Research on the Method of Network Intrusion Detection Based on Data Mining Technology

Wei Wang[a, *], Cheng Yin[b]

Shandong Water Conservancy Vocational College, Rizhao China

[a]79553268@qq.com, [b]jyjsyc@163.com

Keywords: Intrusion detection; Data mining; Cluster analysis; Apriori; Ripper

Abstract: With the progress of network, the importance of security is become more and more obvious, the traditional security device firewall has difficult in protecting network security alone. However, current intrusion detection systems lack effectiveness, adaptability and extensibility, and especially, they become ineffective in the face of detecting new kind of attacks, and it already can't adapt to the trend of endless new attacks and increasing data quantity. In this paper, firstly, intrusion detection and data mining techniques are studied. By studying and analyzing the flaws of traditional IDS. Secondly, the course of data mining and the traditional intrusion detection are integrated to design an intrusion detection system based on the data mining technology. It designs the modules function of the system. The models design the rules database and the much emphasis is put on the design of the cluster analysis module and anomaly detector. Finally, it develops the Apriori and Ripper processes, which are combined in IDS by analysis of data mining technology and its specialties.

1. Introduction

With the growing popularity of computer networks and related technology continues to evolve, more and more businesses, governments and individuals doing business on the open Internet, access to information. But the Internet has brought great convenience to the public.[1-3] At the same time, the security has become a bottleneck for its development. Although in recent years has been more concerned about network security, but also the types of network intrusion, more and more techniques have become increasingly complex, always in the open is difficult to resist the external firewall to resist attacks against the firewall guard against hackers means of renovation, and many from the internal network attacks, making the original difficult to separate the firewall protection of network security, which constitutes a significant security threat.[4]

Currently, there are many network applications insecurity mainly as information leakage, information tampering, illegal use of network resources, illegal information infiltration. Computer network security risks and more are widespread anti-hacker weak. Government, business network was attack the events have occurred, causing enormous economic losses. [5] Therefore, the network information security and prevention become increasingly important.

Intrusion detection is a dynamic security protection measures, it can take the initiative to find intrusion signal to the network system to external attacks, internal attacks and misuse of security. Intrusion detection is divided into data collection, data analysis and response in three parts. [6] In order to find traces of intrusion and data acquisition from multiple points of network systems, including system log collection, network packets, important documents and the status of user activity and behavior. Data analysis is by pattern matching, anomaly detection and integrity of the testing by means of three techniques to analyze the data collected. Once intrusion detection system intrusion, and immediately enters the response process, including logging, alarm and security control. [7-8] This is a firewall, identification and authentication, and encryption, and many other security policies can not be done. Therefore, the introduction of IDS can make up for lack of other security policies, making the whole security system more perfect. [9] However, IDS Although the

  DOI: 10.25236/iccse.18.038

development has been relatively large, but the overall situation is not satisfactory. The biggest problem is the current intrusion detection products is relatively low detection accuracy, false positive and false negative compared to the situation more, and a large amount of data in the system response time is not ideal. [10] From a large number of audit data and network data packets that meaningful information will become very difficult, there has been data rich and information poor phenomenon. As a result, produced in the traditional intrusion detection technology based on the use of data mining technology is through analysis of multiple audit data submitted by the detection engine to find more complex intrusions.

2. Network Intrusion Detection Theory And Method

Intrusion detection is to find the behavior and deal with them, mainly for surveillance, collection of user behavior information, and analyze their behavior. Intrusion detection system is an important part of the protection system. If we can quickly detect the intrusion, in addition to confirm the intruder, the intruder threatened the system can also drive the system before the intruder. And even if not the first time detected the intruder, the faster the detection of illegal intrusion to the system losses are also smaller, and also the faster recovery system.

2.1 The Basic Principles of Intrusion Detection

Intrusion detection system is the core module of the two parts, which are mainly to complete the analysis and processing of events. Analysis module using the various methods of analysis and processing of existing events, to determine whether there are attacks that event, if there is an alarm, if you can not accurately judge to give a suspect value. Based on the analysis of the results, analysis module to decide which his doubts whether the data associated with the module to be sent to the integration of data processing, data fusion is aimed at combining different analysis modules have been transmitted from the value given in the event of suspected, to judge these events Whether there is a distributed attack. [11] If the analysis results are not transmitted to the associated module to make decisions directly to the management module; if sent to the associated modules, by association analysis module will further give management module.

When the alarm is spread to other information management module, management module is to decide whether to respond, if taken to respond to the response. Management by coordinating the interaction interface response information, the results with other security components of information sharing and response interaction. Man-machine interface configuration, maintenance system, the administrator can manually process information or respond.

2.2 The Basic Method of Intrusion Detection

Intrusion detection methods can generally detect signatures of known existing or pre-defined pattern of a suspected attack, as well as the normal pattern of activity from the observed detection of the abnormal behavior. Intrusion detection audit record is the foundation of practice is an ongoing record of each user operation, and record input to the intrusion detection system. The following is a common method of intrusion detection:

2.2.1 Anomaly detection method

Anomaly detection method is the use of statistical analysis in advance by the user on the computer usage habit, or use the flow of information within the network. It makes a statistical record in a time interval, the use of these behavioral data generated test rules, and then these rules detect whether the user appears unlawful. This method should define the frequency of events. Another statistical anomaly detection method is to record the behavior of each user, and then in order to detect abnormal user behavior.

2.2.2 Misuse detection method

Misuse detection system is rules-based approach, we must first define the rules of the attack, the rules defined by the system administrator to go out on the computer system or network characteristics of the environment, cause harm, and then rely on feature matching approach to determine May harm the system or network intrusion, attack. However, this method must be frequently updated signature, or for the invasion not defined is not easy to be detected out. In addition, expert systems can also be used to identify suspicious behavior.

2.2.3 Hybrid detection method

We use method of data mining. Mainly because of data mining method to generate new detection model, however, the new detection model can detect new attacks, and even to predict unknown attacks, anomaly detection can be applied to the Detection of certain well-known port.

2.3 The Common Technology of Intrusion Detection

Intrusion detection technology is an information identification and detection techniques. Therefore, the intrusion detection system can also be identified using the traditional information technology. However, the information and the identification and detection is usually compared to, intrusion detection also has its own characteristics. Because intrusion detection, not only the order information is important, and the information generated is also an important variable of time. Intrusion detection is the core of the invasion behavior analysis, intrusion detection technology is mainly determined on the invasion of behavior analysis and research, the intrusion of technology.[12] Intrusion detection technology currently used as follows:

2.3.1 Intrusion detection technology based on statistical method

Intrusion detection technology based on statistical method is evidence or a priori model, detection rules based on statistical methods of a detection technology. History of user behavior through modeling, real-time detection of the case users use the system, based on user behavior within the system saved the probability of a statistical model to detect user behavior, real-time detection system for unusual behavior, in order to determine whether the system attack.

2.3.2 Intrusion detection technology based on expert system

Network security experts suspicious behavior analysis experience formed the basis of a set of inference rules are based on expert system based on intrusion detection technology to build those rules automatically on detection expert system intrusion involved in analytical work.

2.3.3 Intrusion detection technology based on neural network

Since user behavior is highly complex, in order to accurately match the user's current behavior and user behavior model of probability and statistics is very difficult, based on statistical methods there are some intrusion detection techniques can not overcome the shortcomings. Statistical data on the statistical algorithm inaccurate assumptions result in many false positive alerts. Therefore, people using technology based on neural network for intrusion detection.

2.3.4 Intrusion detection technology based on model reasoning

The attacker before the invasion of a system usually acts by certain procedures, such as using a password guessing program to get the system login password, you can use these acts to constitute a program has certain behavioral characteristics of intrusion model. According to this model have included some sort of invasion of the behavioral characteristics of intent, there can be real-time detection of malicious intrusion attempts. People can use the model-based reasoning approach for the establishment of a specific model of certain acts, which can detect the invasion of those activities with a specific behavior.

3. The Intrusion Detection Model Based On Data Mining Technology

3.1 The Idea of Model Design

In this model, data mining technology is primarily used in cluster analysis. The non-supervised clustering is data mining anomaly detection system, a commonly used method. Anomaly detection model is the behavioral characteristics of the user's habits are stored in the feature database, and then the user behavior and characteristics of the current features of the database comparison, if the deviation between the two is large enough, then the invasion happened. Unsupervised anomaly detection rather than the method proposed is based on two premises: One is the normal data in the network data is far greater than the invasion of the number of data, and the other provided that the invasion of normal data and there is a big difference between the data. This method can be labeled from a set of data is not found in any invasion, and not worry about the data source is pure.

3.2 The System Structure And Work Flow

Based on the above design, this data mining technique used in network intrusion detection system, based on the proposed building in the Snort data mining based network intrusion detection system. The system includes the following specific functional modules:

3.2.1 Packet sniffer

The collection of data to the network, it is only a simple interface to capture information. Packet sniffer determines the location of intrusion detection level of local processing.

3.2.2 Decoder

When a packet is captured, the need for data link layer to decode the original packet. In the decoding process, the captured data to Packet data structure to the pre-processor for subsequent analysis and detection engine in preparation.

3.2.3 Data preprocessing

It is responsible for connecting the original data or data mining methods need to convert the data format. Include: further filtering, noise cancellation, third-party testing tools to detect known attacks.

3.2.4 Exception analyzer

It is responsible for using the network model of normal behavior after pretreatment test packets, discarding those that meet the model of normal data packet, the packet will be exceptions to the rules and regulations to match the list, if the match is successful, indicating an intrusion, this time alarm information. If abnormal data packet does not match with all the rules, then that may be unknown type of packet data generated by intrusion packet, it may be normal behavior unknown network packets, these packets will be sent to the cluster analysis module

3.2.5 Clustering Analyzer

Packet of these abnormal cluster analysis are in the clustering process will produce the new network model of normal behavior to abnormal parser, and for the failure to form a network model of normal behavior abnormal data packets, will be a record to the rules builder.

4. The Improved Apriori And Ripper Algorithm

In statistical analysis test of this hypothesis, there are usually two types of errors: false alarm rate and false negative rate. The corresponding error probability is expressed as α and β. Usually these two types of errors are difficult to be estimated. The threshold method was given as:

$$\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta} \le D(P_1 \| P_0) \tag{1}$$

This type of errors include two types of threshold hypothesis testing related to the probability distribution of entropy: greater related entropy means the stronger detection capability. In order to make statistical analysis system security, statistical analysis needs to reduce the related entropy, or even make it to be zero to obtain a perfect secure statistical analysis system. On the contrary, in order to design good statistical analysis algorithm, we need to look for characteristics of the probability distribution of carrier signal and the statistical analysis. While on the signals (images, audio, etc.) modeling, recent studies have made great progress, but on a unified model of the signal has not been established. However, given the situation that contains two kinds of signals (the original carrier signal and the statistical analysis signal); this problem can be solved through supervised learning approach. Therefore, statistical analysis faces enormous challenges in order to avoid changing the statistical features of cover signal when secret message has been embedded; conversely, statistical analysis is to seek the statistical difference between the feature vectors caused by information hiding.

As a dollar evolution of Gaussian probability density function, Gaussian mixture model (GMM) can be approximation of any probability density distribution of arbitrary shape, which is widely used in speech recognition. This paper also uses it to model the wavelet coefficients. In general, the Gaussian mixture distribution model can be the following using limited form of distribution and said:

$$f_k(x) = \sum_{j=1}^{k} \pi_j \phi(x, \theta_j) \tag{2}$$

Here, $\phi(x, \theta_j)$ is the j^{th} component of GMM model, θ_j is the vector of the mixture parameters which consists of weight π_j, mean μ_j, variance σ_j^2. The weight π_j must satisfy:

$$\pi_1 + \ldots + \pi_k = 1, \pi_j \geq 0 \tag{3}$$

As to a random variable X, its probability density function is denoted by $p(x)$. If it is Gaussian random variable with mean μ and variance σ^2, then its probability density is denoted by $N(\mu, \sigma^2)$. The characteristic function defined as follows:

$$\Phi(t) = \int_{-\infty}^{\infty} p(x) e^{jtx} dx \tag{4}$$

Here, $j = \sqrt{-1}$. Corresponding probability density can also use the following type:

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(t) e^{-jtx} dt \tag{5}$$

The above discussion of the model concerns the generic situation, but it did not tell me if it is sensitive to the statistical and analysis operation. Even if we do not know the exact statistical model and which embedded algorithm is used, but as described above, statistical analysis personnel can capture the signals prior to knowledge of statistical models and general statistical characteristics of statistical analysis algorithms.

5. Conclusion

Intrusion detection is the main area of network security technology research direction. Intrusion detection system network security has now become an integral part of, a firewall for the network's last strong security. Data mining techniques to intrusion detection systems can automatically large amount of network data from the new model was found to reduce the manual preparation of the invasion patterns of behavior and normal behavior patterns of the workload.

Data mining and intrusion detection research is the study of the two hot topics at home and abroad, there are still a series of theoretical and practical application of issues to be resolved, a

number of key technologies to be done to further study. Using data mining technology design and implement a highly efficient and accurate intrusion detection system is a large, complex project, not a personal power alone can accomplish in a short time, but also work in the future and improve gradually.

References

[1] Yun Wang; Weihuang Fu; Agrawal, D.P. Intrusion detection in Gaussian distributed Wireless Sensor Networks. IEEE 6th International Conference on Mobile Ad-hoc and Sensor Systems. pp: 313-321. 2013

[2] Yongquan Mo; Yizhong Ma; Liang Xu. Design and implementation of intrusion detection based on mobile agents. IEEE International Symposium on IT in Medicine and Education. pp: 278-281. 2015

[3] Yu-Xin Ding; Min Xiao; Ai-Wu Liu. Research and implementation on snort-based hybrid intrusion detection system. IEEE International Conference on Machine Learning and Cybernetics. pp: 1414-1418. 2016

[4] Lin Ying; Zhang Yan; Ou Yang-jia. The Design and Implementation of Host-Based Intrusion Detection System. The Third International Symposium on Intelligent Information Technology and Security Informatics. pp: 595-598. 2016

[5] Qingqing Zhang; Hongbian Yang; Kai Li; Qian Zhang. Research on the intrusion detection technology with hybrid model. International Conference on Environmental Science and Information Application Technology. pp: 646-649. 2017

[6] Duanyang Zhao; Qingxiang Xu; Zhilin Feng. Analysis and Design for Intrusion Detection System Based on Data Mining. Second International Workshop on Education Technology and Computer Science. pp: 339-342. 2016

[7] Farid, D.M.; Rahman, M.Z. Learning intrusion detection based on adaptive bayesian algorithm. 11th International Conference on Computer and Information Technology. pp: 652-656. 2015

[8] Meijuan Gao; Jingwen Tian; Mingping Xia. Intrusion Detection Method Based on Classify Support Vector Machine. Second International Conference on Intelligent Computation Technology and Automation. pp:391-394. 2015

[9] Tian-rui Li; Wu-ming Pan. Intrusion detection system based on new association rule mining model. IEEE International Conference on Granular Computing. pp: 512-515. 2015

[10] WenJie Tian; JiCheng Liu. A new network intrusion detection identification model research. 2nd International Asia Conference on Informatics in Control, Automation and Robotics. pp:9-12. 2016

[11] Nie J. Research on Task Scheduling Strategy Based on Cloud Computing Environment[J]. Journal of Applied Science and Engineering Innovation, 2018, 5(1): 9-12.

[12] Li R, Yu R, Wang X. Information Resources Sharing Security in Cloud Computing[J]. Journal of Applied Science and Engineering Innovation, 2018, 5(3): 65-68.