ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 132

2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2017)

The Application of High Dimensional Data Mining Based on Big Data to Intrusion Detection

Jinhua Liu

Modern Educational Technology Centre,Xinyu University,Xinyu City,jiangxi,China

***Abstract*—Under the historical background of big data, the traditional data processing methods can no longer meet the requirement proposed by the intrusion detection of systems. With the current development of information technology, the data in systems become increasingly complex, and the workload of data mining also becomes increasingly large in the intrusion detection of systems, which invisibly increases the difficulty in intrusion detection. In this paper, issues related to the application of high dimensional data mining technology based on big data to intrusion detection are analyzed.**

Keywords-high dimensional data mining; big data; intrusion detection

I. INTRODUCTION

Big data refers to the data information that is difficult to be processed or analyzed with traditional methods and tools. It is not only the explosion of relevant information data, from the original ERP/CRM data to the Internet data and then to the sensors of the Internet of Things, but also the improvement of data complexity [1]. The high dimensional data mining technology based on big data can define every attribute of data as a dimension in the space, by which an n-dimensional space can be formed. In addition, the attribute value can be expressed by it with vector and then transformed into the matrix form, meaning that the information of every record will be expressed with matrix. In this paper, subject extraction is conducted on the high dimensional data based on big data, and the high dimensional data mining technology is applied to the intrusion detection.

II. BRIEF ANALYSIS ON INTRUSION DETECTION TECHNOLOGY

In reality, the intrusion detection technology refers to a new network security technology designed and employed based on intrusion detection systems. A set of software and hardware data processing structure is included in the intrusion detection technology, which not only can make up the deficiency of traditional firewall's safety protection in practical application, but also can provide corresponding intrusion detection means for the network systems that need to be protected to maintain their operation security. In the current era of big data, the data mining and application ability of people should be improved. However, due to the increase in data volume, some deficiencies emerge in traditional intrusion detection technologies. Therefore, the application value of using high dimensional data mining technology in intrusion detection will certainly be shown in the future.

III. DESIGN OPTIMIZATION OF INTRUSION DETECTION TECHNOLOGY BY USING THE HIGH DIMENSIONAL DATA MINING BASED ON BIG DATA

A. *Design Optimization of Intrusion Detection Model*

In the era of big data, the design of intrusion detection system can be optimized and the intrusion detection model can also be built by using the high dimensional data mining method. In this model, data collection engine is included, the main functions of which are dealing with the ambiguity of data semantics, solving the problem of data missing and cleaning data in intrusion detection. In the module of data selection, it can distinguish the data set that needs to be analyzed, which can reduce the range of data processing. Furthermore, in the module of data preprocessing, it can adopt the high dimensional data mining technology, which can effectively overcome the limitations in the application of current data mining tools and algorithms[2]. During this process, the data can be first processed simply, and then useful information can be further collected. By contrast, in the module of correlation analysis, it can extract the correlation features among the intrusion behaviors, so that data reference can be provided for the further excavation and detection of intrusion behaviors and the formulation of intrusion protection mechanism. Similarly, the reference information base is also included in the model, and it is mainly used for storing some feature data related to the users' common habitual behaviors. In this process, if intrusion behaviors are found, it can give an alarm and leave the evidence of intrusion, while it will continue the detection, if the user behavior is normal.

B. *Design of Algorithms*

In this study, design optimization is carried out with the high dimensional data mining technology based on big data. In terms of the intrusion detection technology, the nearest neighbor method is adopted, which is shown as follows:

If a surface scattered point set, P={Pi(xi, yi, zi), i=1,2 n}, is given and a certain point is assumed to be V(xv, yv, zv), then the k points that are the nearest to the point V in P can be viewed as the m neighboring point set of V, which is recorded as :MNB|V|={P1,P2, Pm} and named as the k-nearest neighbor of V. It reflects the local information of point V, and every point in the k-nearest neighbor is called the neighbor point of V.

The specific implementation of the algorithm is as follows:

1) Firstly: The eigenvector can be constructed from the message of http type. Then the original message data can be

obtained from the data collection engine, and some feature data that can be used to determine the intrusion behavior can be extracted and converted to vectors.

2) *Secondly:* The similarity between the records of the eigenvector in historical and practical behavior base can be figured out. Every historical record in the historical behavior base of intrusion detection system can be assumed as an n-dimensional vector. It means that every record can be seen as a point in the space of the n-dimensional vector. The eigenvector to be detected can be denoted by x, the behavior center vector of the ith type in historical behavior by di, the dimensions in the eigenvector by nj, the j-dimension weight by wj, while the jth dimension in x and di by xj and dij, respectively. The calculation formula is shown as follows:

$$S\text{imilarity}(\text{x}, d_i) = \sqrt{\sum_{j=2}^{n} w_j\,(x_j - d_{ij})^2}$$

3) *Finally:* The types of user behavior in intrusion detection can be judged. Specifically, K records of “a” that are the most similar to the eigenvector x, i (neighbor), can be selected from the training records according to the similarity figured out in step b. Then the weight of x in every type of behavior can be tested successively. The specific calculation formula is given as follows:

$$\text{p(i, j)} = \begin{cases} if \sum\limits_{a_i \in K-NN} Similarity(x, a_i)\ \ y(a_i\,, c_i) - b \le 0 \\ \qquad\qquad 0 \qquad others \end{cases}$$

In practical data mining, the high dimensional data mining technology is able to detect the severity of intrusion behavior and provide corresponding judging result according to the type of user behavior as well as the weight of intrusion behavior. Then it can determine the types of alarm, attack, record and safety in the judging result and work out relevant intrusion prevention measures for the system.

IV. Specific Experimental Application Analysis

A. Training on Data Mining Engine

In practical application, the data mining engine needs to be trained before the experiment of data intrusion detection. This means that some known attack modes and their corresponding records can be first marked as corresponding types in the database of the system’s historical behaviors, and then the mining engine of high dimensional data can be trained. In this way, it can be made sure that the data mining engine based on big data will be enabled to have certain ability of effectively judging unknown attacks as well as self-learning and self-development ability of improving its detection capacity.

For example, in the training module, on the basis of data mining, a high-dimensional data mining engine is established to acquire and learn knowledge. Attack behavior characteristics are extracted from a large number of attack data [5]. The training steps are as follows:

Firstly, the training text set is vector quantization and the set of features is obtained.

Secondly, with the feature subset extraction algorithm, an optimal feature subset is extracted from the feature set (the evaluation algorithm is used to determine the "optimal" subset).

Then, the training text represented by the feature subset is classified according to the classifier, and the performance of the classification feature subset is evaluated.

Finally, in the classification module, the test text is represented by the optimal feature subset, then the classifier is used to classify, and the class with the highest posterior probability is taken. The unknown sample category is predicted by using the given training sample set [6].

B. Training Data Source

According to the evaluation data of intrusion detection system provided by the United States National Defense in 1998, the experimental result shown in Table 1 is obtained by analyzing its Snort detection result after it has received the data training of about 10MB. In this table, the rows of Snort that denote the attack types represent the test result obtained by using the Snort software in which the data mining module has not been added, while the rows of DM represent the test result after the data mining module has been added in the Snort software.

TABLE I. RESULT TABLE

Attack types	***Normal events P times***	***Times of detection***	***Abnormal events P times***	***Times of false report***	***Times of failure in report***	***Accuracy of detection P%***
Impersonation attack (Snort)	400	145	200	17	55	72.5
Legal user attack (Snort)	400	157	200	10	43	78.5
Impersonation attack (DM)	400	159	200	17	41	79.5
Legal user attack (DM)	400	182	200	16	18	91.0

Through the analysis on the experimental data in Table 1, it can be found that in high dimensional data mining based on big data, the times of failure in reporting impersonation attack and legal user attack can be reduced, which are decreased from 55 and 43 times to 41 and 18 times, respectively. At the same time, the accuracies of detecting intrusion attack are increased from 72.5% and 78.5% to 79.5% and 91.0, respectively. Therefore, the experimental result proves that the practical application of high dimensional data mining technology based on big data to intrusion detection can improve the accuracy of intrusion detection and increase the positive influence of intrusion

detection system on detecting illegal attack.

C. High-dimensional Data Mining Technology Used in Various Fields

In the fields of financial securities, e-commerce, teaching quality control and e-government based on Internet applications, high-dimensional data mining technology is used to analyze intrusion detection information from the massive data resources accumulated in these areas [7]. At the same time, in practice, we can also use high-dimensional data mining technology to detect network security intrusion in various fields such as financial securities, electronic commerce, etc., and deeply analyze and monitor the Internet operation status in these fields [8], to avoid potential security threats in these application areas, to improve the effectiveness of network security intrusion detection and to play an active application value.

In practical teaching, the use of high-dimensional data mining technology can monitor the quality of teaching from different dimensions (score distribution, typical correlation) , analyze performance and statistical relationship between the marks and learning environment, and dig hidden pattern of the data, so as to monitor the dynamic teaching adequately and to create a comfortable learning environment for students[9]. Moreover, the use of high-dimensional data mining technology can also help improve the teaching curriculum decision-making and adjust the teaching program to enhance the overall quality of teaching combined with the actual situation.

V. CONCLUSION

In conclusion, the application performance of intrusion detection system can not only be improved, its positive application advantage also can be given into full play by using the high dimensional data mining technology based on big data that has self-learning and self-development ability in intrusion detection system to extract the features of aggressive behaviors from a large number of attack data.

REFERENCES

[1] Ouyang Kaicui, Zeng Linghua. An Intrusion Detection Scheme Based on High Dimensional Data Mining [J]. Journal of Chongqing University of Science and Technology (Natural Science), 2005,7(4):61-64.

[2] Meng Yuyu, Sun Chuanqing, Zheng Liying, et al. The Application and Research of Data Mining Technology in Intrusion Detection [J]. Automation and Instrumentation, 2012,(3):1-3,12.

[3] Guo Chun. Research on Key Technologies of Network Intrusion Detection Based on Data Mining [D]. Beijing University of Posts and Telecommunications, 2014.

[4] QIU Hui-ying.An intrusion detection method based on principal component analysis and fuzzy clustering [J]. Chinese Science Bulletin, 2012,28 (12): 51-53.

[5] WANG Qian, TANG Rui.Application of outlier detection based on frequent pattern in intrusion detection [J] .Application Research of Computers, 2013,30 (4): 1208-1211.

[6] Liu Peiqi, Sun Jing, Duan Zhongxing et al .. Research on Outlier Detection Algorithm in High Dimensional Space [J] .Microelectronics and Computer, 2013, (7): 68-71,77.

[7] DONG Fei.Study on the improvement of outlier mining method for high dimensional data [J]. Journal of Computer CD Software and Application, 2013, (4): 108.

[8] YANG Feng-zhao.Study on frequent closed-pattern mining algorithms for high-dimensional data [J] .Application of Computer System, 2011,20 (11): 231-235.

[9] Zhao Peng.Study on frequent item set mining algorithm under massive high dimensional data [J].Journal of Computer Applications and Software, 2012,29 (7): 150-153.

[10] Zhang Haitao, Huang Huihui, Xu Liang et al.Progress in privacy protection data mining [J] .Application Research of Computers, 2013,30 (12): 3529-3535.