Advanced Science and Technology Letters
Vol.53 (AITS 2014), pp.429-433
http://dx.doi.org/10.14257/astl.2014.53.89

A Network Intrusion Detection Method Based on Improved K-means Algorithm

Meng Gao[1,1], Nihong Wang[1],

[1] Information and Computer Engineering College, Northeast Forestry University, Harbin, China
gaomeng0916@126.com，wnh@mail.nefu.edu.cn

Abstract. K-means algorithm could be used in intrusion detection, and selection of initial cluster centers was one of the most important factor that influenced the clustering performance, traditional method had a certain degree of randomness in dealing with this problem, therefore, information entropy was introduced into the process of cluster centers selection, and a fusion algorithm combining with information entropy and K-means algorithm was proposed, information entropy value was used to measure the similarity degree among records, it could help to choose a least similar record to be a cluster center. Comparison results show that the detection ratio and false alarm ratio of the proposed method is better than traditional K-means algorithm.

Keywords: K-means; Information entropy; Intrusion detection

1 Introduction

Network intrusion detection is a process which includes series of actions, such as collecting data related to network status and behaviors from key nodes, analyzing these data, discovering abnormal behavior as well as providing early warning [1-2], it can achieve the purpose of monitoring network behavior and defending network intrusion. As intrusion behaviors tend to have uncertainly in some degree, so, it is of great significance to identify unknown behaviors by extracting hidden information in intrusion data [3-4]. Li Wenhua proposed a FCM cluster network intrusion detection model based on fuzzy c-means [5]; Zhang Guosuo proposed an improved FCM cluster algorithm, it could solve the boundedness in dealing with big dataset by using traditional FCM [6]; Reda M. Elbasiony used random forests and weighed k-means algorithm to build intrusion patterns and choose anomalous clusters [7]; Luo Min researched on the non-supervised intrusion detection model based on K-means algorithm [8]; Li Heling proposed the improved K-means algorithm and carried out experiments aiming at the problem of uneven data distribution [9]; Researches above focused on solving the problem of data size that the algorithm can deal with, they

[1] Meng Gao, female (1989-), Ph.D., mainly engaged in forestry informatization and system security, E-mail: gaomeng0916@126.com.

ISSN: 2287-1233 ASTL

ignored the kernel of algorithm itself. This paper uses K-means algorithm to detect intrusion behaviors, as selection of initial cluster centers is the key factor that influences the cluster results, the information entropy technology is introduced to auxiliary determine cluster centers, experiments show that the improved fusion algorithm has a good detection ratio and false alarm ratio.

2 Algorithm Combing with Information Entropy and K-means

2.1 K-means Algorithm

This paper uses Euclidean distance to measure the similarity among records, and use formula (1) to evaluate the clustering results.

$$E = \sum_{i=1}^{k} \sum_{x \in d_i} | x - \frac{\sum_{x \in d_i} x}{| d_i |} |^2 . \quad (1)$$

where E is the sum of all objects' mean squared error; d_i is a data cluster i; x is a data object; $| d_i |$ is the number of objects in d_i; k is the number of clusters; the smaller value of E, the better of the clustering effect.

Data clustering process using K-means algorithm can be described as follows:

(1) Define the number k of clusters to be finally generated;

(2) Choose k records to be the initial cluster centers;

(3) Divide the original data into the k clusters, and recalculate the center of each cluster;

(4) Break the clustering result in last stage, put object j into the corresponding cluster according to the principle of minimum Euclidean distance, then, form the new clusters, and calculate the value of E at the same time;

(5) Repeat stage (4) until the new clusters are same as the previous clusters.

We can know that performance of the algorithm is mainly determined by stage (1) and (2), cluster number k is often determined according to actual situations [10-11], therefore, selection of initial cluster centers is the key factor that influences the algorithm performance.

2.2 Information Entropy

Information entropy is used to measure the uncertainty of a random variable information, the bigger of it, the more disordered of the data; otherwise, the more ordered and similar of the data [11-12]. If using information entropy to evaluate clustering effect, then the smaller of the entropy, the more similar of data in a same cluster and better of the clustering effect [13-14].

Information entropy of a random variable X can be described as:

$$E(X) = -\sum_{x \in s(x)} \log_n(p(x)) \quad . \tag{2}$$

Where $S(X)$ is the possible value set of X ; $p(X)$ is the probability function of X .

2.3 Improved K-means algorithm based on information entropy

Assume that sample space M includes n records, first, calculate the information entropy value of each record, and then start from the first record, compare the value of current record with other records, finally, regard the minimum value as the information entropy baseline of the current record, the comparison matrix is shown as Table 1.

Table 1. Comparison matrix of information entropy value

i \ j	1	2	3	…	n	Baseline
1	$E(M_1,M_1)$	$E(M_1,M_2)$	$E(M_1,M_3)$	…	$E(M_1,M_n)$	min $E(M_1,M_j)$
2	$E(M_2,M_1)$	$E(M_2,M_2)$	$E(M_2,M_3)$	…	$E(M_2,M_n)$	min$E(M_2,M_j)$
3	$E(M_3,M_1)$	$E(M_3,M_2)$	$E(M_3,M_3)$	…	$E(M_3,M_n)$	min$E(M_3,M_j)$
…	…	…	…	…	…	…
N	$E(M_n,M_1)$	$E(M_n,M_2)$	$E(M_n,M_3)$	…	$E(M_n,M_n)$	min$E(M_n,M_j)$

Calculate the information entropy baseline set $Base(M) = \{\min E(M_1, M_j), \min E(M_2, M_j), ..., \min E(M_n, M_j)\}, 1 \le j \le n$, and order the baseline from big to small, get the ordered baseline set $SortBase(M)$, the bigger of the information entropy value, the less similar between the corresponding record and other records, and the more suitable to be center of the initial cluster. Combine with the cluster number k determined in stage (1) of K-means algorithm, choose the top-k records corresponding with the information entropy values in $SortBase(M)$ as the least similar records, and these records can be regarded as the initial cluster centers.

2.4 Network Intrusion Detection Algorithm Based on IE-K-means

The process of detecting network intrusion using IE-K-means algorithm can be described as:

(1) Define the number k of clusters to be finally generated, and set the instance threshold $instanceLine$ of clusters;

(2) Choose k records as the initial cluster center $C_i(\text{i <= k})$ using IE-K-means algorithm, calculate the Euclidean distance $d(\text{i, j})$ between C_i and other records;

(4) According to the minimum $d(i, j)$, divide each record into clusters with the minimum Euclidean distance, and generate new clusters, recalculate C_i of the new clusters, and record the instance number $Instance_i$ of each cluster.

(6) Break the clustering result in last stage, and repeat stage (3)-(5) until the current clusters are the same as the previous clusters.

(7) Record C_i and $Instance_i$ of the each generated cluster;

(8) If $Instance_i < instanceLine$, mark C_i as the center of abnormal cluster $C_{i-abnormal}$; if $Instance_i > instanceLine$, mark C_i as the center of normal cluster $C_{i-normal}$;

(9) When new connection is coming, calculate the Euclidean distance $d(C_i, C_{new})$ between new connection and each C_i, if $d(C_i, C_{new})$ is closer with $C_{i-abnormal}$, mark the new connection as the abnormal intrusion; If $d(C_i, C_{new})$ is closer with $C_{i-normal}$, mark the new connection as the normal intrusion.

3 Simulation Experiment and Analysis

Use KDDCUP99 data packets to verify the feasibility and effectiveness of IE-K-means algorithm, choose 7200 DoS attack data, of which 5500 records are used as training data for training model, and the other 1700 records are used as testing data for testing the effectiveness of the intrusion detection model.

The experiment adopts different cluster number k, cluster the training data at first to get the cluster center set, and then send the testing data into the anomaly detection system for intrusion detection, calculate the $DetectRate$ and $FalseDetectRate$ of each data set at the same time, experiment results are shown in Table 2.

It can be seen that the network intrusion detection model based on IE-K-means is feasible, and the improved algorithm is better than traditional K-means algorithm in detection ratio and false alarm ratio based on different cluster amount.

Table 2. Comparison experiment results

k	K-means algorithm		IE-K-means algorithm	
	DetectRate/%	FalseDetectRate/%	DetectRate/%	FalseDetectRate/%
20	85.21	3.10	87.33	0.05
30	87.53	6.64	90.46	0.24
40	95.62	9.86	98.23	0.33

4 Conclusions

According to the characteristics of network intrusion data, aiming at the problems existed in the current intrusion detection researches, this paper proposes up a network intrusion detection method based on the fusion algorithm combining with information entropy and K-means, experiment results show that the fusion algorithm has improved the detection ratio and reduced the false alarm ratio compared with

traditional K-means algorithm. However, the implementation of the fusion algorithm did not consider the algorithm execution efficiency, which requires the further study.

Acknowledgments. This work is supported by Special Fund for Scientific Research in the Public Interest (201104037) and The Fundamental Research Funds for the Central Universities (2572014AB22).

References

1. Jonathan, J.D., Andrew, J.C.: Data Processing for anomaly based network intrusion detection: A review. Computers & Security. 30, 353--375 (2013)
2. Liao, S.H., Chu, P.H., Hsiao, P.Y.: Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications. 39, 11303--11311 (2012)
3. Mohammand, S.A., Hamid, M., Jafar, H.: Design and analysis of genetic fuzzy systems for intrusion detection in computer networks. Expert Systems with Applications. 38, 7067--7075 (2011)
4. Chen, X.H.: Intrusion Detection Method Baed on Data Mining Algorithm. Computer Engineering. 36, 75--81 (2010)
5. Li, W.H.: Network Intrusion Detection Model Based on Clustering Analysis. Computer Engineering. 37, 121-128 (2011)
6. Zhang, G.S., Zhou, C.M., Lei, Y.J.: Improved fuzzy C-means clustering algorithm and its application to intrusion detection. Journal of Computer Applications. 29, 44-51 (2009)
7. Reda, M.E., Elsayed, A.S., Tarek, E.E., Mahmoud, M.F.: A hybrid network intrusion detection framework based on random forests and weighed k-means. Ain Shames Engineering Journal. 4, 2239-2249 (2013)
8. Luo, M., Wang, L.N., Zhang, H.G.: An Unsupervised Clustering-Based Intrusion Detection Method. ACTA ELECTRONICA SINICA. 31, 158-166 (2003)
9. Li, H.L.: Study on Application of data mining in network intrusion detection. JiLin University, JiLin (2013)
10. Li, Y.: Application of K-means Clustering Algorithm in Intrusion Detection. Computer Engineering. 33, 127-135 (2007)
11. Du, Q., Sun, M.: Intrusion detection system based on improved clustering algorithm. Computer Engineering and Applicatins. 47, 166-171 (2011)
12. Ye, Z.W.: The Research of Intrusion Detection Algorithms Based on the Clustering of Information Entropy. Procedia Environmental Sciences. 12, 1329-1344 (2012)
13. Feng, J., Sui, Y.F., Cao, C.G.: An Information entropy-based approach to outlier detection in rough sets. Expert Systems with Applications. 37, 6338-6344 (2010)
14. Jin, C.X., Li, F.C., Li, Y.: A generalized fuzzy ID3 algorithm using generalized information entropy. Knowledge-Based Systems. 64, 13-21 (2014)

