A Method for Detecting Large-scale Network Anomaly Behavior

Huimin Hu[1,], *Wenping* Ma[1], and *Wei* Luo[1]

[1]State Key Laboratory of Integrated Services Networks**,** Xidian University, Xi'an, China

Abstract. A clustering model identification method based on the statistics has been proposed to improve the ability to detect scale anomaly behavior of the traditional anomaly detection technology. By analyzing the distribution of the distance between each clustering objects and clustering center to identify anomaly behavior. It ensures scale abnormal behavior identification while keeping the processing mechanism of the traditional anomaly detection technology for isolation, and breaking through the limitation of the traditional anomaly detection method assumes that abnormal data is the isolation. In order to improve the precision of clustering, we correct the Euclidean distance with the entropy value method to weight the attribute of the data, it optimizes the similarity evaluating electric of the nearest neighbor clustering algorithm, and simulated. Experimental results show that the statistical method and the improved clustering method is more efficient and self-adaptive.

1 Introduction

With the rapid development of the Internet and the information technology, the network application and information sharing are becoming more and more extensive. In the face of the increasingly complex network environment, the static defense which relies solely on the firewall is not enough to deal with all kinds of attacks, but the network intrusion detection technology with active monitoring system can effectively make up the shortage of the firewall, which can be divided into host-based system and network-based anomaly detection system according to the different monitoring objects. The host-based system anomaly detection detects the abnormal behavior of the network by monitoring and analyzing the host system log file, but it is poor interaction. At present, the anomaly detection technology based on network behavior is a research hotspot of the network-based anomaly detection. Judging whether the data flow in the network is abnormal by data mining technology, that is, the clustering is regard as normal behavior by clustering algorithm and the isolated is abnormal, and the detection performance is obviously superior to the host-based intrusion detection system. However, due to the rapid expansion of the network scale, various newly and large-scale attacking methods are endless, and it has obvious defects in flexibility, adaptability and so on. They cannot to deal with the changing network intrusion, so it is an important challenge of how to effectively detect the new large-scale attack.

· Corresponding author: 1304237049@qq.com

There have been many scholars have made improvements to this, the literature [1-3] are based on data mining network intrusion detection technology, by reducing data used in the network and the dimension of attributes to raise the accuracy of clustering to improve the efficiency of network intrusion detection through the clustering process, but the algorithm focuses on the clustering data itself, ignoring the clustering. In the literature [4-5], they detect the anomaly behavior of the online network by constantly updating the behavior pattern library, but the method of distinguishing between normal and abnormal behavior is not given in essence. In literature [6], the anomaly behavior is similar to that of isolated points, ignoring the case of scale anomaly behavior would be clustered. The improvements above to the network anomaly detection are essentially based on the fact that the abnormal behavior is very small compared with the normal behavior, and the two are very far apart, that is, assuming that the abnormal behavior of the data is the isolated point. But the actual network anomaly behavior is often a large number of occurrences by camouflage, and close to the normal data cluster. Using the improved method above to detect the anomaly behavior effect is not ideal, and the anomaly behavior clusters are regarded as normal data and cause attacks successful invasion.

In the information theory, information entropy is used to measure the degree of chaos of information, and the concepts used in literature [7] solve multi-decision problems. This article uses it to measure the degree of discretization of individual attributes. In the clustering algorithm, the larger the degree of discretization is, the smaller the information entropy is, the greater the amount of information it provided, and the larger the weight. Otherwise, contrary to the above; based on the entropy method, the nearest neighbor clustering algorithm is improved, and the entropy method is used to evaluate the attributes of the objects. The weights of each attribute are calculated to measure the importance of the attributes. Euclidean distance is modified to make the clustering results more objective. The clustering algorithm completes the clustering of data, but the clustering does not indicate which clusters are normal behavior and which are clustering of abnormal behavior.

In order to solve the problem that the improved algorithm is not effective for large-scale abnormal behavior detection, this paper proposes a clustering model identification pattern method, which combines the statistical model with the traditional cluster-based anomaly detection technology. The distribution of the euclidean distance is analyzed to distinguish the normal and abnormal data clusters, and preserving the large-scale abnormal behavior recognition while keeping the processing mechanism the traditional anomaly detection of isolated points as abnormal data. Analyzing how to apply it to abnormal behavior and the improvement of the clustering method .In the end, it was verified by experiments.

2 Algorithm analysis

For the nearest neighbor clustering algorithm, the number of clusters does not need to be given in advance which can be dynamically generated in the process of clustering, with great flexibility and adaptability. The accuracy of clustering algorithm directly affects the effect of network anomaly detection. Considering the contribution of each attribute of data object generated by network behavior to clustering, an attribute weighting method based on entropy method is proposed. The attribute weights are determined objectively according to the characteristics of each object's attribute itself to improve the accuracy of clustering. Based on the statistical class cluster **pattern recognition** method to identify the cluster model and to detect scale anomaly behavior.

The traditional nearest neighbor clustering algorithm is only suitable for dealing with numerical attributes. For the nominal attribute, the frequency of each state of the cluster attribute is taken as the corresponding attribute value. To analyze the algorithm, we assume that there is a data set which have m objects, each data object is characterized by n

attribute values, which constitute the data matrix X, each row corresponds to an object, which is the j dimension of the i data object, $i=1,2,3,...,m$, $j=1,2,3,...,n$. The distance between the two objects is the matrix $D(i,j)$, the larger the $d(i,j)$, the smaller similarity between the two objects. So if $d(i,j)=0 \,\&\&\, d(i,j)=d(j,i)$,the matrix is symmetrical. As follows,

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & x_{ij} & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tag{1}$$

$$\boldsymbol{D(i,j)} = \begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(m,1) & d(m,2) & \cdots & 0 \end{pmatrix} \tag{2}$$

In order to improve the stability and accuracy of clustering and reduce the error caused by the existence of extreme values, the maximum and minimum values of each column in the data matrix X are set to the average of the remaining $m-2$ values.that is,

$$x_{ak\,max} = x_{bk\,min} = \sum_{j(j\neq a,\,b)}^{m} x_{jk} \Big/ (m-2) \tag{3}$$

Which $x_{ak\,max}$ and $x_{bk\,min}$ are the maximum and minimum values of the columns $a,b\in[1,m]$.

2.1 Improved nearest neighbor clustering algorithm by entropy method

2.1.1 Calculate the weight of each attribute by entropy method

The entropy method determines the weight of each attribute according to the degree of difference between the attribute values of the characterization object, which is determined by the data object itself avoiding the influence caused by the human factors, making the objectivity more accurate and the clustering result more accurate. The calculation steps are as follows.

1. Normalized the attribute value

The similarity between the calculation object and the center of each cluster is the key of the clustering algorithm. The difference of the attribute units of the object will directly affect the clustering result. The attribute of the larger initial range is relatively large. The clustering object attributes must be normalized before compiling the similarity between objects. In order to maintain the correlation between the original data, we use the minimum-maximum normalization method to linearly transform it. For the j dimension attribute, we use the following formula to calculate it,

$$x_{ij}^{'} = n\_min_j + \frac{(x_{ij} - x_{j\,min})}{x_{j\,max} - x_{j\,min}}(n\_max_j - n\_min_j) \tag{4}$$

The attribute values x_{ij} are mapped to $[n\_max_j, n\_min_j]$, and the attribute values are normalized.

2. *Calculate the proportion $x_{ij}^{'}$ of the j attribute value of object i*

$$p_{ij} = x_{ij}^{'} \Big/ \sum_{i=1}^{m} x_{ij}^{'} \tag{5}$$

3. *Calculate the j dimensional attributes of the entropy*

$$e_j = -\sum_{i=1}^{m} p_{ij} \ln p_{ij} \Big/ \ln n \tag{6}$$

4. *Calculate the weight of each dimension*

$$w_j = (1 - e_j) \Big/ \sum_{j=1}^{n} (1 - e_j) \tag{7}$$

That, $\sum_{j=1}^{n} w_j = 1, 0 \le w_j \le 1.$

2.1.2 Similarity measure

In the nearest neighbor clustering algorithm, we need to calculate the similarity between objects. In this paper, Euclidean distance is used as the basis of similarity measure [8], Euclidean distance formula is

$$d(i,j) = \sqrt{\sum_{k=1}^{m} \left(x_{ik}^{'} - x_{jk}^{'}\right)^2} \tag{8}$$

Where the k attribute values of the data objects i and j are the sum of all the attribute distances between the data objects i and j. The formula (1-6) ignores the differences in the proportion of each attribute in the clustering algorithm, which leads to the difficulty of clustering. In this paper, we use the entropy method to calculate the weight of each attribute of the object, the weighted Euclidean distance is used as the basis for the similarity measure, and the euclidean distance formula after the weight correction is

$$d_w(i,j) = \sqrt{\sum_{k=1}^{m} w_k \left(x_{ik}^{'} - x_{jk}^{'}\right)^2} \tag{9}$$

Which w_k is the weight of the attribute calculated by (1-5). The value of the weights reflects the role of the attributes of the objects while clustering, so that the measurement of the distance between objects is more accurate and the nearest neighbor.

2.2 Cluster-based pattern recognition based on statistical methods

The traditional network anomaly behavior detection assumes that the abnormal behavior is a small amount and is very different from the normal behavior data. It can find the isolated point by cluster analysis to detect the occurrence of the anomaly, but does not recognize the clustering model. The number in real network of the normal and abnormal behavior cannot be predicted. In the real network environment, if the attacker send a large number of intrusion data with high similarity to normal data, the detection is low efficiency, or even failure.

To break the limitation of the traditional network anomaly detection, we detect the abnormal behavior by detecting the pattern of the cluster of the number of balanced between the abnormal and the normal behavior in the real network environment and preventing the large-scale abnormal data object invasion. Compared with the traditional, we add the clustering pattern recognition mechanism of the statistical, and the isolated point is the same as the traditional anomaly detection, which is regarded as abnormal data. In this paper, we assume the normal behavior is independent and random, and the large-scale anomaly data generated by the intruder does not meet the assumption of independent randomness [9]. In probability theory, the central limit theorem (CLT) establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. In the clustering algorithm, the Euclidean distance is the sum of the distances between the various attributes of the object. In the case of large sample, theoretically approximating the normal distribution, we confirm the conclusion by Matlab simulation experiment.

We randomly generate the matrix with $1000*12$ (1000 objects of 12 attributes), and select an object as a central point, then calculate the distance d of the remaining objects to the center and the distance d_w using the entropy method. Figure 1 is the distribution of the characteristics d and d_w. This result verifies that both are similar to the normal distribution.

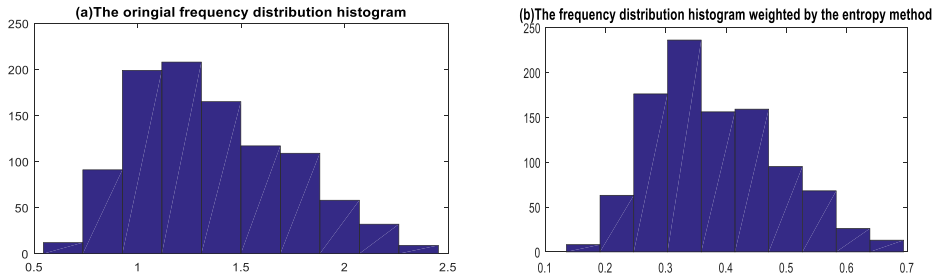

Fig. 1. The frequency distribution histogram of d and d_w

Based on the above characteristics, the clusters of network behavior are detected by statistical distribution, the normal behavior cluster and abnormal behavior cluster can be detected efficiently.

$$f(x,\mu_{dX},\sigma_{dX}) = \frac{1}{\sqrt{2\pi\sigma_{dX}^2}} exp\left(\frac{-(x-\mu_{dX})^2}{2\sigma_{dX}^2}\right) \tag{10}$$

Algorithm description: record the intermediate data generated by the nearest neighbor clustering algorithm: the mean μ_{dX} and variance σ_{dX}^2 of $d_w(i,j)$, and analyze whether the Euclidean distance $d_w(i,j)$ of each cluster obeys the normal distribution of (1-8), if they obey ,the data produced for normal behavior, else the data generated by abnormal behavior. Since the statistical method does not need to know the prior knowledge about the abnormal behavior object, it can detect new intrusion and improve the detection rate and adaptability of intrusion detection.

3 Experiments and results analysis

To verify the effectiveness of the improved algorithm and the detection effect of the abnormal data object, this paper uses Matlab programming to simulate and compare the experiment from the following two aspects:(1) the clustering effect of the improved algorithm, the higherthe detection rate of the algorithm, the better the clustering; (2) cluster pattern recognition effect, and simulating all kinds of cluster distribution to judge whether the various types of cluster model is correct intuitively.

3.1 Data

The experiment using the "corrected.gz" data set in the KDD CUP99 [10] , select three classes of neptune, smurf, mailbomb, each class has 2000 anomaly data objects.

Table 1 List of attributes and weights

num	attribute	Attribute description	weight
1	duration	Connection duration (seconds)	0.0872
2	Protocol_type	Protocol Type	0.0602
3	service	The type of network service of the target host	0.0405
4	flag	Connect to normal or wrong state	0.0914
5	Wrong_fragment	The number of pieces of error	0.1023
6	Num_failed_logins	The number of failed attempts to log in	0.1077
7	Num_file_creation	The number of times to create file	0.0906
8	count	The number of connections with the same destination host is connected to the current connection Over the past two seconds	0.0724
9	srv_count	The number of connections with the same service as the current connection Over the past two seconds	0.0917
10	serror_rate	The same as the current connection host and the percentage of SYN error connections Over the past two seconds	0.0781
11	srv_serror_rate	The same as the current connection host and the percentage of SYN error connections Over the past two seconds	0.0872
12	diff_srv_rate	The same percentage of the current connection to the target host but the different connections Over the past two seconds	0.0907

For the normal data, randomly extract 2000 packets by wireshark. There are 8,000 objects as experimental data set. Each data object in the data set is characterized by several attributes. Because the properties of the object are too complicated to concisely depict the behavior of the network. It is not ideal to use all the attribute clustering directly. ①The increase of dimension will leads to the difficulty of processing the data of clustering algorithm. ②Some properties are very small and even affect the clustering effect. Therefore, the choice of attributes directly affect the effect of clustering, so we must select a set that can distinguish normal and abnormal behavior efficiently. In this paper, we use feature selection method to select the attributes of clustering [11]. Table 1 lists the 12 feature attributes selected for each object and their corresponding weights.

3.2 Results analysis

Next, compare the detection results of the nearest neighbor clustering improved by entropy method with the original method by matlab programming. And verification the result through the statistical clustering model.

3.2.1 The detection rate of the nearest neighbour clustering improved by entropy method

The clustering analysis of the same experimental data set is respectively carried out by using the traditional nearest neighbor clustering algorithm and the improved nearest neighbor clustering algorithm. The clustering algorithm is estimated by comparing the detection rate of the two experimental methods. And the detection rate is defined as

$$\text{Detection rate}=\frac{\text{The correct number of objects}}{\text{The total number of objects}}*100\% \tag{11}$$

Table 2. Comparison of two algorithms

Num	Data Type	Traditional algorithm detection rate	Improved algorithm detection rate	Increment
1	neptune	83.94%	88.32%	4.38%
2	smurf	82.35%	86.76%	4.41%
3	mailbomb	76.31%	78.95%	2.64%
4	normal	89.90%	93.94%	4.04%

From the table above we can know that the data is clustered into four clusters through the clustering algorithm, which coincides with the experimental data set. Compared with the traditional nearest neighbor clustering algorithm, the improved clustering algorithm is better. The accuracy of the nearest neighbor clustering algorithm is improved by 3.87% and treat the isolated points as abnormal data according to the traditional clustering algorithm.

3.2.2 Cluster model detection based on statistics

Figure 2 is the sum of the attribute similarity that the improved frequency distribution of the distance $d_w(i,j)$ between the clusters and the centers. The clusters can be detected to be normal and abnormal efficiently by statistical analysis.

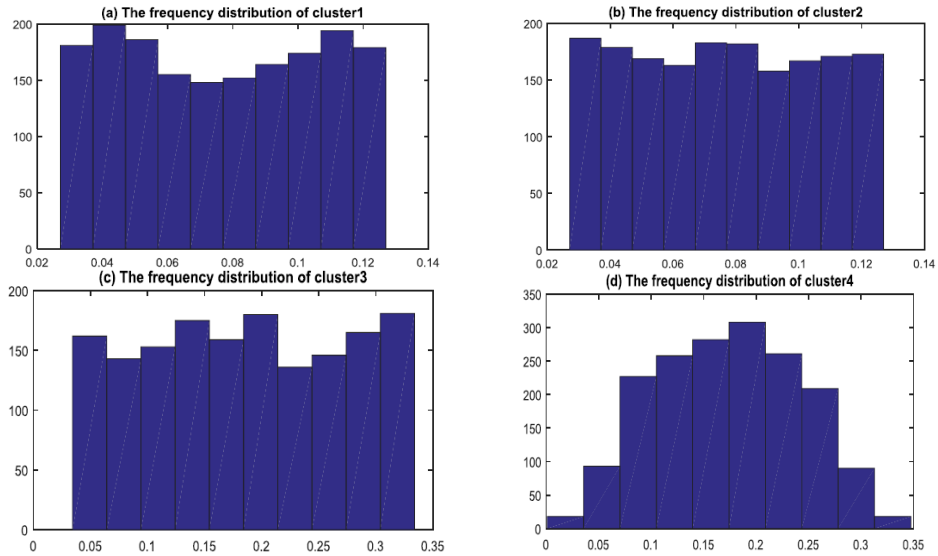

Fig. 2. The frequency distribution histogram of $d_w(i,j)$

It can be seen from the experimental simulation that cluster4 approximates the normal distribution and is normal. The other three categories are not the abnormal. It is consistent with the dataset that cluster1, cluster2, cluster3 for the neptune, smurf, mailbomb data clusters, cluster4 for the normal data cluster. There are two obvious advantages of class clustering pattern recognition based on statistics, ①the complexity of the algorithm is low, ②the normal and abnormal behaviors are identified from the nature and the detection ability of network anomaly detection in large-scale abnormal behavior is improved. It is proved that the statistical detection method is reliable and efficient for cluster detection.

4 Summary

The author studies how to improve the accuracy of clustering and the detection efficiency of large-scale abnormal behavior based on the traditional method of network anomaly detection and proposed that the weights of clustering data object attributes are set by entropy method and clustering pattern recognition method based on statistical analysis. Using the entropy method to calculate the attribute weights and modify the similarity measure formula of the nearest neighbor clustering algorithm which improved the objectivity of the attribute value and the accuracy greatly. For the clustering algorithm, the cluster-based pattern recognition mechanism based on the statistical method is added. The normal and abnormal behavior of the network is efficiently identified by analyzing the statistical distribution characteristics of the parameters. The isolation point is still think as abnormal data based on the traditional intrusion detection method to improve the reliability and applicability of intrusion detection and to overcome the low efficiency for large-scale network abnormal behavior detection, and achieved the expected aims.

Acknowledgement: This work was funded by National Key R&D Program of China under grant No. 2017YFB0802400, National Natural Science Foundation of China under grant No. 61373171 and 111 Project under grant No. B08038.

References

1. Guo Chun. Research on Key Technologies of Network Intrusion Detection Based on Data Mining [D]. Beijing University of Port and Telecommunications for the Degree of Doctor of Engineering.(2014).
2. Cui Wen Ke, Design and Implementation of Intrusion Detection System Based on Clustering Algorithm [D]. A Master Thesis Submitted to University of Electronic Science and Technology of China.(2016).
3. Sun Jian-hua ,Jin Hai,Chen Hao,Han Zong-fen , MA-IDS: A Distributed Intrusion Detection System Based on Data Mining[J]. WuHan University Journal Of Natural Sciences-english,(2005,10(1):111-111).
4. Juliette Dromard;Gilles Roudière;Philippe Owezarski .Online and Scalable Unsupervised Network Anomaly Detection Method[J]. IEEE Transactions on Network and Service Management,(2017,14(1):34-47).
5. Martin Grillab;Tomáš Pevnýab;Martin Rehakab, Reducing false positives of network anomaly detection by local adaptive multivariate smoothing[J]. Journal of Computer and System Sciences,(2017,83(1):43-57).
6. He Z, Xu X, Deng S. Discovering Cluster-Based Local Outliers [J]. Pattern Recognition Letters, (2003, 24(9):1641-1650).
7. Kordecki, H;Karmowski, M;Knapik-Kordecka, M;Karmowski, A;Gworys, B.Method of Component Importance Evaluation in Complex Data Structure Analysis[J].Advances in clinical and experimental medicine,(2011,20(2):205-209).
8. Nguyen T T, Chang K, Hui S C. Supervised Term Weighting Centroid-Based Classifiers for Text Categorization [J]. Knowledge and Information Systems, (2013,35(1): 61-85).
9. Liu Shuai.Research on Analysis and Detection of Multi-layer Abnormal Behavior for Network Data Stream[D].A Dissertation Engineering University for the Degree of Doctor of Engineering.(2015).
10. Hettich S, Bay S D. Kdd cup 1999 data. UCI KDD Archive [DB/OL], 1999. http://kdd.ics.uci. edu//databases/kddcup99/kddcup99. html].
11. Amiri F, Rezaei Yousefi M M, Lucas C, et al. Mutual Information-Based Feature Selection for Intrusion Detection Systems [J]. Journal of Network and Applications, (2011, 34(4):1184-1199).