View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Samara University

Security detection of network intrusion: application of cluster analysis method

W.H. Yang[1]
[1] Railway Signal and Information Engineering Department, Shandong Polytechnic,
Jinan, Shandong 250104, China

Abstract

In order to resist network malicious attacks, this paper briefly introduced the network intrusion detection model and K-means clustering analysis algorithm, improved them, and made a simulation analysis on two clustering analysis algorithms on MATLAB software. The results showed that the improved K-means algorithm could achieve central convergence faster in training, and the mean square deviation of clustering center was smaller than the traditional one in convergence. In the detection of normal and abnormal data, the improved K-means algorithm had higher accuracy and lower false alarm rate and missing report rate. In summary, the improved K-means algorithm can be applied to network intrusion detection.

Keywords: clustering analysis, K-means, cross entropy, network intrusion.

Citation: Yang WH. Security detection of network intrusion: application of cluster analysis method. Computer Optics 2020; 44(4): 660-664. DOI: 10.18287/2412-6179-CO-657.

Introduction

With the rapid popularization of the Internet and the rapid development of its technology, the Internet has penetrated into people's lives, not only providing convenience, but also adding entertainment means [1]. The Internet has developed rapidly because of the open platform, but it has a higher risk also because of the open platform [2]. By setting up a data monitoring system similar to the wall between multiple regional networks, traditional firewall technology [3] shields attacks from the outside world as far as possible to build a closed and secure computer environment. However, the development of network technology makes it difficult for passive and closed firewalls to provide secure network defense. Intrusion detection system monitors the real-time information flowing in the network, identifies malicious behaviord, and responds to it [4]. Ponomarev et al. [5] proposed a method to detect network intrusion data by measuring and verifying the data transmitted through the network rather than the data used in the transmission protocol network and found through simulation experiments that the accuracy of the method reached 94.3 %. Tao et al. [6] proposed a network intrusion detection algorithm combining spectral clustering with deep neural network algorithm and found through simulation experiments that the detection accuracy of this algorithm was better than that of algorithms such as Back Propagation (BP) neural network, support vector machine and random forest. Aiming at the optimization of point-to-point network intrusion detection, Wang [7] put forward an improved ART2 intrusion detection method; the simulation results showed that the improved algorithm had high detection rate and low error rate, which meet the needs of error detection and anomaly detection. This paper briefly introduced the network intrusion detection model and K-means clustering analysis algorithm and improved them. Then two clustering analysis algorithms were simulated and analyzed on MATLAB software.

Network intrusion detection

Fast and convenient Internet not only provides a convenient platform for the general public, but also opens to illegal elements, which makes the general users in the network bear the risk of network intrusion.

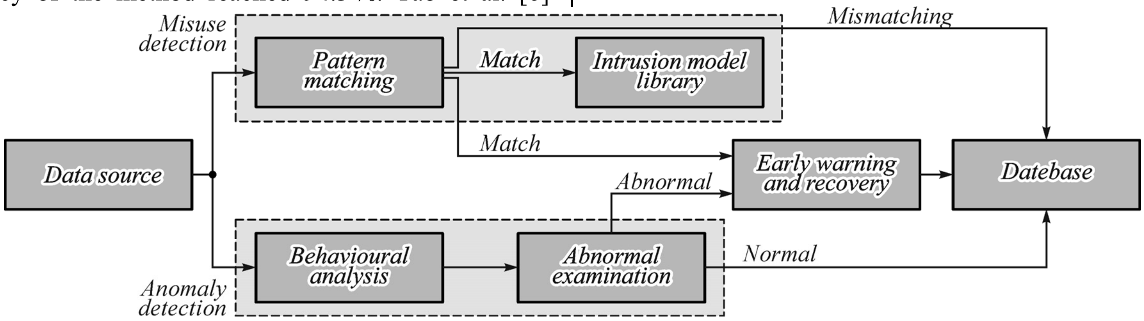

Fig. 1. General model of intrusion detection system

The general model of intrusion detection is shown in Fig. 1, which contains two main intrusion detection methods: misuse detection and anomaly detection [8]. The process of misuse detection in the model in Fig. 1 is mainly as follows. Firstly, data sources that need to be detected are collected, their features are extracted and transformed into patterns that can be identified in the intrusion model library, and they were compared with the intrusion feature patterns stored in the intrusion model library. If the similarity reaches a set threshold, a warning response is given until the exception is processed, and then the data is transmitted to the database. Moreover the intrusion mode is stored in the intrusion model library as the contrast object. If the similarity does not reach the set threshold, there is no abnormality, and the data are transmitted to the database. The process of anomaly detection in the model in Fig. 1 is mainly as follows. Firstly, the data source to be detected is collected, and its characteristics were extracted and compared with the characteristics of normal data, so as to train the detection system. Then the data to be detected and normal data were compared and analyzed to determine whether the data is abnormal or not. If the data is abnormal, the early warning response will be given until the abnormal processing is completed, and the data is transferred to the database; if normal, the data is transferred to the database.

The main functions of these two intrusion detection models [9] are: (1) to monitor the network environment in real time and active interception before the intrusion data takes effect; (2) to minimize the loss as far as possible when the intrusion data cannot be completely intercepted; (3) to record and analyze the intrusion data characteristics after intrusion for the next time of intrusion protection.

Clustering analysis algorithm

As shown above, in the intrusion detection system model, the detection part used for analyzing and judging intrusion data is the core function of the whole system model. Its performance determines the efficiency of the whole system in dealing with intrusion data. As far as intrusion detection system is concerned, it is faced with massive data in the Internet. Moreover, the diversity of data also determines that conventional judgment methods are difficult to adapt to intrusion data detection. In this study, K-means clustering algorithm [10] was applied for detecting the intrusion data in the Internet. The calculation procedures are as follows.

(1) Firstly, *K* clustering center generate randomly according to the input *K* value.

(2) Then the distance between the data and clustering center is calculated according to equation (1):

$$d(x,Z_j)=\sqrt{\sum_{i=1}^{n}(x_i-Z_{ji})^2}, \tag{1}$$

where $d(x,Z_j)$ is the distance between data x and clustering center Z_j, x_i is the i-th dimensional data of x, and Z_{ji} is the i-th dimensional data of Z_j. The distance between every data and clustering center is calculated. Then the data are allocated to the cluster that the clustering center belongs to according to the distance.

(3) After clustering, the clustering center of each kind of data set is recalculated, and then the second procedure is repeated for reclassification.

(4) Procedure (1), (2) and (3) are repeated until the clustering criterion function reached the preset standard, and its expression is:

$$\begin{cases} J(I)=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_i^j-Z_j(I)\right\|^2 \\ \left|J(I)-J(I-1)\right|<\xi, \end{cases} \tag{2}$$

where $J(I)$ refers to the square error of the I-th clustering result, x_i^j refers to the i-th data in j cluster, $Z_j(I)$ refers to the clustering center of the j-th cluster of the I-th clustering, and ξ is the threshold for determining whether the iteration stopped.

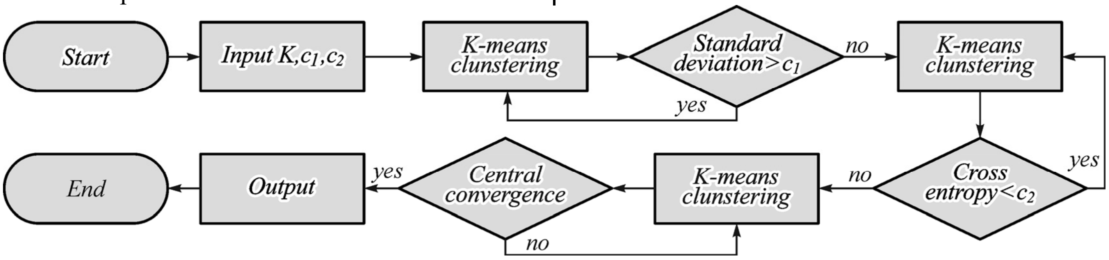

Fig. 2. The flow chart of improved K-means algorithm

The traditional K-means algorithm is simple in steps, but the accuracy of its retrieval depends on the *K* value. The selection of *K* value is often based on experience, and it is easy to fall into local minimum value in the process of classification, which affects the accuracy of classification. Therefore, in order to improve the accuracy and efficiency of forensics, the traditional algorithm is improved by adding standard deviation and cross entropy [11] steps in combination with the theory of hierarchical clustering. The flow of the improved K-means algorithm is shown in Fig. 2.

(1)K,$c1$,$c2$ and data to be classified are input, and $c1$ and $c2$ are the threshold value of standard deviation and cross entropy respectively.

(2) Then *K* clustering centers are selected from the data to be classified, and then the data are processed by clustering division using the principle of proximity according to clustering centers to obtain *K* categories of data.

(3) The standard deviation between data in every cluster is calculated after classification. The clustering center of the cluster whose standard deviation is not larger than c_1 remains unchanged. The clustering center of the cluster whose standard deviation was larger than c_1 needs to be reselected, and two points which are the nearest to the original clustering center are selected as the new clustering centers. Then the data were processed by clustering

division using the principle of proximity according to the newly selected clustering center. The above steps repeat until the standard deviations of all categories is not larger than c_1. That procedure is to split the original K categories to obtain suitable classification number.

(4) Aggregation of cluster number: Clustering division was performed using the principle of proximity according to the clustering center obtained after the final split at the last step. Then the cross entropy between any two clusters is calculated. When the cross entropy between two clusters is smaller than c_2, the two clusters are merged, and the clustering center after merge is calculated and taken as the new clustering center. Then the data are processed by clustering division using the principle of proximity according to the reselected clustering center. The above procedure repeats until the cross entropy between any two clusters is no smaller than c_2.

(5) K-means clustering division is performed according to the clustering center obtained from the final aggregation at the last step. The steps are the same with the traditional K-means algorithm. The results are output until center convergence.

The calculation formulas of standard deviation and cross entropy between clusters in the improved K-means algorithm are as follows:

standard deviation:

$$\theta = \sqrt{\frac{\sum_{i=1}^{m} d(a_i, C)}{m}}, \tag{3}$$

where θ is the standard deviation, C is the clustering center, m is the number of data in the cluster, and αi is the i-th data in the cluster.

cross entropy is:

$$\begin{cases} D_R = -\log\left(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} G(a_i - b_j, \sigma^2)\right), \\ G(a_i - b_j, \sigma^2) = e^{\left(-\|a-b\|^2\right)/2\sigma^2}, \end{cases} \tag{4}$$

where DR is the cross entropy, $G(\bullet)$is Gaussian kernel function, m is the number of all data in one cluster, n is the number of all data in the other cluster, α is the data set of one kind of clustering, b is the data set of another kind of clustering, and σ2 is the variance of Gaussian function.

The improved K-means algorithm is shown in Fig. 2. Firstly, K value and eigenvector data were input to calculate the center points, then K was divided into appropriate number of categories by standard deviation and aggregated by cross entropy, and finally they were classified by K-means algorithm until the center converges.

Simulation analysis

1. Experimental environment

The algorithm model was compiled using MATLAB simulation platform [12]. The experiment was carried out on a laboratory server. The server configurations were Windows 7 system, I7 processor and 16G memory.

2. Data preparation

KDD99 data set was used in this study [13]. Each data in the data set had 42 features. The first 41 features were the characteristic attributes of the data, and the last one was the decision attribute, indicating whether the data was abnormal, and it was used for detecting the performance of the algorithm.

The data set includes normal data and Dos, R2L, U2R and Robe intrusion data, which could simulate the real network environment. Four kinds of intrusion data are a big classification of the existing network intrusion data. Dos is a denial of service attack, which makes the object unable to process normal requests by occupying network computing resources; R2L attack can achieve no permission to enter the target server by some means; U2R attack takes ordinary privileges as springboards to steal high-level privileges; Probe attack collects system information about objects, similar to spy. 1000 normal data, 800 Dos data, 600 U2R data and 400 Probe data were randomly selected from KDD99 data set, and 30 % of them were used as training samples and 70 % as testing samples.

3. Experimental steps

(1) Data preprocessing: Among the 41 features of data in KDD99 data set, only 38 features were digital, and the remaining 3 features were characters, which could not be directly used for calculation. First, character features were converted into digital features, and finally 41 features were converted into 122 digital features. Then, in order to eliminate the problem of too large numerical span between different data, the average variance method [14] was used for standard operation of data:

$$y = (x - \bar{x})/x_{\text{var}}, \tag{5}$$

where y stands for standardized data, x is the data that needs standardization, $\bar{x}$ is the average value of the data, and x_{var} is the variance of the data.

(2) Parameter setting of clustering algorithm: K value of the traditional K-means algorithm was 6; the initial K value of the improved K-means algorithm was 6, the initial standard deviation was set as 0.6, the splitting parameter was set as 0.6, splitting parameter c_1 was set as 2.1, and aggregation parameter c_2 was set as 0.4.

(3) Training model: The traditional and improved K-means algorithms were trained using training samples, the normal data and abnormal data among the training samples were classified, and the clustering centers of respective categories were obtained.

(4) Model test: Two algorithms were tested using the test samples after training.

4. Evaluation indicator

The performance evaluation of the intrusion detection algorithm [15] was usually expressed by three data, which are accuracy rate B_C, false alarm rate E_A and missing report rate N_A For the performance of intrusion detection algorithm, the higher the accuracy, the better; the lower the false alarm rate and missing report rate, the better. The calculation formulas are:

$$B_C = \frac{C_P + C_N}{C_P + C_N + M_P + M_N}, \tag{6}$$

$$E_A = \frac{M_N}{C_N + M_N}, \tag{7}$$

$$N_A = \frac{M_P}{C_P + M_P}, \tag{8}$$

where C_P refers to the attack data which are correctly classified, C_N refers to the normal data which are correctly classified, M_N refers to the normal data which are wrongly classified, and M_P refers to the attack data which are wrongly classified.

4. Simulation results

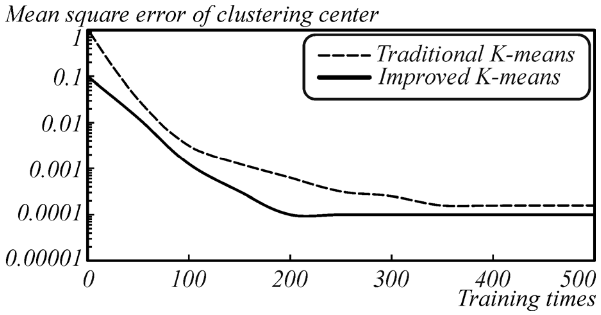

Fig. 3. Central convergence curves of two clustering algorithms during training

As shown in Fig. 3, the mean square deviations of cluster centers before and after updating gradually decreased and tended to be stable with the increase of training times in the training phase of the two clustering algorithms; the traditional K-means algorithm tended to be stable after 350 times of training, and the improved K-means algorithm tended to be stable after 200 times of training. It was found from Fig. 3 that the traditional K-means was higher than the improved K-means when the mean square deviation of the clustering center was stable.

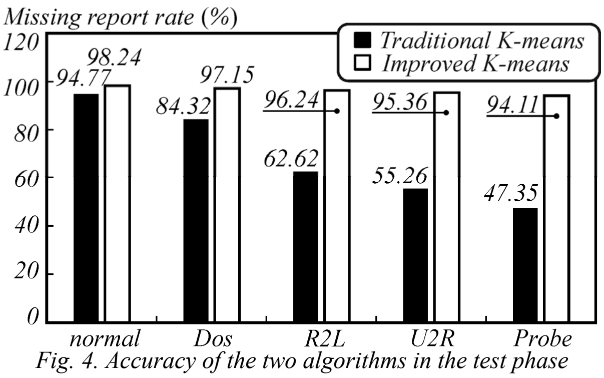

Fig. 4. Accuracy of the two algorithms in the test phase

As shown in Fig. 4, the accuracy of the traditional K-means was 94.77 % in detecting normal data, 84.32 % in detecting Dos, 62.62 % in detecting R2L, 55.26 % in detecting U2R, and 47.35 % in detecting U2R; the accuracy of the improved K-means algorithm was 98.24 %, 97.15 %, 96.24 %, 95.36 % and 94.11 % respectively in detecting normal data, Dos, R2L, U2R and Probe. It was clearly seen from Fig. 4 that the accuracy of the improved K-means algorithm was higher whether normal data or attack data were detected.

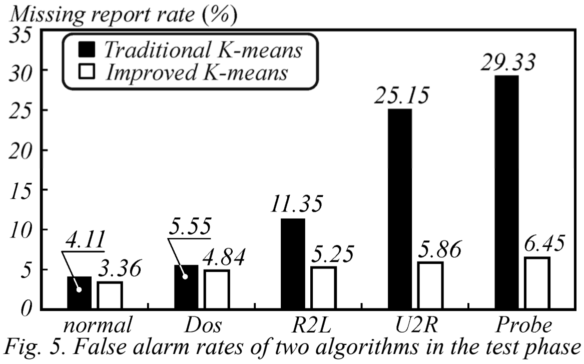

Fig. 5. False alarm rates of two algorithms in the test phase

As shown in Fig. 5, the false alarm rate of the traditional K-means algorithm was 4.11 %, 5.55 %, 11.35 %, 25.15 % and 29.33 % in detecting normal data, Dos, R2L, U2R and Probe; the false alarm rate of the improved K-means algorithm was 3.36 %, 4.84 %, 5.25 %, 5.86 % and 6.45 % in detecting normal data, Dos, R2L, U2R and Probe. It was clearly seen from Fig. 5 that the false alarm rate of the Dos attack data was close to that of the normal data, and the false alarm rates of the improved K-means algorithm were lower than the traditional K-means algorithm in detecting other kinds of data.

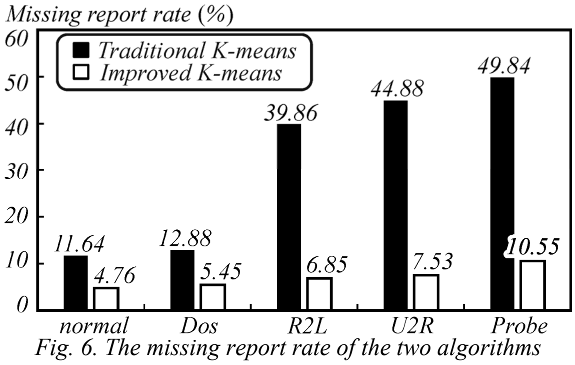

Fig. 6. The missing report rate of the two algorithms in the test phase

As shown in Fig. 6, the missing report rate of the traditional K-means was 11.64 %, 12.88 %, 39.86 %, 44.88 % and 49.84 % in detecting normal data, Dos, R2L, U2R and Probe; the missing report rate of the improved K-means was 4.76 %, 5.45 %, 6.85 %, 7.53 % and 10.55 % in detecting normal data, Dos, R2L, U2R and Probe. It was seen from Fig. 6 that the detection accuracy of the improved K-means algorithm was higher, but the accuracy in detecting the normal and Dos data were relatively closer. Moreover, comparing the three detection indicators of different types of data, it was found that both models showed higher accuracy and lower false alarm rate and missing report rate in detecting normal data and Dos data, which was because that the training and testing samples of normal and Dos data were more and the training was more. But the influence of the small size of training sample on the improved K-means algorithm was smaller; hence it was more stable than the traditional K-means algorithm in recognition.

Conclusion

This paper briefly introduced the principle of network intrusion detection and K-means algorithm and made simulation analysis on two clustering analysis algorithms on MATLAB software. The results are as follows: (1) in the model training stage, the improved K-means algorithm could complete the center convergence faster, and the mean square error of clustering center was smaller than the traditional K-means algorithm; (2) in the model testing stage, the improved K-means algorithm had higher accuracy and lower false alarm rate and missing report rate than the traditional K-means algorithm in the detection of normal and abnormal data.

References

[1] Keegan N, Ji SY, Chaudhary A, Concolato C, Yu B, Jeong DH. A survey of cloud-based network intrusion detection analysis. Human-centric Computing and Information Sciences 2016; 6(1): 19.

[2] Qiao L, Ryan M. A hybrid approach for supply chain analysis: An application of network and cluster analysis. Incose International Symposium 2017; 27(1): 746-762.

[3] He ZY. Research on network intrusion detection based on data mining technology. Appl Mech Mater 2015; 713-715: 2081-2084.

[4] Ganesh S, Ramar K. A cluster based intrusion detection system for homogeneous and heterogeneous mobile ad hoc network. J Comput Theor Nanosci 2017; 14(9): 4249-4254.

[5] Ponomarev S, Atkison T. Industrial control system network intrusion detection by telemetry analysis. IEEE Trans Dependable Secure Comput 2016; 13(2): 252-260.

[6] Ma T, Wang F, Cheng J, Yu Y, Chen X. A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks. Sensors 2016; 16(10): 1701.

[7] Wang X. Compulsory coverage network intrusion detection algorithm based on rough set theory. J Comput Theor Nanosci 2016; 13(12): 9480-9483.

[8] Vahid S, Ahmadzadeh M. KCMC: A hybrid learning approach for network intrusion detection using K-means clustering and multiple classifiers. Int J Comput Appl 2015; 124(9): 18-23.

[9] Ravale U, Marathe N, Padiya P. Feature selection based on hybrid anomaly intrusion detection system using K Means and RBF kernel function. Procedia Comput Sci 2015; 45(39): 428-435.

[10] Verma A, Ranga V. Statistical analysis of CIDDS-001 dataset for network intrusion detection systems using distance-based machine learning. Procedia Comput Sci 2018; 125: 709-716.

[11] Kang SH, Kim KJ. A feature selection approach to find optimal feature subsets for the network intrusion detection system. Cluster Comput 2016; 19(1): 1-9.

[12] Hao X, Zhang X. Research on abnormal detection based on improved combination of k-means and SVDD. IOP Conf Ser: Earth Environ Sci 2018; 114: 012014.

[13] Laftah Alyasee W, Ali Othman Z, Ahmad Nazri MZ. Hybrid modified K-Means with C4.5 for intrusion detection systems in multiagent systems. Sci World J 2015; 2015(2): 294761.

[14] Zhang Y, Wang K, Gao M, Ouyang ZY, Chen SG. LKM: A LDA-based K-means clustering algorithm for data analysis of intrusion detection in mobile sensor networks. Int J Distrib Sens Netw 2015; 2015(2): 7.

[15] Elssied NOF, Ibrahim O, Osman AH. Enhancement of spam detection mechanism based on hybrid ***k***-mean clustering and support vector. Soft Comput 2015, 19(11): 3237-3248.

Author's information

Wenhu Yang (b. 1974) has gained the master's degree. He is now an associate professor in Shandong Polytechnic. He is interested in cloud computing, big data and network security. E-mail: *whywenhu@yeah.net* .

Received October 24, 2019. The final version – December 13, 2019.