International Journal of Security and Its Applications
Vol.8, No.6 (2014), pp.285-294
http://dx.doi.org/10.14257/ijsia.2014.8.6.25

A Network Intrusion Detection Model Based on K-means Algorithm and Information Entropy

Gao Meng[1], Li Dan[1], Wang Ni-hong[1] and Liu Li-chen[2]

[1]*Information and Computer Engineering College, Northeast Forestry University, Harbin, China*
[2]*College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, China*
gaomeng0916@126.com, ld725725@126.com, wnh@mail.nefu.edu.cn, liulichen_521@tom.com

Abstract

Many factors could influence the clustering performance of K-means algorithm, selection of initial cluster centers was an important one, traditional method had a certain degree of randomness in dealing with this problem, for this purpose, 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, in which, information entropy value was used to measure the similarity degree among records, the least similar record would be regarded as a cluster center. In addition, a network intrusion detection model was built, it could make cluster centers change dynamically along with the network changes, and the model could real-time update the cluster centers according to actual needs. Experiment results show that the improved algorithm proposed is better than the traditional K-means algorithm in detection ratio and false alarm ratio, and the network intrusion detection model is proved to be feasible.

Keywords: *Information entropy, K-means algorithm, Dynamic cluster center, Intrusion detection model*

1. Introduction

Network intrusion detection is a series of processing actions, it includes collecting data related to network status and behaviors from key nodes, analyzing the collected data, discovering abnormal behaviors as well as providing early warning [1-2], it can achieve the purpose of monitoring network behaviors and defending network intrusions. As intrusion behaviors tend to have uncertainly in some degree, therefore, it is of great significance to identify the unknown behaviors by extracting hidden information existing in intrusion data, cluster analysis technology can help to achieve this goal [3-4].

Li Wenhua proposed a FCM cluster network intrusion detection model based on fuzzy c-means, it could deal with numerical attributes and symbol attributes, its detection ratio was 85% and false alarm ratio was 1.5% [5]; Zhang Guosuo proposed an improved FCM cluster algorithm by improving the objective function of traditional FCM algorithm, it could solve the boundedness of traditional FCM in dealing with condition of uneven sample distribution, its detection ratio was 98.71% and false detection ratio was 0.034% [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 an non-supervised intrusion detection model

ISSN: 1738-9976 IJSIA

based on K-means algorithm, classification of training dataset would not depended on manual work or other methods [8]; Li Heling proposed an improved K-means algorithm, experiments showed that it could solve problems caused by uneven data distribution, its detection ratio is 87.6% and false detection ratio is 5.7% [9]; Other intrusion detection methods include Hidden Naïve Bayes [10], decision tree [11-13], association rules [14], hierarchical clustering [15], support vector machine [15-16], etc. Researches above mainly focused on algorithm improvements, most of them aimed at solving problem of data size that an algorithm can deal with, they ignored the kernel of algorithm itself. This paper uses K-means algorithm to detect intrusion behavior, in consideration of selection of initial cluster centers is the key factor that influences the clustering results, so, the information entropy technology is introduced to auxiliary determine the initial cluster centers, and a network intrusion detection model based on the IE-K-means algorithm is built, experiments show that the improved fusion algorithm has a good detection ratio by using the established intrusion detection model.

2. Fusion Algorithm Combing with Information Entropy and K-means

Traditional K-means algorithm has randomness in selection of initial cluster centers, this paper considers using information entropy to calculate the similarity of records, and choose the top-k records with minimum information entropy values as the initial cluster centers, and this method can help to improve the intrusion detection efficiency.

2.1. K-means Algorithm

2.1.1. Basis of K-means Algorithm: It needs to standardize the collected data before clustering, and define the standards for evaluating cluster results, these standards will be used to terminate or continue the algorithm execution [4].

(1) Standardization of data objects

$$S_j = \frac{1}{n}\sum_{i=1}^{n} | x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij} | \tag{1}$$

$$x'_{ij} = \frac{x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij}}{S_j} \tag{2}$$

where n is the number of data objects; x_{ij} is the value of attribute j of object i; x'_{ij} is the standardized value of attribute j of object i. Standardized data can remove the dimensional effects on clustering process.

(2) Euclidean distance

$$d(i, j) = \sqrt{| x_{i1} - x_{j1} |^2 + | x_{i2} - x_{j2} |^2 \ldots\ldots + | x_{in} - x_{jn} |^2} \tag{3}$$

where $d(i, j)$ is the distance between object i and the cluster center which i belongs to; $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ are the two-dimensional data objects.

(3) Evaluation function

$$E = \sum_{i=1}^{k} \sum_{x \in d_i} | x - m_i |^2 \tag{4}$$

$$m_i = \frac{\sum_{x \in d_i} x}{| d_i |} \tag{5}$$

where E is the sum of all objects' mean squared error; d_i is a data cluster i; x is a data object; m_i is the mean value of all objects in d_i; $|d_i|$ 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.

2.1.2. Process of K-means Algorithm: Use K-means algorithm to choose the initial cluster centers, this process can be described as follows:

(1) Define the amount value 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 the previous stage, and according to the principle of minimum Euclidean distance, put object i into the corresponding cluster, 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.

It can be known from the above that performance of the algorithm is mainly determined by stage (1) and (2), cluster number k is often determined according to actual situations [17-18], therefore, selection of initial cluster centers is the key factor that influences the algorithm performance. In consideration of cluster centers selection has large randomness; this paper introduces information entropy to auxiliary select cluster centers, which can optimize the clustering 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 [18-19]. If using information entropy to evaluate clustering effect, then the smaller of the entropy, the more similar of data in a same cluster, and the better of the clustering effect [20-21].

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

$$E(X) = -\sum_{x \in s(x)} \log_n(p(x)) \tag{6}$$

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

If $X = \{x_1, x_2, ...x_n\}$ includes multiple attributes, its information entropy can be described as:

$$E(X) = -\sum_{x_1 \in s(x_1)} \sum_{x_2 \in s(x_2)} ..... \sum_{x_n \in s(x_n)} p(x_1, x_2, ..., x_n) \log_n(p(x_1, x_2, ..., x_n)) \tag{7}$$

and if x_1, x_2, $\ldots x_n$ is mutual independence, then $E(X)$ can be described as:

$$E(X) = -\sum_{x_1 \in s(x_1)} \sum_{x_2 \in s(x_2)} ..... \sum_{x_n \in s(x_n)} p(x_1)p(x_2)....p(x_n) \log_n(p(x_1)p(x_2)...p(x_n)) \tag{8}$$

2.3. IE-K-means Algorithm based on Information Entropy

For purpose of reducing the randomness in the selection process, this paper mainly studies using information entropy to optimize the selection of initial cluster centers, which can achieve the goal of better clustering, results.

Assume that a 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 baseline set $Base(M_j) = \{\min E(M_1, M_j), \ldots, \min E(M_i, M_j)\}$, $1 \le i \le n, 1 \le j \le n$, order the information entropy baselines from big to small, and get the ordered baseline set $SortBase(M_j)$, the bigger of the information entropy value, the less similar between the corresponding record and other records [22], and the more suitable to be center of the initial cluster. Combining with the clusters amount 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.

The pseudocode of calculating initial cluster centers can be described as:

```
for i=1 to n
  for j=1 to n
    if i<>j
      Calculate E(Mi,Mj);
for i=1 to n
  minE = E(Mi,M1);
    for j=1 to n
      if(E(Mi,Mj)< minE)
minE=E(Mi,Mj);
        minESet[]=minE;
for i=1 to n
  for j=i to n-1
    if(minESet[j]>minEset[j+1])
      s=minESet[j];
      minESet[j]= minESet[j+1];
      minESet[j+1]=s;
for i=1 to k
  Centerset[i]=minESet[i];
```

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 amount value k of clusters to be finally generated, and set the instance threshold $instanceLine$ of clusters;

(2) Choose k records as the initial cluster center $C_i (i <= k)$ using IE-K-means algorithm;

(3) Calculate the Euclidean distance $d(i, j)$ 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.

(5) Recalculate C_i of the new clusters, and record the instance number $Instance_i$ of each cluster.

(6) Break the clustering result in the previous 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;

(10) 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. Network Intrusion Detection Model based on IE-K-means Algorithm

Based on the fusion algorithm combining with information entropy and K-means (IE-K-means), this paper builds a network intrusion detection model, it regards the network intrusion data as input and marked anomaly intrusion tag as output, the model includes 4 kernel units: data standardization processer, IE-K-means clustering tool, updater of cluster centers and anomaly detection system. The structure of this model is shown as Figure 1.

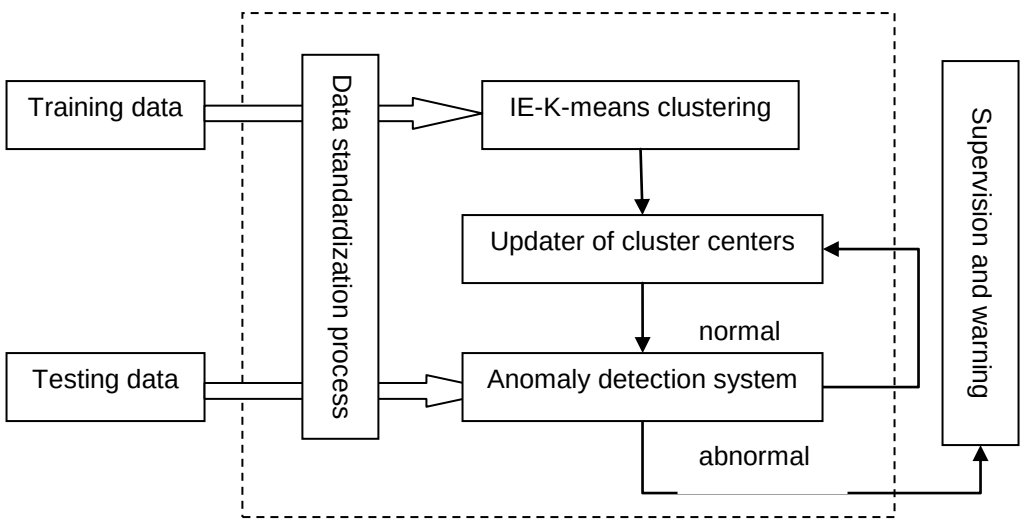


Figure 1. Network Intrusion Detection Model based on IE-K-means Algorithm

3.1. Data Standardization Processer

As network intrusion data may contain both numerical and symbolic attributes, therefore, it needs to quantize the symbol attributes, and standardize them together with numerical attributes, this process can help to unify the dimension of these attributes. Standardization process can avoid the situation that big number is dominant absolutely and small number can not fully play roles. Standardization method has been discussed in 2.1.1.

3.2. IE-K-means Clustering Tool

This unit is the most important part of the model, which uses the IE-K-means algorithm as kernel for training the model. Combining with the improved algorithm and the process of K-means algorithm above, automatically generate the normal cluster center set $NormalC = \{nc_1, nc_2, nc_3, ..., nc_n\}$ and the abnormal cluster center set $AbNormalC = \{anc_1, anc_2, anc_3, ..., anc_m\}$, and save them into the updater of cluster centers for anomaly detection.

3.3. Anomaly Detection System

When new connection is coming, this unit uses formula (4) to respectively calculate the distance between new connection and elements in $NormalC$ as well as $AbNormalC$, according to the property of cluster center with minimum distance, mark the attributes of the new connection, on one hand, send the connection into the supervision and warning system for daily monitoring and warning the network intrusion; on the other hand, send the marked result into the updater of cluster centers for dynamically updating the cluster centers along with the network changes.

3.4. Updater of Cluster Centers

This unit simultaneously receipts the trained cluster center sets from IE-K-means clustering tool and the marked connections from anomaly detection, and it can realize the dynamically updating of cluster centers along with the network change, which can promote the anomaly detection accuracy. If connection $X = \{x_1, x_2, ..., x_n\}$ is closest with cluster center set $C = \{c_1, c_2, ..., c_n\}$, then mark X with C, and update C at the same time by revising its cluster center to the mean value of attributes in connection X and cluster center of C.

4. Simulation Experiment and Analysis

Use KDDCUP99 data packets to verify the feasibility and effectiveness of IE-K-means algorithm and the network intrusion detection model, choose 7200 DoS attack records, of which 5500 records are used as training data and the other 1700 records are used as testing data; choose 6990 Probing attack records, of which 5200 records are used as training data and the other 1790 records are used as testing data. Based on the 2 data groups above, respectively use K-means and IE-K-means algorithm to train the model and validate the effectiveness of the intrusion detection, then finish the comparison analysis.

Algorithm performance evaluation function can be described as:

$$DetectRate = \frac{detectednum}{total\ num} \times 100\% \tag{9}$$

$$FalseDetectRate = \frac{false\ detectednum}{total\ normal\ num} \times 100\% \tag{10}$$

where $DetectRate$ is the detection ratio; $detectednum$ is the detected intrusion number; $total\ num$ is the total intrusion number; $FalseDetectRate$ is the false alarm ratio; $false\ detectednum$ is the wrong detection number; $total\ normal\ num$ is the total number of normal records.

The experiment adopts different cluster amount 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, the comparison experiment results of the 2 data groups are respectively shown in Table 2. and Table 3.

Table 2. Algorithm Comparison Experiment Results based on DoS Attack Data

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

Table 3. Algorithm Comparison Experiment Results based on Probing Attack Data

k	K-means algorithm		IE-K-means algorithm	
	$DetectRate$ /%	$FalseDetectRate$ /%	$DetectRate$ /%	$FalseDetectRate$ /%
20	82.63	4.21	85.76	0.12
30	85.78	6.89	88.64	0.28
40	92.94	9.22	95.35	0.36

Experiment results show that the performance of traditional K-means algorithm and the improved IE-K-means algorithm are consistent on DoS attack and Probing attack, as the larger of cluster amount, the attack detection ratios are all improved and higher than 80%; the false alarm ratios are also increased, but they are all under 10%. In addition, when cluster amount is the same, the performance of IE-K-means algorithm is better than traditional K-means algorithm, which has a higher detection ratio and lower false alarm ratio. As is shown in Table 2, based on DoS attack data, when k is 40, the detect ratio and false alarm ratio are all the highest, they are respectively 95.62% and 9.86% when using traditional K-means algorithm; however, the detect ratio is 98.23% and the false alarm ratio is 0.33% when using IE-K-means algorithm. As is shown in Table 3, based on Probing attack data, when k is 40, the detect ratio and the false alarm ratio are also the highest, they are respectively 92.94% and 9.22% when using traditional K-means algorithm; however, the detect ratio is 95.35% and the false alarm ratio is 0.36% when using IE-K-means algorithm.

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.

5. Conclusions

Aiming at the problems existed in the current intrusion detection researches, and combining with the characteristics of network intrusion data, this paper proposes up a network intrusion detection model based on the fusion algorithm combining with information entropy and K-means, experiment results show that this model is feasible, and comparing with traditional K-means algorithm, the fusion algorithm has improved the detection ratio and reduced the false alarm ratio in recognizing the anomaly network intrusion. However, the implementation of the fusion algorithm and network intrusion detection model did not consider the execution efficiency, in the future; it needs to further research on the implementation method for detecting intrusion in a shorter time.

Acknowledgements

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

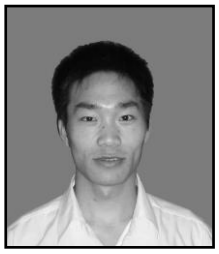
[1] J. J. Davis and A. J. Clark, "Data Processing for anomaly based network intrusion detection", A review Computers & Security, vol. 30, **(2013),** pp. 6-7**.**
[2] S.-H. Liao, P.-H. Chu and P.-Y. Hsiao, "Data mining techniques and applications", A decade review from 2000 to 2011, Expert Systems with Applications, vol. 12, no. 39, **(2012).**
[3] M. S. Abadeh, H. Mohamadi and J. Habibi, "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks", Expert Systems with Applications, vol. 6, no. 38, **(2011).**
[4] C. Xiao-Hui, "Intrusion Detection Method Baed on Data Mining Algorithm", Computer Engineering, vol. 17, no. 36, **(2010).**
[5] L. Wen-Hua, "Network Intrusion Detection Model Based on Clustering Analysis", Computer Engineering, vol. 17, no. 37, **(2011).**
[6] Z. Guo-Suo, Z. Chuang-Ming and L. Ying-Jie, "Improved fuzzy C-means clustering algorithm and its application to intrusion detection", Journal of Computer Applications, vol. 5, no. 29, **(2009).**
[7] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighed k-means", Ain Shames Engineering Journal, vol. 4, no. 4, **(2013).**
[8] L. Min, W. Li-Na and Z. Huan-Guo, "An Unsupervised Clustering-Based Intrusion Detection Method", ACTA ELECTRONICA SINICA, vol. 11, no. 31, **(2003).**
[9] L. He-Ling, "Study on Application of data mining in network intrusion detection", JiLin University, JiLin, **(2013)**, pp. 26-30.
[10] L. Koc, T. A. Mazzuchi and S. Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier", Expert Systems with Applications, vol. 18, no. 39, **(2012).**
[11] G. V. Nadiammai and M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal, vol. 1, no. 15, **(2014).**
[12] P. Louvieris, N. Clewley and X. Liu, "Effects-based feature identification for network intrusion detection", Neurocomputing, vol. 9, no. 121, **(2013).**
[13] F. Amiri, M. M. R. Yousefi, C. Lucas, A. Shakery and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems", Journal of Network and Computer Applications, vol. 4, no. 34, **(2011).**
[14] L. Hanguang and N. Yu, "Intrusion Detection Technology Research Based on Apriror Algorithm", Physics Procedia, C, vol. 24, **(2012).**
[15] S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L. Lai and C. D. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, vol. 1, no. 38, **(2011).**
[16] P. Sangkatsanee, N. Wattanapongsakorn and C. Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches", Computer Communications, vol. 18, no. 34, **(2011).**
[17] L. Yang, "Application of K-means Clustering Algorithm in Intrusion Detection", Computer Engineering, vol. 14, no. 33, **(2007).**
[18] D. Qiang and S. Min, "Intrusion detection system based on improved clustering algorithm", Computer Engineering and Applicatins, vol. 11, no. 47, **(2011).**
[19] Y. Zheng-Wang, "The Research of Intrusion Detection Algorithms Based on the Clustering of Information Entropy", Procedia Environmental Sciences, B, vol. 12, **(2012).**
[20] F. Jiang, S. Yu-Fei and C. Cun-Gen, "An Information entropy-based approach to outlier detection in rough sets", Expert Systems with Applications, vol. 9, no. 37, **(2010).**
[21] J. Chenxia and L. Fachao and L. Yan, "Generalized fuzzy ID3 algorithm using generalized information entropy", Knowledge-Based Systems, vol. 64, **(2014).**
[22] L. Jiye, Z. Xingwang, L. Deyu, C. Fuyuan and D. Chuangyin, "Determining the number of clusters using information entropy for mixed data", Pattern Recognition, vol. 6, no. 45, **(2012).**

Authors

Gao Meng, Female, born in 1989, Ph.D., studying in Information and Computer Engineering College of Northeast Forestry University, mainly engaged in forestry informatization, system security.

Wang Ni-hong, Female, born in 1952, Ph.D. supervisor, working in Information and Computer Engineering College of Northeast Forestry University, mainly engaged in forestry informatization, Internet of Things.

Li Dan, Male, born in 1983, lecturer, working in Information and Computer Engineering College of Northeast Forestry University, mainly engaged in forestry informatization.



Liu Li-chen, Male, born in 1983, Ph.D., studying in College of Mechanical and Electrical Engineering of Northeast Forestry University, mainly engaged in mechanical and electrical engineering , vehicle control.