IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 3 (Mar. - Apr. 2013), PP 20-23
www.iosrjournals.org

Application of Data Mining Technique in Invasion Recognition

M.Charles Arockiaraj
Asst.Professor, Computer Science Department Arakkonam Arts and Science College, Arakkonam-631003,

Abstract: *The article introduced the importance of invasion recognition, as well as the traditional invasion recognition's type and the limitation. Also, according to the invasion recognition's general process and data mining characteristic, it establishes a data mining-based model of network invasion recognition which is designed for its flaw. As a result, the missing report reduces greatly; the examination rate enhances; and network system's security strengthened. Finally, the article lists several hot topics which need to be further studied.*
Keyword: *Invasion recognition; Data Mining; Information Security*

I. Introduction

Along with the development and application of the computer network technology, the network intrusion event is getting more and more frequent, the harm which it creates is also getting more and more serious, the network security problem is increasingly prominent. Presently, the method of solving the network security problem includes the firewall, the data encryption, the identity validation, the invasion recognition and so on.

For those attempts to attack system's behavior in a normal way, the first three method can provide good guard function, but for the behavior, such as using the unusual method,making use of the system software's mistake or flaw, even using legal identity to harm the system safety,they actually seems helpless.

Under this kind of demand background, the invasion recognition has rapidly development. The invasion recognition, as the name suggests, is to detect intrusion behavior. It collects and analyzes the information from the certain key point in the computer network or computer system, in order to discover whether the network or system's behavior violates the security policy or any signs of being attacked.

The invasion recognition system is the combination of the software and hardware. It uses the intrusion detection technology to monitor the network and network system, and performs the different security action according to the result; in order to reduce the possible intrusion harm as much as it can.

II. Invasion Recognition System's Imitation

According to the detection principle, the invasion recognition technology may divide into Misuse Detection and Anomaly Detection.

2.1 Anomaly Detection

Anomaly Detection is the deviation between detection and acceptable behavior. If may define each acceptable behavior, each unacceptable behavior should be the intrusion. Summarizes characteristic which the normal operation should have when the user activity have the significant deviation from the normal behavior, namely considered as an intrusion. In this kind of detection, the rate of missing report is low, although can detect unknown intrusion effectively, the rate of false alarm is high

.

2.2 Misuse Detection

Misuse Detection is the match degree between detection and the unacceptable behavior. The security expert first collects the behavior characteristic of the unusual operation, builds the related characteristic library, Then at the time the monitored user or the system behavior match the record of library, the system regard this behavior as the intrusion. In this kind of detection, the rate of false alarm is low, but the rate of missing report is high. For the known attack, it may reports the attack type detailed and accurately; but For the unknown attack, it's function is limited, moreover, the characteristic library must be renew continually.

III. Data Mining Technology

Data Mining is refers to the process, it extracts effective, updated, latent, useful, and the understandable pattern from a lot of incomplete, noise, non-stable, vague and random data. In the invasion recognition system, the important information comes from the host log, the network data package, the system's log data against applications, and alarm messages that other invades the detection system or the system monitors. Because the

variety of data source and format, the complexity of the operating system increased and the big growth of network data traffic, audit data increased shapely and the data analysis duty is arduous.

The data mining technology has the huge advantage in the data extracting characteristic and the rule, so it is of great importance to use data mining technology in
the invasion recognition. The data mining technology was first applied in the invasion recognition research area by Lee and Salvatore J.Stolfo, Columbia University Wenke. Its idea is: Through analysis mining of the network data and the host call data discover misusing detection rule or exception detection model.

Applying the data mining technology in the invasion recognition can widely audit the data to obtain the model, thus it enables you to catch the actual invasion and the normal behavior pattern precisely. This automatic method does not need the manual analysis any more and the coding intrusion pattern, and when building normal skeleton, it no longer choose the statistical method by experience as before.

The main benefit is that the same data mining tool may apply in many data streams, so it is advantageous to build strong invasion recognition system.An important problem of Invasion recognition is how to effectively divide the normal behavior and
the abnormal behavior from a large number of raw data's attributes, and how to effectively generate automatic intrusion rules after collected raw network data. To accomplish this, various data mining algorithms must be studied, such as correlation analysis of data mining algorithms, sequence analysis of data mining algorithms, classification of data mining algorithms, and so on.

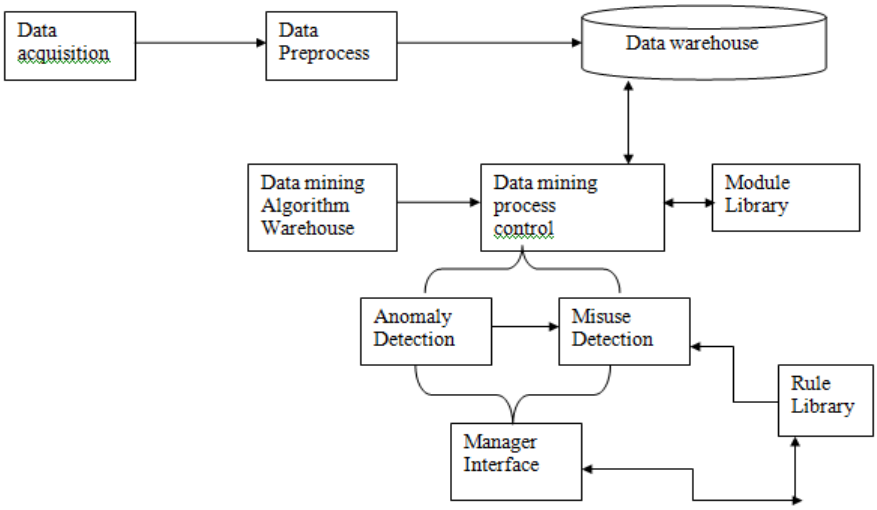
Correlation Analysis algorithm can be used to discover the relationship of attributes in network connection record, sequence analysis algorithms can discover the timing relationship of network connection records. Correlation analysis and sequence analysis of the data mining data mining algorithm can be used to obtain the normal behavior pattern, which is used in anomaly invasion recognition. Classification of data mining algorithm can mine the rules from trained data, which can identify normal behavior and the intrusion.

IV. System Design

First, the retried data is preprocessed by the data pre-processing module, to produce the data which conform to the data mining required, and store in the data warehouse ; the process control modules in data mining analysis the data in the warehouse, chooses the right data mining algorithms , and gives the rules that mining generates to Invasion recognition module ; according to the existing rule of the rule library, the invasion recognition modules will match the rules which data mining module outputs, and deliver the test results to the manager interface module; The manager interface module carries on the artificial judgment to the unusual pattern or the unknown pattern, to decide the measure.

4.1 System Model

Based on the general process of invasion recognition and the characteristics of data mining, we establish the application of data mining technology network invasion recognition system model (as shown in Figure 1).



4.2 Functions of the System Components

(1) Data acquisition

Intercepts the data which will be used in examining on the network, sets the Internet options to "promiscuous pattern", and catches every packet on the network for further processing.

(2) Data Preprocess Module

In the data mining, the data's quality directly impacts on the accuracy of the extracted user features and the derived rules. In data preprocess module, the data packet will be converted into suitable mining forms by using many techniques such as data cleaning, data integration, data reduction techniques, and using standardized data analysis.

(3) Data Mining Module

Data warehouse holds the data which is obtained from the data preprocess modules, and mines the training data which the control module produces. Because the data collection carries on unceasingly, data in data warehouse becomes also more and more rich. So it may provide all kinds of data to the system and mine more useful information. Data mining algorithm library is a set of many different data mining algorithm. Including sequence pattern analysis algorithm, connection rule algorithm, fuzzy clustering parsing algorithm, K average value cluster algorithm, sorting algorithm, neural network algorithm and so on.

For the efficiency reason, methods in the algorithm library should have be good time complexity, extensible, in order to instruct the search process of the data mining, and be used for the forecast or the increase mining at the next time. Data mining process control is playing a very important role in the system, it is the system extendibility key, it is responsible to choose the appropriate mining algorithm from the data mining algorithm library. Because the present various algorithms are inconsistent in each data set's performance. Not an algorithm in all data sets is on a good performance.

The algorithm choice dependents on the data type experimental result and expert's experience. It also contains a set of mining functional module. Because the traditional data mining-based intrusion detection system needs to mark the data as normal data or intrusion behavior, which means it needs well-marked data to obtain the training data set, but the expenses of obtains training data set is huge.

Therefore when the system that this article designs is in situation of having the data training set, the data mining control module may extract the characteristic and the detection pattern through the algorithm of the algorithm library; IF when the system begins running, there is not the training data set, the data mining control module may be trained by calling data of the data warehouse, and by using clustering algorithm, the data is marked as normal or attack data, and return the data warehouse as a training data.Finally, the mining results will be delivered to invasion recognition module.

(4) Invasion recognition Module Rules Library:

it stores the rules that invasion recognition needs; these rules will be used for match with data mining output modules. Anomaly Detection Module: This module summarizes the characteristic which the normal behavior should have, compare the rules from the current network data stream with the rules of the rules library, if the detected data is beyond the threshold, that is a intrusion;if not, that it is a normal behavior. Misuse detection modules: it assumes that all invasion behavior has detectable characteristics. When the detected system behavior matches records of the rule library, the system believes that such behavior is an intrusion.

(5) Manager Interface Module

The module is responsible to artificially make judgment on the unusual pattern or the unknown pattern. If the judgment result is the normal pattern, adds it to its close normal pattern in the regular library. If the judgment result is the unusual pattern, add it to its close unusual pattern in the regular library, and carries out the essential processing measure immediately.

V. Conclusion

The invasion recognition technology has gradually improved and matured in the research for many years, it becomes the important part of network security technology, and it marked the development process from network security technology to dynamic defense. Obviously, studding the intrusion detection technology and taking advantage of benefit of its dynamic defense, has the important practical significance to improve the network security. In future studies, the following questions should be considered deeply:

(1)Raw intrusion model rules can be obtained by the data mining, but those rules is only some allowed intrusion model rules. For how to validate mined rules dynamically and effectively, needs to study further.

(2) The need to improve the efficiency of data mining, the correlation rule mining algorithms and classification algorithms need to be further improved, in order to process magnanimous information effectively.

(3) Intelligent Invasion recognition that is, using intelligent ways and means for invasion recognition.More consistent solution is the combination of the invasion recognition systems of efficient conventional and intelligent detection software.

(4) Comprehensive Defense Schema: Deals with the network security problems with the ideas and methods of safety engineering risk management, treats the network security as a whole project Considers several aspects

including the management of the network, structure encryption, access, firewall, virus protection and invasion recognition to provide a comprehensive assessment for network, and then gives some feasible solutions.

References

[1] Julisch K. Data mining for invasion recognition: A critical review. IBM Research, Zurich Research Laboratory.

[2] E1-Sayed M, Ruiz C, Rundensteiner E A. FSMiner: efficient and incremental mining of frequent sequence patterns in Web logs. ACMWIDM'04, Washington DC, November 2004:12-13.

[3] Lee W, Stolfo S, J Mok K W. Data mining approaches for invasion recognition. Proceedings of the 7th USENIX Security Symposium, 1998.

[4] Cai Y, Clutter D, Pape G, et al. MAID: mining alarming incidents from Data Streams.ACM-SIGMOD Int Conf Management of Data (SIGMOD04), New York: ACM Press, 2004, 919-920.

[5] Chatzigiannakis V, Androutidakis, G, Maglaris B. A distributed invasion recognition prototype using security agents. HP Open View University Association, 2004.

[6] Kumar S, Spafford, E H.A pattern matching model for misuse invasion recognition. Proceedings of the 17th National computer Security Conference, 1994. 1277

[7] Boyer R.Moore J S.A fast string searching algorithm. Communication of the ACM, 1971,20(10):762-772.

[8] Jiawei Han, Micheline Kamber. Data mining concepts and techniques. Beijing: Mechanical industry publishing house 1-23, 70-94,152-168,188-196.

[9] Wayne A. Jansen. Invasion recognition with Mobile Agents Computer Commucications.2002 (25):96-99.