, destination port number and the time to form the second-level database tables according to three different size of minutes, hours and days.

Table type2: Aggregating data according to upper layer protocol identifier, protocol and time to form the second-level database tables according to three different sizes of minutes, hours and days.

Table type3: Aggregating data according to the source IP address, destination IP address, upper layer protocol identifier, protocol and time to form the second-level database tables according to three different size of minutes, hours and days.

In the process of network running, when analyzing network traffic according to one or more of the features of the source IP, the destination IP, source port, destination port, we can directly analyze data by the time size in the table type1; when analyzing network traffic according to the protocol analysis of network traffic characteristics, we can directly analyze data by the time size in the table type2; when analyzing network traffic according to one or more of the features of the source IP, the destination IP and protocol, we can directly analyze data by the time size in the table type3.

Second: Frequent Pattern Mining Algorithm of Network Traffic

According to data collected and dealed with previously and the data warehouse, we use Frequent pattern mining algorithm—CLOSET [1] to check uncertain data in network again. System collects information of network traffic in every port from the data warehouse regularly, and calculates the frequent pattern regularly. In order to adapt features of the traffic treatment, we add time marker in the acquired frequent patterns, record generation time of frequent patterns, reflect the existence time and priority of frequent pattern by time marker of frequent pattern, mine out frequent sequence of the abnormal network traffic, get uncertain data sequence of abnormal network traffic, establish efficient sequence of rules, and get the result of intrusion detection. Then the result is trained to detect by the artificial neural network.

CLOSET algorithm：

Input： database, Minimum support support threshold min_support；

Output： Frequent closed item FCI；

method：

BEGIN

(1) Build flist：

Scan the database to extract information;

Calculate the number of same items in item (port,stream) ;

Flist set empty;

Every network traffic in database {

If (item.support>=min_support)

add item to Frequent item list;

}

(2) Flist is descending order according to the number of support: Descend_flist_item (flist) ;

(3) Get Frequent item list in database according to flist and order: transaction_item_descend () ;

(4) FCI=φ;// initializeFCI

(5) Use CLOSET Algorithm, get FCI;

(6) CLOSETφ;

END
Flist is arraytype, every item's data type;
Type fist_node as
{port
traffic
support
point}

**The BP neural network based on the genetic algorithm**

The neural network is selected by its self-learning ability, which can still get a full feature profile; Even the system data is incomplete and variable. By clustering algorithm and flow analysis algorithm, we have already divided determined data set into different clusters by size, and extracted out traffic anomaly of uncertain data, that is, we believe that the invasion of data. According to cluster size, we will sort sub-cluster descending, and make the large cluster in the data as normal data as training data for training neural networks. Meanwhile, we can use detected abnormal data for training, and finally neural network which had been trained is used to distinguish the smaller cluster.

However, the defect of BP algorithm is slow convergence and easy local convergence, and it can be solved by genetic algorithm can solve exactly the problem. The excellent individual characteristics of network intrusion data have been well preserved by data mining and traffic analysis, which niche genetic algorithm technology required. We combine genetic algorithm with BP algorithm[4], the advantages of genetic algorithms was used to overcome the slow convergence of BP algorithm and easy local convergence of the defect, therefore, we combined genetic algorithm and BP algorithm. First, we used the advantages of genetic algorithm to overcome BP algorithm's shortcoming, slow convergence and easy local convergence. Then, we used gradient information of BP algorithm to solve the problem, which genetic algorithm only can find the neatly optimal solution in short-time.

The idea is to optimize repeatedly the network data which has been processed with the genetic algorithm, until the average error of fitness function is no longer increase in so far significantly. Then BP algorithm optimizes them again. The basic idea of this combination method is to select the network intrusion data with genetic algorithms roughly, then thin and optimize with BP algorithm. The genetic algorithm which the system used is summarized below:

First: The new data stream is called the parent, which extracted from the network characteristics with the invasion. On each parent chromosome using crossover operator with the genetic algorithm to form its next generation of the offspring, also known as the first generation

Second: We will define all of the characteristics adaptation in the feature library as 1.

Third: We will match the parent and the generated offspring various characteristics of each chromosome with corresponding to the various characteristics of each chromosome in the features library;

Fourth: Substitute the signature features of the recent △ values in the fitness function, and work out the value of the fitness f;

Fifth: Select the larger half of the fitness of individual species in the composed of parent and offspring in the population as a new cross-operating to form the second generation of the individual;

Sixth: Repeat 3, 4, 5 step;

Finally we can consider to have local optimization when the fitness greater than a threshold value after several cycles. Variability of these characteristics until the fitness reaches the goal we have identified. At this point we will submit intrusion data extracted and determination of selected data to neural network for learning, and judge the uncertain data.

The BP algorithm is a training of teachers, in the learning algorithm using the mean-squared error and gradient descent method to achieve the amendment of the network connection weights. The amendatory goal is to least the mean-squared error between actual output and specified output of the network. BP network model consists of three layers, all connections between the layers. Because the BP network training speed is slow for large data sets, we use the algorithm with adaptive modifiable learning rate to train the data. In the negative gradient algorithm, which used in BP network training, the learning rate is a fixed constant, and its value will directly affect the training performance. If the value is too large, the network stability will be reduced; If the value is too small, the training time will be too long.

This algorithm calculated the network output error first, then after each training, we used the learning rate, which we get from training, to calculate the network weights and thresholds, and calculated the network output error. If the ratio of output error and last output error is greater than the predefined parameter, then we reduced the learning rate, on the contrary, we increased the learning rate. Then, we calculated the network weights, thresholds and output error again, until the ratio of the output error and its last output error is less than the parameter.

**Experimental Results and conclusion**

Making use of the well-formed campus intrusion detection system brought by working conditions, we construct data set based on collected network data from campus network in one year. The data set has total of 721,068 pieces of data, among which 5,423 pieces of data have been identified as the intrusion data (0.76 %). After clustering, every sub-cluster will be ordered by size from big to small. The certain data set (72% of total amount of data,) will be regarded as normal data; uncertain data will be collected in data warehouse and analyzed anomaly network traffic for the second time. Abnormal data are finally detected. Experiments' results show that accuracy of intrusion detection is 82%; False alarm rate is less than 6%.

The advantage of this model is that it can be self-learning and has good adaptability. This model has high detection rate for conventional intrusion means, and improve the ability to detect unknown intrusion.

REFERENCES


[1] Pei J, Han J, Mao R.CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets [J]. Proc. 2000 ACM-SIGMOD Int. Workshop onData Mining and Knowledge Discovery (DMKD'00) , 2000:132-196.

[2] S d kant. Fast Algorithms for Mining Association Rules and Sequential Patterns[C]. Madison: University of Wisconsin, 2003.24(5):324~355

[3] Qiao X.W, Xin Y, Bin,S, Ge.Anomaly. inrtusion deteetion method basedon HMM. Eleetronies Letters, Vol. 38, No. 13, P. P663—664, 20 Jun 2002

[4] E.Eskin. Anomaly deteetion over noisy data using leanred Probability distributions Proceedings of ICML 2000. Menlo Park, CA, 2002



***Authors**: prof. Yunfeng Dong, Shandong, Polytechnic University, Computer center, ul. Jingshiyi Road, Jinan, Shandong, China, E-mail: dyf@spu.edu.cn; Bei Qi, Shandong, Polytechnic University, Network center, ul. Jingshiyi Road, Jinan, Shandong, China, E-mail: qb@spu.edu.cn; Weiyue Zhu, Shandong, Polytechnic University, Computer center, ul. Jingshiyi Road, Jinan, Shandong, China, E-mail: zwy@spu.edu.cn; Wushi Gao, Shandong, Polytechnic University, Computer center, ul. Jingshiyi Road, Jinan, Shandong, China, E-mail: gws@spu.edu.cn.*