Acta Technica *62* (2017), No. 4A, 635–640

Research on outlier intrusion detection technologybased on data mining

Liang zhu[1,2]

Abstract. With the rapid development of information technology, network security issues become increasingly prominent. Intrusion detection technology is a detection technology for all kinds of network attacks. The intrusion detection system needs to adapt to high-dimensional and massive network traffic. In order to solve this problem, this paper proposes an outlier mining algorithm based on attribute correlation and outlier probability to detect intrusion behavior, and improve the performance of intrusion detection system by combining data mining technology with intrusion detection technology. Finally, the experiment shows that the algorithm can effectively find the intrusion behavior, and improve the performance of intrusion detection system.

Key words. Data mining, outlier, intrusion detection.

1. Introduction

With the rapid development of computer network technology, computer network is widely used in various industries and fields. The computer brings great convenience to people's production and life. However, because the computer network has information sharing, terminal distribution and network open and other functional characteristics, there are many security risks. Therefore, the research on network information security technology has become a very important research topic of information security technologies in the fields such as computer communication.

2. Basic principle of outlier detection

2.1. Outlier mining technology

Data mining[1] refers to the non-trivial process of discovering data that is not known and inherently useful from large amounts of data in the database. Among

[1]Workshop 1 - College of Computer Science, Chongqing University, Chongqing400044,China
[2]Workshop 2 - School of Information Engineering, ChongqingIndustry Polytechnic College, Chongqing401120,China; e-mail: jsly2001@sina.com

them, the outlier mining technology is a key mining technology in the data mining industry, which is included in the basic requirements of data mining. The main task requirement is to discover unique data information in the specified data set. The unique data here is outlier data. Outlier [2−3] is an object that obviously deviates from the observed value of most objects, which is thought to be generated by a different mechanism.

2.2. Intrusion detection technology

Intrusion detection [4] is a technology developed in recent years. It is a security mechanism that can dynamically monitor, predict and defend against system intrusion. Through the dynamic monitoringof the operation of the computer network, system architecture and other aspects to identify the means and purpose of intrusion. Intrusion detection has a variety of features, including real-time monitoring, intelligent control, dynamic response and configuration convenience. The main goal is through dealing with the network behavior characteristics or the system log to distinguish the user behavior those who violate the security policy or threat to the system security, so as to ensure that the system resources are not abused, to prevent the omission of system data, being artificially modified and wantonly destructed.

Intrusion Detection System [5] (IDS) takes intrusion detection as a new security protection, based on the overall characteristics of the target, the system objectives, audit documents, abnormal behavior and activity records. There are several different classification methods for intrusion detection systems from different perspectives.

3. Application and Implementation of Outlier Detection Technology

3.1. Algorithm of outlier intrusion detection technology

An outlier mining algorithm based on outliers [6−7] uses an outlier to represent a data point, and the outlier probability is a number from 0 to 1. For any data set, the maximum and minimum values of the outlier are fixed.

The algorithm first analyzes the attribute relevance of the data set and obtains the relevant information of the attribute set. And then the attribute reduction is carried out according to the relevant information of the attribute set, and the optimal attribute subset which can keep the original information of the data set as much as possible is obtained. Finally, the outlier probability of the data set is calculated on the optimal attribute subset. Select the outlier data according to outlier probability, that is, the final result required. The flow chart of the implementation of specific algorithms is shown in Figure 1.

It can be seen that the algorithm optimizes the attribute set by attribute reduction, and obtains the optimal attribute subset. Therefore, the algorithm is suitable for high-dimensional data sets. In addition, the algorithm uses the outlier probability to select the outlier data, which makes the algorithm keep a uniform standard on different data sets, so the algorithm can adapt well to the vast majority of data

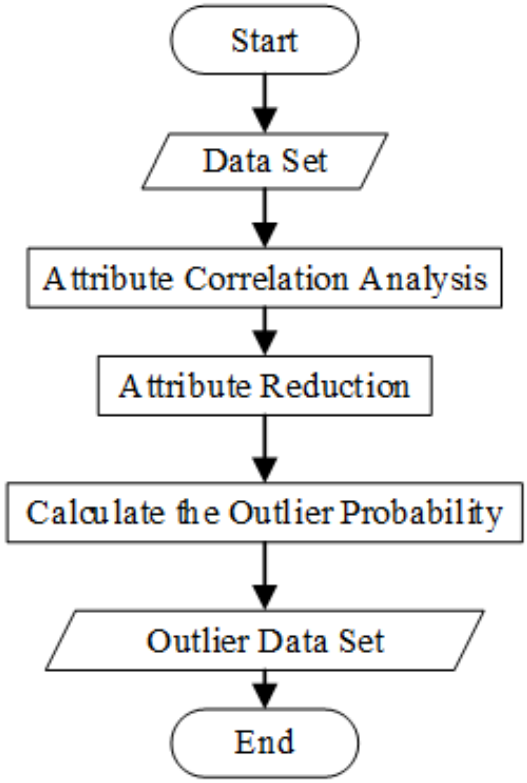

Fig. 1. Flow chart of algorithm implementation

sets.

For a given attribute $A_i A_j A$ the attribute relevance betweenA_i and $A_j$are described through the formula (1):

$$\gamma\ A_i A_j = \int_{k-1}^{N} \frac{P_k}{N} \tag{1}$$

Wherein A is the set of attributes, N is the number of rows of matrix Z n*d . When there is zki=zkj in matrix Z n*d , Pk = 0, otherwise Pk = 1. It can be seen that the higher the γ (Ai, Aj) value, the stronger the correlation between Ai and Aj.

For a given threshold ε, if the correlation between any attribute in the matrix Z (n*d) and the actual data is greater than a given threshold value ε, it is assumed that the attribute is redundant and needs to be deleted from the attribute set. After using the above inference for each attribute in the attribute set, the attribute reduction of the algorithm can get the optimal attribute subset.

$$b_{pq} = \frac{a_{qp}}{\int_{k=1}^{n} a_{qp}} \tag{2}$$

Therefore, the probability that the data point xp belongs to the outlier dataset C0 is expressed as follows:

$$\mathrm{P}\ x_p \in C_0 = \int_{qp} 1 - b_{qp} \tag{3}$$

After adding the weight, the algorithm can more accurately calculate the dissimilarity between two data points. Because in the calculation of the dissimilarity of two data points the different attributes will account for different proportions according to their different importance. The attribute with a larger effect on the dissimilarity values between data points will account for a larger proportion when calculating dissimilarity, whereas the attribute with a smaller effect on dissimilarity between data points will take less when calculating dissimilarity proportion.

3.2. Algorithm Implementation and Feasibility Analysis

In order to evaluate the intrusion detection system, two common performance parameters in the intrusion detection system are discussed: data error rate and data leakage rate. The data error rate corresponds to the case where the normal data is seen as outlier data in the outlier mining algorithm, and the data omission rate corresponds to the outlier data hidden in the normal data in the outlier mining algorithm. Therefore, when the outlier mining algorithm is applied to the intrusion detection system, the ratio of these two cases needs to be considered.

In addition, the first two stages of the algorithm mainly deal with the attribute set of the data, analyze the relevance of the data set, reduce the attribute set, and delete the redundant attributes. The intrusion detection system in the data collection of many attributes, attribute set has a lot of redundant attributes. After the first two stages of the algorithm, the data stream will preserve the main attributes, which is very beneficial to the detection of intrusion data. At the same time, because the attribute reduction of the algorithm does not need to be running all the time, the algorithm can run the first two stages by extracting small sample, and then run the third stage in practice to complete the intrusion data detection. This will be very effective for intrusion detection systems which has a high requirement on real-time performance.

4. Experimental testing and results analysis

4.1. Design and implementation process of the experiment

The attributes in the data set are of both numeric and parameter symbol types. For the attribute in the parameter symbol type, this paper first carries on the numerical correspondence to the data of the parameter symbol type, then carries on the standardized calculation to the mapped data. For attributes that are of continuous numeric type, discrete analysis is needed. In addition, the measurement units of the selected attributes are different, and the results are very obvious in the processing of

the data. In order to facilitate the analysis, this paper uses the normalized interval to measure the difference between the different units.

4.2. Analysis of experimental results

In verification experiment of algorithm in the intrusion detection, two indicators including data detection rate (Detection Rate) and data false rate (False Rate) are selected to represent the efficiency of the implementation of the algorithm. Compare other algorithms to verify the efficiency of the algorithm. The data detection rate is the ratio of the number of true outliers detected by the algorithm to the total number of outliers in the data set, ie the data detection rate (DR) = number of true outliers detected (D) / total number of outliers (N). Data false rate (FR) is the ratio of the number of the normal data which is mistaken as outliers to number of the normal data in the data set, ie the data error rate (FR) = number of the normal data which is mistaken as outliers (F) / number of the normal data (N) * 100%.

Table 1. Experimental results of several algorithms

Algorithm Name	DR %	FR %
Algorithm of This Paper	96.2	4.57
EOS	95.1	3.96
SPOD	94.3	4.21
OMBGI	95.6	4.02
ENBROD	93.9	3.92

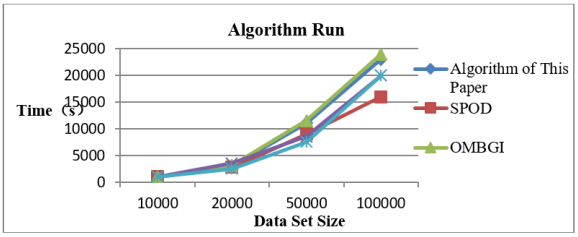

Fig. 2. Algorithm running time comparison diagram

It can be seen from the figure that when the data set is small, the running time of each algorithm is negligible. With the increase of the scale of the data set, the running time of the algorithm is also increasing, and the increasing trend is also increasing. The algorithm proposed in this paper can meet the requirements of intrusion detection and can be effectively applied to the actual intrusion detection system to analyze and process the network data.

5. Conclusion

This paper analyzes the outlier mining technology and intrusion detection technology. It is considered that the data to be processed by the current intrusion detection system is massive and of high dimension with the increasing of the amount of network dat. Therefore, this paper proposes an outlier data mining algorithm based on attribute correlation and outlier probability. The three steps of algorithm implementation are described in detail, and the application and feasibility of the algorithm in intrusion detection system are analyzed.

References

[1] G. H. Orair, D. C. Gupta, W. Meira: *Distance-based outlier detection: consolidation and renewed bearing.* Proceedings of the Vldb Endowment *3* (2010), Nos. 1–2, 1469–1480.

[2] J. S. Tomar, A. K. Gupta: *Survey on Outlier Detection in Data Stream.* International Journal of Computer Applications *98* (1985), No. 2, 257–262.

[3] R. H. Gutierrez, P. A. A. Laura: *Online bad data detection using kernel density estimation.* Power & Energy Society General Meeting. IEEE *18* (1985), No. 3, 171–180.

[4] R. P. Singh, S. K. Jain: *A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data.* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining *7* (2004), No. 1, 41–52.

[5] M. N. Gaikwad, K. C. Deshmukh: *TExploiting Active Sub-areas for Multi-copy Routing in VDTNs,* . Applied Mathematical Modelling *29* (2005), No. 9, 797–804.

[6] S. Chakraverty, R. Jindal, V. K. Agarwal: *Exploiting Active Sub-areas for Multi-copy Routing in VDTNs.* Indian Journal of Engineering and Materials Sciences *12* (2005) 521–528.

[7] N. L. Khobragade, K. C. Deshmukh: *A Time-Efficient Connected Densest Subgraph Discovery Algorithm for Big Data.* Sadhana *30* (2005), No. 4, 555–563.

[8] Y. F. Zhou, Z. M. Wang: *Fast and Scalable Outlier Detection with Approximate Nearest Neighbor Ensembles.* J Sound and Vibration *316* (2008), Nos. 1–5, 198–210.

Received November 16, 2016