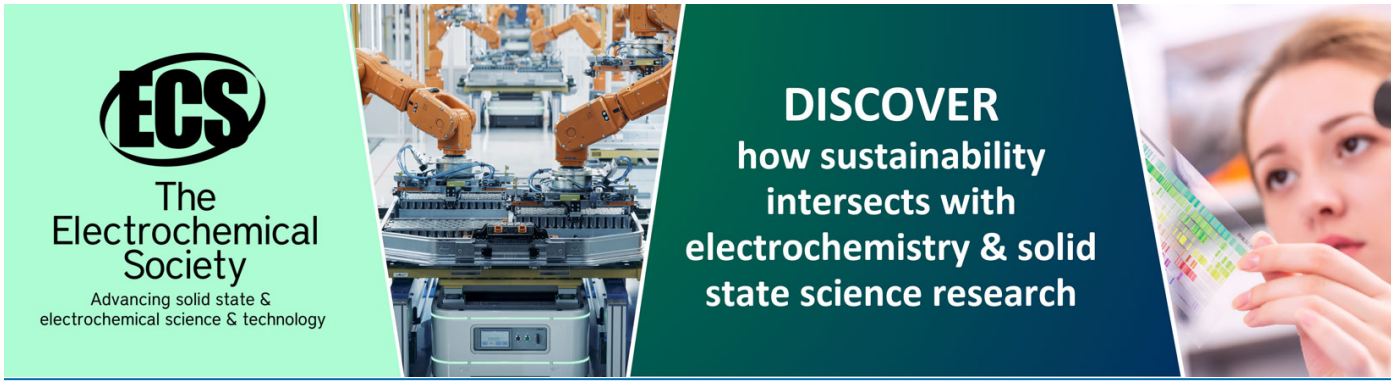
Research on Application of Data Mining Technology in Network Intrusion Detection



Research on Application of Data Mining Technology in Network Intrusion Detection

Haibo Song[1*], Yilin Yin[2]

[1]School of management, Tianjin University of Technology, Tianjin, 300000, China;

[2]School of management, Tianjin University, Tianjin, 300000, China;

[*]Corresponding author's e-mail: 1441317029@qq.com

Abstract. In order to improve the efficiency of intrusion detection systems, data mining technology is applied to network intrusion detection. This article introduces the basic concepts of intrusion detection systems, describes the techniques commonly used in data mining research in intrusion detection systems, proposes an intrusion detection system based on data mining, and an improved k-means algorithm.

1. Introduction

In the context of the era of big data, people's dependence on the Internet continues to increase, and network and information security have become serious issues facing people. At present, four methods are used to solve network security problems: firewalls, data encryption, identity authentication, and intrusion detection. The first three methods have better protection against attacks on the system through normal channels, but they cannot do anything to endanger the security of the system by using abnormal means or legal identities. Therefore, people urgently need to adopt intrusion detection technology to meet the increasing security needs.

2. Data mining technology and intrusion detection system

2.1 Data Mining Technology

Data mining technology is to find valuable data rules or data patterns in a large amount of data, and then provide auxiliary services to decision makers through analysis and processing. Its main steps include data collection, data preprocessing, data mining, and knowledge representation[1]. Data collection refers to the acquisition and selection of relevant data from databases or other information bases; data preprocessing, that is, data cleaning, refers to eliminating noise or interference data, unifying or converting the data into a data format suitable for mining; data mining refers to Relevant algorithms or intelligent methods are used to extract the required data patterns; knowledge representation refers to using a visual or graphical interface to show users the mining knowledge. There are various algorithms for different areas of data mining. The commonly used methods are association analysis, classification analysis, cluster analysis, etc.

2.2 Intrusion detection system

Intrusion detection is to collect information on several key points in computer networks and systems, and then analyze these information to discover various attack attempts, behaviors, or attack results in the network or system[2]. An intrusion detection system is a security system that can identify attacks and

malicious access behaviors in the network in a timely manner and make a certain response, and comprehensively detect behaviors that affect the integrity, confidentiality, and availability of system information. The intrusion detection system mainly has three functional modules: data acquisition, data analysis, and response processing.

2.3 Necessity and Possibility Analysis of the Combination of Data Mining and Intrusion Detection System

The development of intrusion detection systems has been facing many challenges so far. The high false positive rate and false negative rate, low detection efficiency, and lack of intelligence in the intrusion detection system have limited the forward development of intrusion detection technology. Faced with these problems, the technical characteristics of data mining just meet the functional requirements of intrusion detection. Compared with other data processing and analysis technologies, data mining has obvious advantages and possibilities in the field of intrusion detection, mainly in terms of intelligence, efficiency and strong adaptability. The mainstream data mining technologies, such as classification analysis, cluster analysis, and association analysis, can be applied to the field of intrusion detection systems, thereby explaining that data mining technology has great application possibilities in intrusion detection.

3. Application of data mining in intrusion detection

3.1 A new model of intrusion detection system based on data mining technology

There is only a small amount of abnormal data in the network data. If the normal data can be filtered out, the detection efficiency of the intrusion detection system can be improved. In order to filter the normal data in the network, it is necessary to generate accurate behavior patterns that can identify the normal data. Cluster analysis is an effective way to construct the normal behavior pattern of the network. The data packets that do not conform to the normal behavior pattern are regarded as abnormal data packets and further detected by the detector in the system. The feature extraction module is used to analyze the intrusion behavior of abnormal data packets. The new intrusion detection rules are added to the rule base, so the detector can detect new unknown intrusion behaviors.

The new model is to add three modules of feature extraction module, pre-detection module and cluster analysis module to the model of the original intrusion detection system. The behavior analysis module constructs a network normal behavior detection mode based on clustering-related algorithms. The rule base is a repository for intrusion rules and provides the basis for intrusion detection. The feature extraction module correlates and analyzes the data records in the log to generate relevant rules, and can convert the rules into intrusion detection rules that conform to the syntax of the intrusion rules, and finally add them to the rule base. The data packet acquisition module uses Libpcap to capture network data packets. Data packet analysis module: Decodes and analyzes the detected data packets, and stores the analysis results in the specified data structure. Pre-processing module: pre-processing before data matching is performed on decoded data packets by calling related pre-processing functions, such as data fragmentation functions. Pre-detection module: Filter the network data by using the mode of the cluster analysis module to remove normal network data packets. System detection module: Compare the data packet with the rules in the rule base to determine whether the behavior is an intrusion behavior. The new intrusion detection system model is shown in Figure 1.

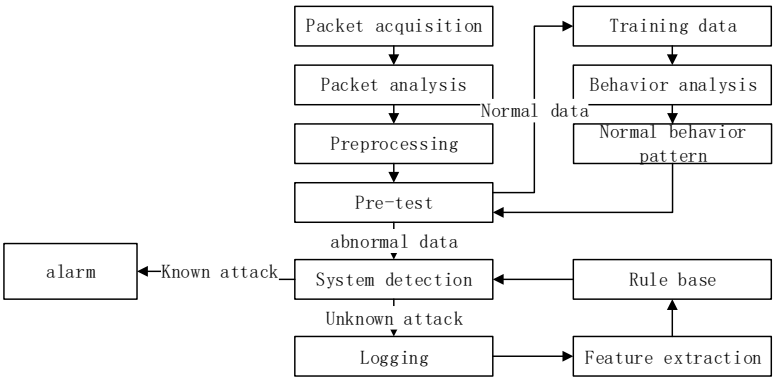


Figure 1. New intrusion detection system model

3.2 Data mining technology to construct an intrusion detection model

3.2.1 General Process of K-Means Algorithm

The k-means algorithm takes k as a parameter and divides n objects into k clusters so that the clusters have a high degree of similarity[3]. It is the most widely used clustering algorithm. K objects are randomly selected, and each object initially represents the average or center of a cluster. For each remaining object, it is assigned to the nearest cluster based on the distance of the remaining cluster centers. Then recalculate the average of each cluster. The process is repeated until the criterion function converges.

3.2.2 Data mining technology applied to Snort intrusion detection system

Selecting different initial clustering centers produces different clustering results with different accuracy rates[4]. If k initial centers can be found, which respectively represent data sets with a large degree of similarity, then the initial cluster centers are found. In order to find a data set that is consistent with the spatial distribution of the data and has a large degree of similarity, take the following steps: calculate the distance between the pair of data objects and set a distance threshold; then find the two data that are closest Objects to form a data object set A1 and delete them from the total data set U; calculate the distance between each data object in A1 and each sample in the data object set U, and find the closest in U to A1 in U Data object, merge it into the collection A1 and delete it from U until the distance between the data object in A1 and the data object in U reaches a certain threshold; then find two samples of the closest data object from U A2, repeat the above process until k object sets are formed; finally, the k object sets are arithmetically averaged to form k initial cluster centers. This improves the accuracy of the clustering results.

3.2.3 Result analysis

As an open data mining work platform, WEKA integrates a large number of machine learning algorithms that can undertake data mining tasks. In this paper, the data set in UCI is used to test the validity of the algorithm. The improved algorithm is implemented in the WEKA platform using Java language, and the algorithm is embedded in the WEKA platform. The experimental data comparison results are shown in Table 1.

Table 1. Comparison of experimental data

	General K-Means Algorithm	Improved K-Means algorithm
Seed	Sum of clustering squared errors	Sum of clustering squared errors
90	1568.4335	1153.2356
95	1543.4366	955.6528
97	1509.6754	801.2576
100	1553.5693	1025.3379

The sum of the squared errors of the clusters is a criterion for evaluating the quality of the clusters. The smaller the value, the smaller the distance between instances of the same cluster. Under the same seed, the sum of the clustered squared error data of the improved K-Means algorithm is small, and the improved K-Means algorithm can effectively improve the accuracy of intrusion detection.

4. Conclusion

At present, the intrusion detection system has developed into a key component of the security network system, and is an important part of the information security strategy of the defense-in-depth system. Data mining-based network intrusion detection systems can solve certain limitations of intrusion detection systems, greatly improve the efficiency of detection, and automatically discover new rule patterns from a large amount of data, greatly reducing the number of previously relying on hand-written patterns by domain experts. The workload is very good.

References

[1] Luo Jiaohuang. Application of data mining technology in intrusion detection system [J]. Journal of Jilin Normal University (NATURAL SCIENCE EDITION), 2016,37 (02): 131-135
[2] Liu Zechen. Application and research of data mining technology in network intrusion detection [J]. Information recording materials, 2019,20 (08): 188-189
[3] Solane Duque, Mohd. Nizam bin Omar. Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)[J]. Procedia Computer Science,2015,61.
[4] Shi Dongsheng. Application of data mining technology in network intrusion detection system [J]. Microcomputer information, 2010,26 (30): 81-82 + 99