The Application of Data Mining Technology in the Intrusion Detection System

Zongpu Jia¹, Shichao Jin²

¹Computer Science and Technology Department, He Nan Polytechnic University, JiaoZuo, China
E-mail: jiazp@hpu.edu.cn

²Computer Science and Technology Department, He Nan Polytechnic University, JiaoZuo, China
E-mail: swaybottle@sina.com

Abstract—This paper analyzes the current situation of the intrusion detection system, which is the basis to put forward that data mining technology is to be applied to the intrusion detection system in terms of the problems of the traditional intrusion detection system. Meanwhile, the paper designs the intrusion detection model of data mining. With the study on intrusion detection and data mining, the algorithm of classical relation with clustering in the characteristics on intrusion detection system is improved and optimized.

Index Terms—Intrusion Detection; Data Mining; Error Detection (Abnormal Detection); Misuse Detection; Relation Rule; Sequence Rule; Clustering Algorithm

I. FOREWORD

With the development of internet technology, more and more people have got the abundant network resources to learn the various patterns of network attack and may implement the seriously destructive attack only with the simple operation. So how to detect and prevent the invasive behaviors has become the highlight of computer field.

There are plenty of methods to strengthen the network security such as setting secret code, VPN and firewall. But most them are static and can’t make the effective protection. However, the intrusion detection technology is a dynamic protective strategy, which can make monitor, attack and counterattack on network security to make up the weakness of the traditional static strategy.

II. THE INTRODUCTION OF INTRUSION DETECTION TECHNOLOGY

The intrusion detection technology monitors the operation state of network system and digs out each attacking attempt, attacking behavior and attacking result so as to ensure the confidentiality, integrity and usability of system resource. The intrusion detection system can be classified into the diverse patterns of being based on mainframe, being based on network, being based on kernel and being based on application. The paper mainly analyzes the structure of the intrusion detection system based on network. The intrusion detection system can be divided into two sorts according to the differences of the data analysis methods.

The first one is misuse detection, also named detection based on quality, which is to establish a quality-base in terms of a known invasive behavior to match the motion having occurred. When the consistent result is got, the invasive behavior is surely proved. The merit of misuse detection is that it has got the low misinformation rate. However, the quality-base will become larger and larger because of too many invasive behaviors and can only detect the known invasive behaviors.

The second one is anomaly detection, also named detection based on behavior, which is to establish a normal quality-base and judge the invasion according to the user’s behavior or the consumption of resource. The merit of anomaly detection is that it has the strong currency and little relation to system and can detect the attacking methods that have never appeared before. However, it still has the high misinformation rate because of the impossibility of the produced outline to comprehensively describe all users’ behaviors of the whole system as well as the variable behaviors of each user.

Therefore, the combination of the two methods can obtain the better performance. Anomaly detection can make the system detect the new, unknown and other situations; misuse detection can protect the integrity of anomaly detection by means of preventing the alteration of behavior patterns that some patient hackers may use to make anomaly detection consider it legal.

The data origin of intrusion detection can be obtained by some specialized wire-shark, for example, Winpcap is generally used to obtain data packet in the system of Windows, while Tcpdump and Arpwatch may be used in the system of Unix. Data mining technology is mainly introduced and will be used in the data analysis phase. The response consists of the active response and the passive response.

III. THE INDUCTION OF DATA MINING TECHNOLOGY

The mission of invasion analysis is to find the invasion trace among the numerous obtained data. The network invasion is judged when the obtained data is input into the detecting system as information and analyzed and disposed by the detecting system. It is a huge intellectual project to establish a rule-base (quality-base), because the current detecting rules are generally made by handcraft,
especially the detecting knowledge of judging the invasion behavior, which is a series of deduction rules coming from the security experts’ analysis experience on the suspicious behaviors to obtain the invasion quality so as to compile IDS. The IDS of this sort has got the obvious fault that the experts need to summarize and obtain the invasion quality continually so as to provide the comprehensive and abundant invasion detecting rule. It proves that the IDS can passively detect the known invasion or attack behaviors only with the outer aid but can’t detect the variants or unknown invasion or attack behaviors. Besides, the reliability of IDS is not steady because of the man-made analysis and experience and the misinformation rate is high.

In order to overcome the current limitations of IDS, a new sort of technology---data mining technology is supposed to be applied. Data mining is the process of mining the unknown knowledge useful to decision from the numerable data sets. The clustering analysis algorithm of data mining is suitable for constructing the normal behavior model of network from the numerous data packets, while the relation analysis algorithm is suitable for describing the relation rules of invasion behavior pattern with which invasion detecting is carried out.

IV. THE REALIZATION OF INVASION DETECTING SYSTEM BASED ON DATA MINING

The invasion detecting system in the paper consists of the pre-processor of data, data partition, relation rule obtaining, rule base and data analysis, and so on. The flow chart is as the following.

![Figure 1. The Pattern of Invasion Detecting System Based on Data Mining](image)

The pre-processor of data is in charge of selecting the reasonable property from the data stream such as the system log data and network data packets to create the data form fit for detecting pattern; data partition partitions the data into the abnormal or normal data set with clustering algorithm. The module of relation rule obtaining makes the quality obtaining of each data set, then creates the rule sets and the rule bases describing the current each data set to store the normal and abnormal rules respectively. Whether the invasion occurs is decided by calculating the matching degree between the rule of the current data set and the rule of the rule bases. After the rule of the unsuccessful matching is input into the data analysis engine which makes analysis and judgment, the correspondent normal and abnormal rules are formed.

V. THE CRUCIAL TECHNOLOGY ANALYSIS

A. Data Partition

The partition of data is to accomplish the following work. That is to form numerous data sets accurately, to make each data set comprise just the normal data or abnormal data, to judge whether each data set is the normal data set or the abnormal data set and utterly to favor the relation rule obtaining.

The clustering algorithm is a common sort of technology of data mining, the core idea of which is to gather the similar (or close) data into a cluster and make the same patterns of data close but the different patterns of data distant. Here the improved clustering algorithm of K-means and the data partition in advance is taken.

The disadvantages of the traditional algorithm of K-means include:

1) The amount K of the ultimate clustering is supposed to be confirmed in advance and the record of the same amount should be designated as the initial clustering center. Then the whole record set is to be scanned by times, the clustering center and the clustering that the record belongs to are to be altered continually before the steadiness of the clustering center. The result of clustering has a direct and close relation to the amount K of clustering. The different clustering amounts will bring the different clustering results, so it is very hard to confirm the clustering amount bringing the best clustering result.

2) Some of the clustering results are likely to be vacancy, that is, the vacancy clustering may be brought without any object similar to the clustering center of such clustering.

3) The initial clustering center at random is likely to be not the best one at the beginning of the clustering and the clustering result is easily influenced by the initial clustering center. The traditional algorithm of K-means takes the arbitrary selection when the initial clustering center is obtained. However, obtaining a better initial clustering center can obtain a better clustering result.

The schedule to improve of the algorithm of K-means:

The improvement is made according to the disadvantages of the algorithm of K-means, that is, the two clustering parameters of the clustering semi-diameter and the nearest value H are to be added. The concrete method is the following. It is to calculate the smallest value of the distance of the current record from all the clustering centers; the record will be regarded as the new clustering center if the smallest value is more than the clustering semi-diameter. The final clustering amount K will be the best clustering amount of the data set instead of the beforehand clustering amount. It is not necessary for this method to confirm the clustering amount in
advance, which can sort out the records of the contiguous distances into the same clustering and isolate the abnormal records of great distances from the other clustering. The method can adjust the clustering amounts in a certain scope and sort out the abnormal records into an individual clustering, which is favorable for marking the abnormal records and cuts down the influences on calculating the clustering center. When there is vacancy clustering in the clustering results, the farthest object from the clustering center will be removed from the current clustering to bring a new clustering center so as to replace the vacancy clustering with the newborn clustering.

For the selection of the initial clustering center, the sample set of the amount T should be taken first and the initial center set of the amount T will be produced after the clustering of K-means on each sample set. Then all the elements of T×K are clustered with the algorithm of K-means with C1 being the initial clustering center and the clustering center sets of T will be obtained, then the best one will be selected as the ultimate initial clustering center.

The advantage of this method is putting forward a method to automatically select the initial clustering center, which reduces the complexity of algorithm time by means of selecting sample instead of the whole data set and can avoid the influence of “the isolated point” by means of the beforehand initial center clustering and multiple sample sets so as to improve the representation of the initial center.

Improved k-means algorithm is described as follows: Input: a database containing n items of data

Input parameters: the initial number of clusters M; cluster radius r; nearest neighbor threshold h

Output: k a cluster

(1) Select the M were the best initial cluster centers (w1, w2, ..., wn) wj = xi, where, j ∈ {1..k}, i ∈ {1...n};
(2) to correspond to each cluster cj and wi.
(3) Calculate the other records xi (i (1 ... n)) to the minimum distance from cluster center min.
(4) If min < r, will be assigned to the nearest wj * xi where cluster Cj *; that | xi-wj * | <= | xi-wj | mj = (1 .. k) Otherwise, create a new cluster, the xi as a new cluster center.
(5) Back (2), until all records are complete.
(6) to each cluster mean replace the original cluster center, namely:
(7) If there is an empty cluster, the furthest point away from the cluster center out to create a new cluster where the cluster center, the new cluster created to replace the empty cluster.
(8) until the same value until the cluster centers.
(9) Finally, calculate the value of all the cluster centers for any two centers The distance between the nearest neighbor with threshold h are compared, if the distance is less than h is to merge these two cluster.
(10) repeat (9), until the distance between any two cluster centers are greater than the value of h up

B. The Birth of Relation Rule

The normal and abnormal data are partitioned into different data sets to obtain after the clustering partition of the data, so the influence of the too high degree of minimum support is reduced so as to be favorable for the relation rule obtaining.

The mining process of relation rule consists of two steps.

(1) The first one is to find out all the frequent item sets, which are used to find out all the item sets whose support degree is not less than the received minimum support value.

(2) The second one is to bring the forceful relation rule by the frequent item sets, which must satisfy the minimum support and the minimum confidentiality. In the process of mining, the first step, which decides the general quality of mining relation rule, is the core of the relation rule finding algorithm.

By means of the mutual quality between the appointed min_sup and min_conf and the hunting algorithm of the frequent item sets as well as the mutuality with the relation rule sets, the users make explanations and evaluations on the mining results. This experiment module realizes the relation analysis with the improved algorithm of Apriori.

Because the traditional algorithm of Apriori needs to scan the whole data base in the process of producing the frequent item sets, a back-up set needs to be formed before the formation of each item set K, which will be the bottleneck of algorithm efficiency when there is a big amount of data. The rule amount mined by the algorithm of Apriori is large. It is reasonable to get informed of the original IP address to have analysis and judgments for the record data of network link, so it is not favorable for the analysis of invasion behavior when the received rule sets include no rules of the original IP address and that should be excluded.

The improved algorithm of Apriori consists of two parts which are bringing the frequent item sets and bringing the relation rule. The process of bringing the frequent item sets comprises linking, which is used to bring the back-up item set, and pruning, which excludes part of the item set elements by means of minimum support. The process of bringing the relation rule is to obtain the relation rule sets with the following formula.

\[
\text{count}(s) \geq \text{min}_\text{count}
\]

With the above considerations, the improved project of the algorithm of Apriori is put forward.

(1) the instantaneous linking and pruning
(2) The core quality parameter P is supposed to be introduced that means the parameter of the original IP address. In the process of bringing the relation rule, it is necessary to check whether there is the original IP address in the rule. Otherwise, the rule bringing should be quit.
C. The General Quality Detecting

We downloaded 3 group tcpdamp form of network traffic data from the site, base on behalf of the normal state of network traffic, net3 is the network traffic that includes analog IP Spoofing attacks, In the data file, an intruder is trying to guess the serial number to IP Gain access to the remote host, net3 is included simulated port scanning Describing the attack traffic .in the data file, the intruder tried to collect information about Web hosts and the services provided information. Experimental data can be easily obtained from the Internet, data is representative. Experiments show: Using the improved Apriori algorithm can improve the speed, the resulting rule set is also smaller.

KDD Cup 99 data set includes four major types of attacks, DOS attacks and port scanning attacks PROBE accounted for more than 80%, respectively, for detection of these two attacks. Select the data in the 15 key numeric attribute clustering, used in the clustering process does not record the type of identification, clustering the data set results can be clustered into different categories. Makes the exception classes and normal classes of data separately. Can be found through the experiment, take different number of clusters a great impact on the results, but can not predict the optimal number of clusters. The improved k-means clustering algorithm can solve the initial issue of the number of difficult choices. Clustering algorithm is introduced and the most close to the radius threshold, the number of clusters does not require pre-input data can be clustered into the best number. Table 1 shows the parameters of the DOS attacks on the introduction of test results. Experiments show that: the improved k-means algorithm DOS attack detection rate significantly increased, better clustering algorithm, and do not need to specify the final number of clusters in advance, you can change most near the threshold to control the particle clustering degree.

VI. CONCLUSION

Invasion detection has been developed fast in the recent years, which is the second security valve after the firewall as a kind of active detecting method of network security. The paper starts from the basic definitions, introduces invasion detection as well as several data mining methods commonly used for invasion detection and theoretically states the module of the invasion detection system based on data mining technology. The paper realizes a mining module based the mainframe log data. Besides, the abnormal visiting relation rule is brought by mining the log files of IIS so as to serve the invasion detection system.

There are various patterns of technology used for invasion detection. The technology of invasion detection based on data mining technology has become the highlight of the current development of invasion detection. However, data mining is still in the developing phase, so it is essential to make the deep research on data mining.

REFERENCES