Application of An Improved K-means Clustering Algorithm in Intrusion Detection

Dongmei Yu[1, a], Guoli Zhang[1 b] and Hui Chen[2, c]

[1] North China Electric Power University, Baoding, 071003, China;

[2] Ji'nan GEELY automobile Co., Ltd, Jinan, 250000, China.

[a]1430892268@qq.com, [b]zhangguoli_hebei@sina.com, [c]980636211@qq.com

Abstract. For the initial clustering center usually choose the randomness of the problem, the pa-per proposes a new initial clustering center selection method. First, the algorithm calculates the Euclidean distance of all data to the origin of the coordinate, and then evenly divide the k class, at last, the average value of each class is calculated, and the k center is selected by this method. And through the experimental comparison of the improved algorithm with the merits of the original algorithm and the improved k-means algorithm has been proposed. The experimental results show that the improved algorithm greatly improves the stability and the computation efficiency of the algorithm.

Keywords: K-Means algorithm; Clustering center; Clustering analysis.

1. Introduction

With the rapid development and wide application of the network, people realize the convenience of network, at the same time, the security of computer network becomes more and more important. In the face of increasingly serious security problems, the intrusion detection technology with the development of network technology and related disciplines and matures, intrusion detection is used to identify the unauthorized use of computer system (such as hackers) and individuals have legal authorization but abusing its users. Most of the existing intrusion detection systems adopt the expert system or the system based approach, which requires a lot of experience. The advantage of the method of data mining is that it can extract the knowledge and rule of the people's interest and unknown knowledge from a large amount of data, but it is not dependent on the experience [1].

Clustering analysis is an important research content in data mining.The K-Means algorithm is a classical algorithm in clustering analysis. In 1957, Lloyd [2] proposed K means algorithm in the literature for the first time. The algorithm is simple and efficient. But the K-Means algorithm is a local search technique, which is easily affected by the initial cluster center. To solve these problems, there have been derived algorithm combined with K-Means algorithm and other algorithms. literature [3] proposed a new method based on the data sample distribution to select the initial cluster center. Although the algorithm has improved the stability and accuracy, it is still not good enough. In order to solve the problem of random initial point, proposed an improved clustering algorithm by calculating the Euclidean distance to coordinate all the data into homogeneous classes, each class is then calculated the mean method of selecting the center. Experimental results show that the proposed algorithm is better than the traditional K-Means method and the improved algorithm proposed by literature [4], which has higher stability and computational efficiency.

Finally, an improved K-means clustering algorithm we proposed is applied to the intrusion detection system, this algorithm can automatically analyze the original data, make inductive reasoning, so as to identify potential patterns, predict customer behavior, classification and description of intrusion behavior. Simulation results show that the method is practical and accurate.

2. The concept of intrusion detection system

Intrusion detection is a test of intrusion behavior. It is through the collection and analysis of information network behavior, security log audit data, and other available on the network and computer systems in a number of key points of information, check the network or system exists in violation of security policy and signs of attack. Intrusion detection system can through to the

administrator from intrusion or intrusion attempts to strengthen the current access control system, such as firewall; identification of firewall usually cannot identify the attacks such as internal attacks, to provide the necessary information after the discovery of intrusion attempts. The framework of intrusion detection [5] is shown in Figure 1.

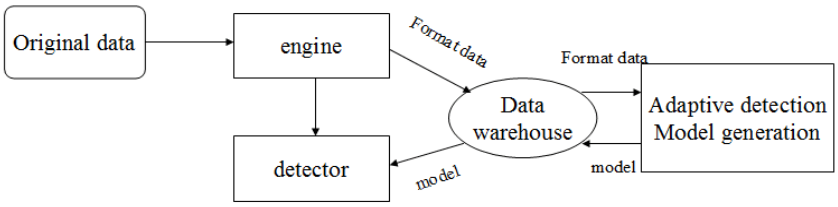

Fig. 1 the framework of intrusion detection

3. Standard K-means algorithm [6] overview

The clustering problem in R^p space can be simply described as follows: for the data set containing n points, the similarity is divided into k clusters according to their similarity.

The set of cluster samples is: $X=\{x_i \,|\, x_i \in R^p, i=1,2,...,n\}$, the k clustering center is $c_1, c_2, ..., c_k$, which makes $G_j\,(j=1,2,...,k)$ to represent k disjoint partition set.

a) Randomly selected k points $c_1, c_2, ..., c_k$ from $X=\{x_i \,|\, x_i \in R^p, i=1,2,...,n\}$, as the initial center of the k cluster set, determine the convergence accuracy of the standard function ϖ;

b) According to the central object of each cluster, the distance between each sample object of rest and the center object is calculated;

c) Use the following formula to calculate the new cluster center $g_1, g_2, ..., g_k$,

$$g_i=\frac{1}{n_{G_i}}\sum_{x\in G_i} x \quad i=1,2,...,k. \tag{1}$$

G_i represents the first i cluster, n_{G_i} represents the number of data within the first i cluster.

d) Calculating average criterion function[7]

$$E_1=\sum_{i=1}^{k}\sum_{x\in G_i}\left\|x-g_i\right\|^2 \quad i=1,2,...,k. \tag{2}$$

e) Calculation of new distribution modes: if

$$\left\|x_j-g_{i^*}\right\|^2<\left\|x_j-g_i\right\|^2, i=1,2,...,k, x_j\in G_i, i^*\neq i,\ j=1,2,...,\ n. \tag{3}$$

The sample x_j is assigned to the cluster G_{i^*},if $i<n, i=i+1$.

f) To calculate the average criterion function again

$$E_2=\sum_{i=1}^{k}\sum_{x\in G_i}\left\|x-g_i\right\|^2 \quad i=1,2,...,k. \tag{4}$$

g) If the $E_1-E_2<\varpi$ is set up, then the calculation is terminated, otherwise, to the first (3) step.

The shortage of the algorithm [8]

K-means algorithm is sensitive to the initial cluster center, and the general initial center is randomly selected, which can cause a lot of difference. In addition, in the K-means algorithm, often K is given in advance, usually, in advance does not know how many categories of the given data set should be divided into appropriate, which is a problem K-means algorithm.

4. Improved K-means algorithm

Literature [4] proposed a specificity of the improved algorithm based on data, in this paper the data, in accordance with the difference between the data of ascending order, specify the distance of the farthest data objects for different clusters, and then the rest of the data according to its associated with each cluster dissimilarity values, assigned to the minimum values of the clusters. Finally, compute the mean of all data objects in each cluster as the center of the cluster. This method can get a very good initial center and reduce the number of iterations without the need of clustering. However, when selecting each cluster member, the method is easy to miss the data points, and the two time allocation of the non grouped data points is needed, and the computational efficiency is not perfect. In order to make all the data points can be divided into the cluster, the algorithm is improved according to the following methods.

The main steps are as follows:

a) First, $data=\{x_i \,|\, x_i \in R^p, i=1,2,...,n\}$ containing n in p dimensional data object data set is calculated for each data point from the origin of the Euclidean distance in $x_0=(0,0,...,0)$.

b) According to the size of the distance of each data point from the origin $x_0=(0,0,...,0)$, the data were from small to large arrangement, get another data set $data2=\{y_i \,|\, y_i \in R^p, i=1,2,...,n\}$.

c) According to the following method, the initial members of each cluster are selected.

$$d_i=[\frac{(i-1)}{k}*n+1],\ n \text{ represents the number of data objects, } d_i \in [1,n]. \tag{5}$$

The class of G_i, which is the initial member of the y_{d_i} cluster, is uniformly separated from the first i cluster.

$$\begin{cases} G_i=\{y_{d_i}, y_{d_i+1},...y_{d_{i+1}-1}\} & 1\le i<k; \\ G_k=\{y_{d_k}, y_{d_k+1},...y_n\} & i=k; \end{cases} \quad i=1,2,...,k \text{ represents a class of } k. \tag{6}$$

$[x]$ represents no more than the largest integer, so that all the data can be divided into a set of identified.

d) At the end of each cluster to calculate the center of all data points. The calculation formula for the center point $(g_1,g_2,...g_j...g_k)$ in the cluster is:

$$g_j=\frac{1}{n_{G_j}}\sum_{x\in G_j} x \quad j=1,2,...,k. \tag{7}$$

Where x represents all data objects within the cluster G_j, n_{G_j} represents the number of data objects in the cluster G_j.

5. Simulation and analysis of experiment

5.1 Description of the experiment.

Following experiments respectively using k-means algorithm, literature [4] algorithm and the improved algorithm in Matlab environment simulation experiment, the experimental environment for hardware configuration for Intel (R) core (TM) i5-3470 CPU@3.20GHz 4.00GB memory of the computer is, development environment for matlab R2013a.

The experimental data consist of a combination of data1 data and a set of Iris data sets in the UCI database. UCI database is a data set which is specially used to test the performance of clustering algorithm. Data1 for the three dimensional distribution of data sets, is divided into 3 classes, each class of data from 50, a total of 150 data, data set distribution such as Figure.2, which black points indicate data set, the black circles indicate the cluster center. Iris data set contains 150 data sets,

divided into 3 classes, each class 50 data, each data contains 4 attributes: the length of the sepals, the width of the sepals, the length of the petal, the width of the petal Accuracy statistics [9].

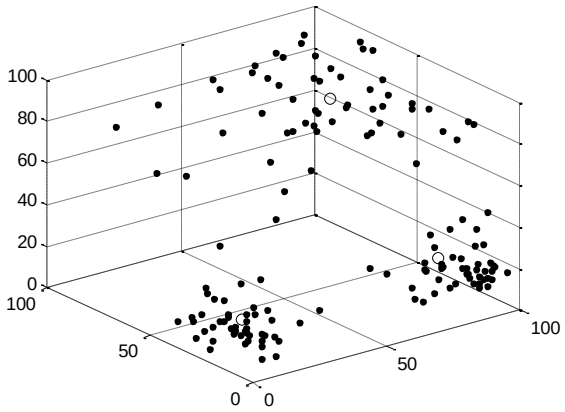

Fig. 2 the 3D distribution of the experimental results of data1

Because the selected data set has determined the number of data sets, so this paper directly determine the K value according to the number of packets. Because of the data in the dataset has its category, may wish to each data do a number, when after the implementation of the clustering algorithm and number of data in each cluster statistics, cluster in most data number is the number of this class, the discrete points individually numbered. And then compare the corresponding numbers of the data objects and their clusters after each execution.

5.2 Experimental results.

Table 1 Test results of three algorithms on data sets

Data set	Clustering algorithm	Initial center	E	Iteration number	Accuracy rate
Iris	Kmeans	5.2,3,5,1.1,0.2;4.7,3.6,1.4,0.1;5.4,3.0,1.7,0.4	124.022	12	60.1%
iris	Kmeans	7.2,3.2,4.0,1.8;6.4,2.8,5.1,1.9;4.9,3.5,4.5,1.7	97.204	13	85.3%
iris	Ref.[4]	5.2,3.4,1.4,0.2;5.5,3.0,4.2,1.2;6.8,3.4,6.1,2.3	97.224	4	87.9%
iris	Proposed method	5.1,3.3,1.6,0.3;5.8,2.8,4.2,1.4;6.8,3.1,5.8,1.9;	97.204	4	89.3%
data	Kmeans	10,51,23;84,79,106;109,55,140;	e+3*5.19	9	57.7%
data	Kmeans	41,49,93;10,80,83;60,59,73;	e+3*3.20	7	89.9%
data	Ref.[4]	13,23,19;89,16,21;69,74,91;	e+3*3.21	4	90.5%
data	Proposed methed	15.6531,25.6735,13.6531;81.36, 15.44,24.56;76.94,62.7,81.22;	e+3*3.21	3	94.6%

5.3 Experimental analysis.

The data data set and iris data set are used in the K-means algorithm, and the improved literature [4] algorithm is proposed in this paper. The proposed algorithm is compared with the improved K-means algorithm proposed in this paper. From the experiment results analysis, through four randomly selected initial clustering centers, respectively, corresponding to the number of iterations, because the test data sets have been identified in the number of packets, when after the implementation of the algorithm, provided in accordance with the literature [9] and the accuracy of calculation method of statistics, the highest reached 94.6% only the lowest 57.7%, which led to the results of clustering are poorer accuracy. Although the second random selection, the accuracy rate reached 85.3% and 89.9%, close to the accuracy of the improved algorithm, but the number of iterations is improved after three times. Compared to the traditional K-means clustering algorithm, the new algorithm proposed in this paper to ensure computational efficiency at the same time, the accuracy and stability in 89%, algorithm presented in literature[4] although reached a relatively high level, in the same number of iterations, in criterion function and the accuracy, this algorithm is superior, and after the improved

algorithm select the initial cluster center has good stability and accuracy, making the efficiency of the clustering algorithm to improve the 33%~67%. This also shows that after the improvement algorithm is a feasible and efficient algorithm.

6. Improved K-means clustering algorithm in the application of intrusion detection

In this paper, using the improved K-means clustering algorithm is proposed for the classification of a small example of the user behavior database [1], Table 2 (in addition to the class column) lists the data [10] 20 network level connection records, showing some of the characteristics of the user login. Class column: "1" means "normal", "2"means "abnormal", "3"means "attack". Table 3 explains the significance of the related characteristic parameters [11].

Table 2 Network connection records and classification

Num	Count	Serror	Same _srv	Diff _srv	Srv _count	Srv _serror	Srv_diff _host	Class
1	18	0.89	1.0	0.1	28.0	0.6	0.6	3
2	1	0.0	1.0	0.0	4.0	0.0	0.5	1
3	15	0.88	0.88	0.12	25	0.5	0.0	3
4	6	0.75	0.3	1.0	6.0	0.5	0.0	2
5	8	0.95	0.25	0.75	7.0	0.6	0.3	2
6	7	0.82	0.18	0.9	6.0	0.5	0.0	2
7	1	0.0	0.55	0.67	2.0	0.0	0.0	1
8	1	0.0	1.0	0.0	2.0	0.0	0.0	1
9	1	0.2	1.0	0.0	3.0	0.0	0.33	1
10	2	0.1	0.9	0.1	5.0	0.0	0.55	1
11	1	0.0	0.75	0.0	6.0	0.0	0.5	1
12	6	0.8	0.1	0.9	6.0	0.2	0.2	2
13	7	0.85	0.05	0.85	6.0	0.1	0.0	2
14	1	0.0	0.85	0.0	5.0	0.0	0.45	1
15	1	0.2	0.65	0.3	4.0	0.0	0.65	1
16	1	0.2	0.5	0.5	4.0	0.0	0.0	1
17	2	0.0	0.6	0.4	5.0	0.0	0.1	1
18	1	0.0	0.67	0.33	6.0	0.0	0.2	1
19	7	0.9	0.33	0.67	7.0	0.6	0.2	2
20	6	0.8	0.0	1.0	6.0	0.0	0.0	2

Table 3 Characteristic parameters and their meanings

Parameter	Meaning	Type
Count	In a time window,the target host is the same number of connections with the current connection .(the following attributes are connected to the same host).	Unit
Serror	Percentage of SYN errors in connection	Unit
Same_srv	Target port (service) the same percentage of the connection	Unit
Diff_ srv	Target port (service) the percentage of different connections	Unit
Srv_count	The destination port (service) is the same number of connections as the current connection (the following attributes are connected to the same service)	Unit
Srv_serror	Percentage of SYN errors in connection	Unit
Srv_diff_host	The percentage of different connection of the target host	Unit
Class	Cluster result	Unit

Table 4 Cluster result

Class	Cluster center
1	(1.18,0.06,0.77,0.21,4.18,0.0,0.29)
2	(6.71,0.84,0.17,0.87,6.29,0.36,0.1)
3	(16.5,0.44,0.94,0.06,15.0,0.25,0.3)

Table 5 Detection rule

Rules	Meaning
Normal	Count ≥ 15;Serror ≥ 88% Srv_count ≥ 25
Abnormal	Count ≥ 6;Serror ≥ 75% Srv_count ≥ 6
Attack	Do not meet the above conditions

This algorithm is realized by MATLAB programming, according to the characteristics of data provided by the program, after 5 iterations, the program identifies 3 types of records: attack, anomaly, and normal.This paper proposes the K-means clustering algorithm to identify the results shown in table 2class column. As can be seen from the table, records 1 and 3 are the only records that have a tendency to attack. While recording 4~6, 12, 13, 19, 20 is a 7 record of abnormal behavior, need to do further observation; the rest of the 2, 7~11, 14~18 is safe.

Table 4 is the result of the clustering algorithm, through the analysis of the results, summarizes 3 kinds of rules are shown in Table 5 and its meaning, mode of normal and aggressive behavior of these rules can be used as intrusion detection model is retained in the data warehouse, as a basis for predicting and judging the validity of user's behavior.

In this paper, the use of clustering analysis of the K-means algorithm analysis of user behavior database, from which to screen out the security of users, but also by virtue of the algorithm in the security level, to help build intrusion detection library. This algorithm because of its simple operation, low requirements for data sets, especially it can optimize or completely abandon the existing model, the re classification of user behavior, from the continuous mining potential new mode, which makes the method has wide application prospect in the field of intrusion detection.

7. Conclusion

Based on the classical K-means algorithm is sensitive to the initial clustering center this problem, we propose an improved K-means algorithm, through the improvement of the algorithm greatly reduces the number of iterations of the algorithm, improve the computational efficiency, ensure the quality of clustering quality and stability. At last, through the experiment of the algorithm and the practical application in the intrusion detection, the results show that the improved K-means clustering algorithm is effective. Compared with the traditional method, this algorithm may still have deficiencies in the detection accuracy, in order to reduce the rate of miscarriage of justice and also for primary outcome after re classification or in conjunction with other algorithm analysis, improve the degree of recognition.

References

[1] Yang Li. Application of K-means clustering algorithm in Intrusion Detection [J]. Computer Engineering.Vol.33 (2007) No.14, p.154-156.

[2] Lloyd S. Least squares quantization in PCM [J].Information Theory, IEEE Transactions on, Vol.28 (1982) No.2, p.129-137.

[3] Yifeng Xu, Chunming Chen, Yunqing Xu. An improved K mean clustering algorithm [J]. Computer Applications and Software.Vol.25 (2008) No.3, p.276.

[4] Qirui Dong. Improvement and implementation of K mean clustering algorithm [D]. Changchun: Jilin University.2015.

[5] Jianbin Hu. Intrusion Detection Technology [DB/OL]. Research Laboratory of network and information security, Peking University, 2004.

[6] Fei Li, Bin Xue, Yalou Huang. Optimization of initial center K-Means clustering algorithm [J]. Computer Science Vol.29 (2002) No.7, p.94-95.

[7] Linhua Lu, Bo Wang. An improved genetic clustering algorithm [J]. Computer Engineering and Applications.Vol.43 (2007) No.21, p.171.

[8] Aiwu Zhou, Yafei Yu, Research on K-Means clustering algorithm [J]. Computer Technology and Development. Vol. 21 (2011) No.2, p.62-63.

[9] Jianyu Zhu. Research and application of K mean algorithm and its application [D]. Dalian University of Technology. 2014.

[10] Yingxia Dai, Yifeng Lian, Hang Wang. System security and intrusion detection [M]. Beijing: Tsinghua University press, 2002.

[11] Shoushan Luo, Intrusion detection [M]. Beijing: press of the University of Posts and Telecommunications, 2004.