# Detection of Network Protection Security Vulnerability Intrusion Based on Data Mining

Jinming Zhang

*(Corresponding author: Jinming Zhang)*

Department of Information Technology, Yantai Vocational College
Yantai, Shandong 264670, China
(Email: jinmingzjm@126.com)

*(Received Feb. 2, 2019; Revised and Accepted Aug. 11, 2019; First Online Oct. 1, 2019)*

## Abstract

With the rapid development of Internet technology, network security has received more and more attention. Therefore, the detection of network protection security vulnerability intrusion has become an urgent task with some practical and guiding significance. In this paper, intrusion detection system (IDS) is taken as the research object to establish a data mining-based IDS model, the experimental results are obtained, and the relevant experimental conclusions are drawn. At the same time, it is compared with the traditional IDS, and six experiments are carried out. The output results of the detection rate, false negative rate and false positive rate of the two different methods in six experiments are obtained. The experimental conclusions that the network protection security performance of IDS using the data mining is better, and the detection capability of network vulnerability intrusion is stronger are drawn. This study provides a new route for the research on the detection of network protection security vulnerability intrusion.

*Keywords: Data Mining; Intrusion Detection System; Protection Security; Vulnerability*

## 1 Introduction

In the modern era, the network is slowly integrated into people's daily life, which has become an indispensable part of people's life. At the same time, the amount of data information on the network has also increased in abundance, which is followed by the frequent occurrence of unlawful incidents of data leakage on the Internet [3]. It not only exposes the privacy of individuals and businesses, but also poses some risks and safety hazards to individuals and businesses. Therefore, intrusion detection system (IDS) is quite important in this network environment, which can improve the security performance of network protection [10]. IDS mainly refers to a new detection mode [6] extending from traditional firewall technology, which relies mainly on intrusion detection technology [12] and monitor events in the network through corresponding working principles. At present, there are more and more researches on IDS, and methods such as data mining, statistical models, *etc.* can be used to improve and optimize them, thereby further improving the detection performance of network protection security vulnerability intrusion and the current network environment [1, 20].

In response to this problem, many experts and scholars have put forward their own opinions and views. Aparicio-Navarro *et al.* [4] believed that new and more powerful detection mechanisms need to be developed as the complexity of cyber attacks increases and proposed that next-generation IDS should be able to adjust its detection characteristics based not only on measurable network traffic, but also on available advanced information related to the protected network to improve its detection results. Chakchai *et al.* [19] believed that with the rapid development of the Internet, the number of network attacks has increased. Therefore, a model of network intrusion detection data mining classification was proposed. Hachmi and Limam [7] proposed a two-stage technology improved IDS based on the data mining algorithm and verified the performance of low false positive rate of the system. The experimental effect was significant. In this paper, IDS is taken as the research object to establish a data mining-based IDS model, the experimental results are obtained, and the relevant experimental conclusions are drawn. This study provides a reference for the research on the detection of network protection security vulnerability intrusion.

## 2 Data Mining

Data mining mainly refers to the process of searching for previously unknown but valuable and meaningful information through algorithms in a large number of data [18], which is an important operation step in knowledge-discovery in databases (KDD) [21] and relies mainly on artificial intelligence technology, statistics, *etc.* [11]. The main objectives of data mining include clas-

sification, clustering, prediction, bias analysis, *etc.* [16]. The main methods of data mining include mathematical statistical analysis, machine learning, *etc.* Data mining technology can mine normal or intrusive behavior patterns from large-scale audit data [23], where audit data is mainly composed of pre-processed and time-stamped audit records [17], and each audit record has some characteristics.

Data mining is widely used in various fields because of its own advantages [9], in which data mining is closely related to the computer field [5]. With the massive growth of network data information, network problems such as data information leakage, *etc.* have emerged one after another, so network security has become an arduous task. Therefore, it needs to combine some new ways to protect network security. In this paper, the detection performance of network protection security vulnerabilities intrusion is specifically studied by the method of data mining to verify its feasibility and practicality.

# 3 Detection of Network Protection Security Vulnerability Intrusion

## 3.1 IDS

The main working principle of IDS [15] is to perform correlation analysis on data information related to security in the network under the condition that the existing network performance is not affected, so as to detect intrusion behaviors. The role of IDS is [22] to identify illegal intrusion behaviors to perform corresponding response operations [14] and to detect system construction, weakness audit and user behaviors [2]. The main features of IDS include accuracy, scalability and fault tolerance [8]. In this study, IDS is taken as the main research object to carry out relevant simulation experiments, in which the network protection security vulnerability intrusion is mainly detected. Figure 1 shows the main components of IDS.

## 3.2 Research Based on Data Mining

In this study, IDS is taken as the main research object. The K-means clustering algorithm in data mining is used to detect network protection security vulnerability intrusion, and a model of IDS based on data mining is established. Firstly, through the corresponding collection system, the required data is selected as the initial clustering center. Then the relevant calculations are performed for each cluster center to obtain the relevant output results. Finally, though the output results obtained by the clustering calculation, the connection records are reasonably and scientifically assigned to distinguish the normal or abnormal connection records, and the relevant data is classified by the normal behavior pattern class and the abnormal behavior pattern class, thereby detecting the network protection security vulnerability invasion.

The main content of the $K$-means clustering algorithm is to use the similarity between the data through the iterative idea as a standard of the measure, to classify the objects into different similarity categories, making the internal similarity of each class high. In this algorithm, the solution of the cluster radius is inseparable from the data set itself. By extracting the distance characteristics of the data set itself, the appropriate cluster radius should be determined before clustering.

$k$ represents the number of clusters, and the inaccuracy of the value of $k$ will affect the quality of the final clustering results in this algorithm. Therefore, determining a suitable value of $k$ is a major focus of the algorithm. When choosing the appropriate value of $k$, it needs to pay special attention to the two parameters of intra-class distance and inter-class distance. Specifically speaking, when the clustering effect is better, the intra-class distance is smaller, but the inter-class distance is larger. Therefore, in order to better balance the relationship between the two parameters, the method of linear combination is mainly adopted to carry out calculation and solution in this study.

In this algorithm, the Euclidean distance calculation method is used in this paper. It is assumed that the size of the data set is $m$, $I$ indicates the number of iteration, $Z_j(I)$ represents the clustering center of category $j$, $X_i^{(j)}$ represents any data object in the class $j$, $Z_j$ represents the new clustering center of class $j$, $I = 1$ is taken, and $K$ initial clustering centers are selected, $Z_j(I)$, $j = 1, 2, \cdots, k$. The Euclidean distance between each data object and cluster center is calculated, $L(X_i, Z_j(I)) = \min\{(X_i, Z_j, (I)), i = 1, 2, \cdots, m, j = 1, 2, \cdots, K\}$. If the obtained Euclidean distance meets $L(X_i, Z_j(I)) = \min\{(X_i, Z_j(I)), j = 1, 2, \cdots, k\}$, $X_i \in W_k$ will be obtained.

The sum of the squares of the distances from all samples in the cluster domain to the cluster center is expressed as $H(c)$, $\delta$ indicates the iteration termination threshold, which needs to be determined whether it meets:

$$|H_c(I) - H_c(I-1)| < \delta. \tag{1}$$

If Equation (1) mentioned above is met, the algorithm will end, otherwise $I = I + 1$, $k$ new cluster centers will be continued to calculate:

$$Z_j = \frac{1}{m} \sum_{i=1}^{m_j} X_i^{(j)}$$

The squared error criterion is also used in the algorithm, the sum of the squared errors of all samples in the data set is expressed as $E$, $o$ represents the point in space, and $m_i$ represents the average value of cluster $C_i$. $E$, the optimal result, can be defined as:

$$\min E = \min \sum_{i=1}^{k} \sum_{o \in C_i} |o - m_i|.$$

The detection performance of IDS is deeply analyzed through relevant parameters, wherein $P_a$, $P_b$, $P_c$ indicate

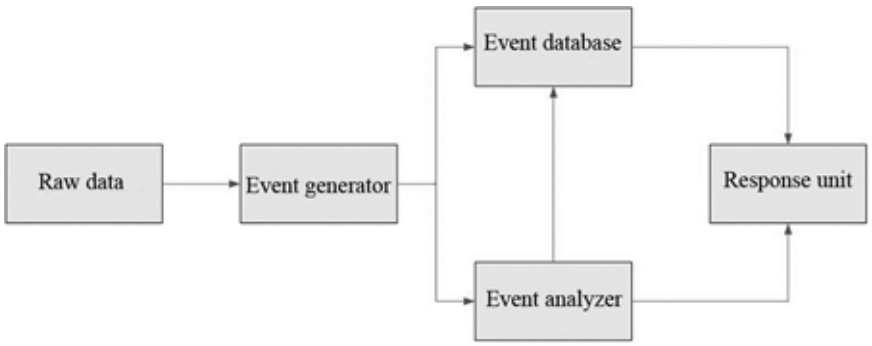


Figure 1: Main components of IDS

the detection rate, false negative rate, and false positive rate of the system respectively, $N_d$ indicates the number of intrusion events detected correctly, $M$ indicates the number of all intrusion events, $M_{total}$ indicates the number of all events, $N_e$ indicates the number of false negative intrusion events, and $N_f$ indicates the number of false positive intrusion events. The formula shown below can be obtained:

$$P_a = \frac{N_d}{(}M) \times 100\%$$
$$P_b = \frac{N_e}{(}M) \times 100\%$$
$$P_c = \frac{N_f}{(}M_{total}) \times 100\%$$

# 4 Simulation Experiment

## 4.1 Experimental Methods and Parameters

In this study, the problem of the detection of network protection security vulnerability intrusion based on data mining is mainly researched. In this simulation experiment, the used related software is Matlab7.0, and the used experimental data set is KDD Cup 99, in which the test data includes 5000 pieces, and the training data includes 600 pieces. In the set of test data, normal data accounts for 75%, and vulnerability data accounts for 25%. Corresponding cluster analysis is performed on the data, and the vulnerability data in the experimental data is detected to obtain the output results of relevant parameters. At the same time, it is compared with the traditional IDS [13], the both methods are run six times respectively, the results of related parameters such as the detection rate of six different experiments are obtained, and the corresponding experimental conclusions are drawn for reference.

## 4.2 Experimental Results

(1) Cluster analysis Cluster analysis is performed on the data set of this simulation experiment, and the value of the cluster radius is adjusted to obtain the corresponding output results and experimental conclusions. Table 1 and Figure 2 can be obtained.

It can be seen from Table 1 that the total number of clusters shows a decreasing trend as the cluster radius increases. When the cluster radius is 1, the total number of clusters reaches the maximum value, i.e., 199. When the cluster radius is 10, the total number of clusters reaches the minimum value, i.e., 127. It can be obtained that the larger the cluster radius is, the smaller the total number of clusters is. Therefore, the cluster radius is inversely proportional to the total number of clusters, the more clusters are, the more detailed the cluster analysis is, and the smaller the false positive rate of IDS is.

It can be seen from Figure 2 that the cluster accuracy decreases as the cluster radius increases. When the cluster radius is 1, the cluster accuracy reaches the maximum value, i.e., 83.69%. When the cluster radius is 10, the cluster accuracy reaches the minimum value, 66.71%. It can be found that the smaller the cluster radius is, the better the accuracy of the algorithm is. Adjusting the cluster radius can effectively improve the cluster effect of the algorithm and have more obvious detection effect of vulnerability intrusion.

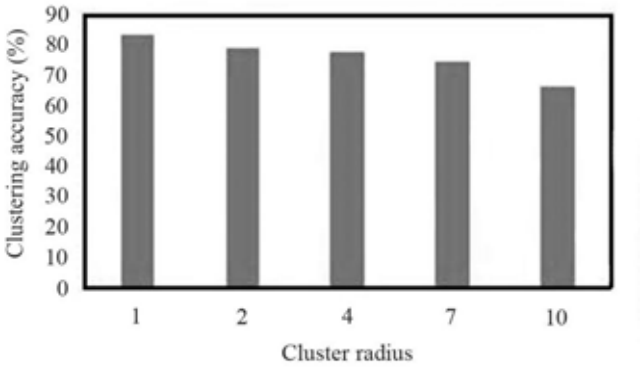


Figure 2: Comparison of the clustering accuracy

(2) Detection efficiency Figure 3 shows the detection

Table 1: The number of clusters

| Cluster radius | Normal behavior pattern class | Abnormal behavior pattern class | Total number of clusters |
|---|---|---|---|
| 1 | 1 | 198 | 199 |
| 2 | 0 | 196 | 196 |
| 4 | 2 | 189 | 191 |
| 7 | 6 | 172 | 178 |
| 10 | 11 | 116 | 127 |

time of the vulnerability data by IDS. It can be seen from Figure 3 that the detection time also shows a growing trend as the number of data increases. When the number of data reaches 700, the detection time of IDS reaches a maximum value, i.e., 16.5 s. The rate of increase shows a downward trend and gradually stabilizes although the time of detection is constantly increasing. Therefore, it can be obtained that the IDS using data mining has high detection efficiency for vulnerability data and remarkable experimental effect.

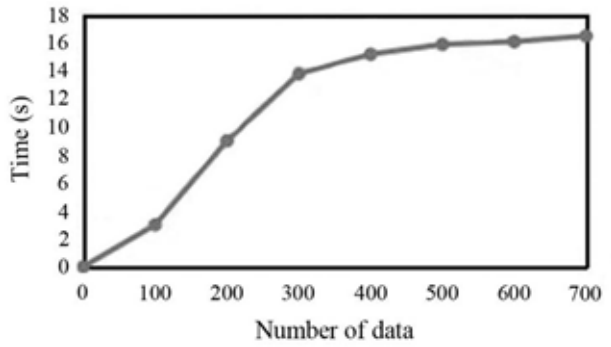


Figure 3: Detection time of vulnerability data by IDS

### 4.3 Comparative Analysis

In order to verify the performance of the IDS using data mining mentioned in this paper, it is compared with the traditional IDS, and Figure 4 and Table 2 are obtain. Figure 4 shows the comparative analysis of the relevant parameters between the two different methods in six experiments. In the six-time experiment, the IDS using data mining reaches the maximum value of the detection rate in the third experiment, i.e., 96.45% and the minimum value of the detection rate in the second experiment, i.e., 94.69%. The traditional IDS reaches the maximum value of the detection rate in the first experiment, i.e., 88.3% and the minimum value of the detection rate in the fifth experiment, i.e., 84.26%. Compared with the traditional IDS, the detection rate of the IDS using data mining is higher, and the experimental effect is more obvious. Therefore, it can be obtained that the ids using data mining has better detection performance, which is beneficial to the detection of network protection security and vulnerability intrusion.

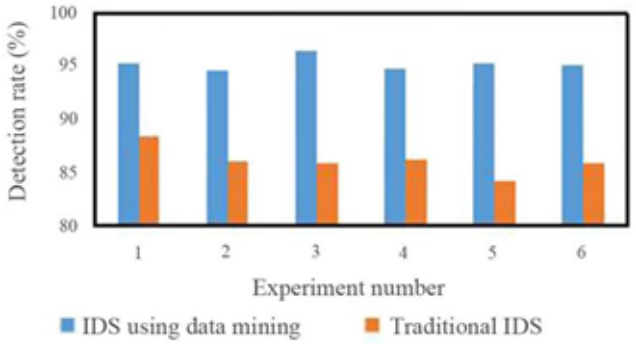


Figure 4: Detection rate in two different ways

It can be seen from Table 2 that the average detection rate of the traditional IDS is 87.75%, the average false negative rate is 28.49%, and the average false positive rate is 4.72%; while the average detection rate of the IDS using data mining is 95.63%, the average false negative rate is 20.23%, and the average false positive rate is 2.81%. Compared with the traditional IDS, the average detection rate of the IDS using data mining is higher, and the average false negative rate and average false positive rate are lower. Therefore, it can be obtained that the performance of the IDS using data mining is better, and the optimization effect on the experiment is more obvious. Therefore, in the future, the IDS using data mining has better development space and potential for the detection of network protection security vulnerability intrusion, but the traditional IDS needs continuous improvement and optimization.

## 5 Conclusion

Nowadays, network security issues are getting more and more attention. Therefore, in this paper, IDS is taken as the research object, and the corresponding simulation experiments are carried out. The obtained experimental results are as follows: the relationship between the cluster radius and the total number of clusters is inversely proportional; the smaller the cluster radius is, the better

Table 2: Comparison of the related parameters in two different ways

| | Average detection rate (%) | Average false negative rate (%) | Average false positive rate (%) |
|---|---|---|---|
| IDS using data mining | 95.63 | 20.23 | 2.81 |
| Traditional IDS | 87.75 | 28.49 | 4.72 |

the accuracy of the K-means clustering algorithm is; as the number of data increases, the detection time shows a growing trend, but the growth rate tends to be stable. At the same time, through comparing with the traditional IDS, the experimental results that the average detection rate of the IDS using data mining is higher than that of the traditional IDS, and the average false negative rate and false positive rate are lower are obtained. The experimental conclusion that the IDS using data mining has better detection performance for network protection security vulnerability intrusion is drawn. This study provides a new model for the research on the detection of network protection security vulnerability intrusion.

## References


[1] A. A. Al-khatib, W. A. Hammood, "Mobile malware and defending systems: Comparison study," *International Journal of Electronics and Information Engineering*, vol. 6, no. 2, pp. 116–123, 2017.

[2] N. Y. Almusallam, Z. Tari, P. Bertok, *et al.*, "Dimensionality reduction for intrusion detection systems in multi-data streams — A review and proposal of unsupervised feature selection scheme," in *Emergent Computation*, pp. 467–487, 2017.

[3] M. H. R. Al-Shaikhly, H. M. El-Bakry, and A. A. Saleh, "Cloud security using Markov chain and genetic algorithm," *International Journal of Electronics and Information Engineering*, vol. 8, no. 2, pp. 96–106, 2018.

[4] F. J. Aparicio-Navarro, J. A. Chambers, K. Kyriakopoulos, *et al.*, "Using the pattern-of-life in networks to improve the effectiveness of intrusion detection systems," in *IEEE International Conference on Communications*, pp. 1–17, 2017.

[5] W. Chen, H. R. Pourghasemi, S. A. Naghibi, "Prioritization of landslide conditioning factors and its spatial modeling in Shangnan County, China using GIS-based data mining algorithms," *Bulletin of Engineering Geology and the Environment*, vol. 77, no. 2, pp. 611–629, 2017.

[6] B. Elhadj, W. Thomas, H. Walaa, "A critical review of practices and challenges in intrusion detection systems for IoT: Towards universal and resilient systems," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3496–3509, 2018.

[7] H. Fatma, L. Mohamed, "A two-stage technique to improve intrusion detection systems based on data mining algorithms," in *IEEE the 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO'13)*, pp. 1–6, 2013.

[8] H. Hammami, H. Brahmi and S. Ben Yahia, "Security insurance of cloud computing services through cross roads of human-immune and intrusion-detection systems," in *The 32nd International Conference on Information Networking (ICOIN'18)*, pp. 174-181, 2018.

[9] C. Helma, T. Cramer, S. Kramer, *et al.*, "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds," *Journal of Chemical Information and Modeling*, vol. 44, no. 4, pp. 1402–1411, 2004.

[10] S. Islam, H. Ali, A. Habib, N. Nobi, M. Alam, and D. Hossain, "Threat minimization by design and deployment of secured networking model," *International Journal of Electronics and Information Engineering*, vol. 8, no. 2, pp. 135–144, 2018.

[11] B. Kang, J. Lijffijt, R. Santos-Rodríguez, *et al.*, "Subjectively interesting component analysis," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1615–1624, 2016.

[12] A. Keramatpour, A. Nikanjam, H. Ghaffarian, "Deployment of wireless intrusion detection systems to provide the most possible coverage in wireless sensor networks without infrastructures," *Wireless Personal Communications*, vol. 96, no. 3, pp. 1–14, 2017.

[13] M. Keshk, N. Moustafa, E. Sitnikova, *et al.*, "Privacy preservation intrusion detection technique for SCADA systems," in *IEEE Military Communications and Information Systems Conference (MilCIS'17)*, pp. 1–6, 2017.

[14] T. Miquel, J. Condomines, R. Chemali and N. Larrieu, "Design of a robust controller/observer for TCP/AQM network: First application to intrusion detection systems for drone fleet," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'17)*, pp. 1707-1712, 2017.

[15] N. Moustafa, J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Military Communications and Information Systems Conference (MilCIS'15)*, pp. 1–6, 2015. DOI: 10.1109/MilCIS.2015.7348942.

[16] R. Mousheimish, Y. Taher, K. Zeitouni and M. Dubus, "PACT-ART: Enrichment, data mining, and

complex event processing in the internet of cultural things," in *The 12th International Conference on Signal-Image Technology & Internet-Based Systems*, pp. 476–483, 2016.

[17] L. G. Noemí, A. S. Santiago, M. R. A. Blanca, M. O. M. Montserrat, T. Y. Chiang, "Divide and conquer! Data-mining tools and sequential multivariate analysis to search for diagnostic morphological characters within a plant polyploid complex (Veronica subsect. Pentasepalae, Plantaginaceae)," *Plos One*, vol. 13, no. 6, pp. e0199818, 2018.

[18] G. Omid, R. Hashem, B. Thomas, *et al.*, "A new GIS-based data mining technique using an adaptive neuro-fuzzy inference system (ANFIS) and k-fold cross-validation approach for land subsidence susceptibility mapping," *Natural Hazards*, vol. 94, no. 2, pp. 497–517, 2018.

[19] C. So-In, N. Mongkonchai, P. Aimtongkham, K. Wijitsopon and K. Rujirakul, "An evaluation of data mining classification models for network intrusion detection," in *International Conference on Digital Information & Communication Technology & Its Applications*, pp. 90-94, 2014.

[20] B. Subba, S. Biswas, S. Karmakar, "Enhancing performance of anomaly based intrusion detection systems through dimensionality reduction using principal component analysis," in *IEEE International Conference on Advanced Networks & Telecommunications Systems*, pp. 1-6, 2016.

[21] K. R. Thorp, G. Wang, K. F. Bronson, *et al.*, "Hyperspectral data mining to identify relevant canopy spectral features for estimating durum wheat growth, nitrogen status, and grain yield," *Computers and Electronics in Agriculture*, vol. 136, pp. 1–12, 2017.

[22] S. Vakili, J. M. P. Langlois, B. Boughzala, *et al.*, "Memory-efficient string matching for intrusion detection systems using a high-precision pattern grouping algorithm," in *Proceedings of Symposium on Architectures for Networking and Communications Systems (ANCS'16)*, pp. 37–42, 2016.

[23] Q. Wang, G. Z. Yao, G. M. Pan, *et al.*, "Analysis of on medication rules for Qi-deficiency and blood-stasis syndrome of chronic heart failure based on data mining technology," *China Journal of Chinese Materia Medica*, vol. 42, no. 1, pp. 182–186, 2017.

## Biography

**Jinming Zhang** born in Yantai, male, from Yantai, Shandong, China, has gained the master's degree. He is now working in Yantai vocational college. He is an associate professor and senior engineer. He is interested in computer network technology, information security technology and cloud computing technology.