Efficient K-means Algorithm in Intrusion Detection

Wenjun Yang
Zhejiang Yuexiu University of Foreign Languages, Shaoxing, 312000, Zhejiang, China

***Abstract*—In order to improve the detection rate of invasion, reduce false detection rate and put forward a method based on density and maximum distance of k means clustering algorithm, the clustering results used in intrusion detection, improved the original algorithm in the choice of initial clustering center, simplify the computational complexity of the algorithm. Finally simulation experiments using KDD Cup 99 data set. Results show that the model can obtain ideal intrusion detection rate and false detection rate.**

Keywords- intrusion detection, K-means, cluster, average density

I. INTRODUCTION

The network Intrusion Detection (IDS) is in continuous monitoring, behavior of computer, network and the malicious behavior recognition and response network and information security technology. So the study of intrusion detection algorithm has important theory and the very strong practical application value. [1]

Clustering analysis method to without any tag data for training and learning, will have similar characteristics of classified data together, so the clustering analysis method has the ability of self learning, can be in does not have complete knowledge of intrusion detection under the conditions of work. If applied in clustering analysis of training data in the normal amount of data is far greater than abnormal data volume, and there are large differences between normal data and abnormal data, then using clustering analysis method can automatically identify the normal data and abnormal data. If used in clustering analysis of training data is purely normal data or abnormal data, then by clustering analysis can extract the normal network activity or abnormal network activity attribute. [2]

K - means algorithm is a kind of clustering algorithm based on partition, is also one of the most classic, the most widely used clustering algorithm. Classes and select the algorithm fixed k initial clustering center, according to the principle of minimum distance will be assigned to each sample k one in the class, after class heart constantly and adjust the pattern of categories, eventually make the samples to its for minimizing the sum of square of the center distance. But K means clustering algorithm are inevitably exist disadvantages: cannot determine the appropriate number of clustering, in advance to clustering quality is not high. According to the above problem, a paper on the basis of [3] is improved.

This paper proposes a clustering algorithm based on K - means high efficiency. Calculate the density of each sample data set parameters, at the beginning of the algorithm is divided into the data set of isolated points, point shall not participate in all kinds of sample mean contained in the process of clustering, the selection of clustering center to search in the greater than the average density of the collection, to calculate the sample set is greater than the average density of a subset of the m sample points the distance between the two, and the furthest distance in the two sample points as the initial clustering center. In the remaining (m - 2) sample points, selection into the center of the front two initial clustering centers of their maximum distance product the sample points as the third initial clustering centers. Likewise, in the rest of the sample (m - 3) on your site, select three initial cluster centers to the front of the respective product maximum distance of the sample points as the fourth initial clustering centers. K can be found and so on, the initial clustering center. According to the principle of minimum distance to the rest of the sample points is assigned to the nearest cluster center, until the complete classification of data sets.

II. PRELIMINARY KNOWLEDGE

Definition 1: [4] in the space distance of any two data objects are as follows:

$$d_{i,j} = \left(x_i - x_j\right)^T \Delta^{-1}\left(x_i - x_j\right) \tag{1}$$

Is covariance matrix of sample

Definition 2: [5] the average distance between sample points as follows:

$$avgD = \frac{1}{C(n,2)}\sum d_{i,j} \tag{2}$$

Definition 3: the density parameters of the data objects:

$$Dens(X_i) = \sum_{j=1}^{n} V\left(avgD - d_{i,j}\right) \tag{3}$$

$$V(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Definition 4: For the average of sample density:

$$MDens = \frac{1}{n}\sum_{i=1}^{n} Dens(X_i) \tag{4}$$

Definition 5: (isolated point) in the data set D samples, If

$$Dens(X_i) < r * MDens, 0 < r < 1 \tag{5}$$

Then this point as the isolated point.

III. Efficient K-Means Algorithm Based on Density and Maximum Distance

A. The Determination of Cluster Number K:

Under normal circumstances, the K value in the range of$\boldsymbol{[2, \sqrt{n}]}$. The number of objects where n for the data set. Shan-lin Yang in [6] to the k value given by upper bound $\boldsymbol{k_{max}}$ conditions, and theoretically proves the rationality of the experience rule $\boldsymbol{k_{max} \le \sqrt{n}}$.

B. Clustering Validity Index Function

Effective cluster number k search range is determined, choosing the appropriate cluster validity index is the key. Clustering validity index mainly has two kinds, namely external validity index and internal validity index. External validity index to reflect the results with the original data clustering differentiate between alignments; Internal validity index is in the raw data classified evaluation of clustering results under the condition of unknown quality of a measurement.

With the help of internal validity index to implement the determination of k. Based on clustering validity index calculation under different k value, the optimal clustering results of the clustering number as the optimal clustering number, so as to finally determine the effective value of k. The Silhouette in the commonly used clustering validity index function [7] indicators for its more excellent performance and low computational complexity, has been widely used. Silhouette index is a composite index, index of Silhouette are defined as follows:

$$sil(X_i) = \frac{b(X_i) - a(X_i)}{max\big(b(X_i), a(X_i)\big)} \tag{6}$$

The sample average Silhouette are defined as follow:

$$sil = \frac{1}{n}\sum_{i=1}^{n} sil(X_i) \tag{7}$$

$a(X_i)$ is the sample I and class's average distance from all the rest of the sample. $b(X_i)$ as samples to other samples in each class I mean distance of minimum. Silhouette index value in the range of [-1, 1], the greater the average Silhouette said clustering results the better the quality, the maximum value of the class number is optimal clustering number of expectations.

C. The Improved Algorithm

Input: The training sample set

Output: The data after clustering

Step 1: According to the formula (1) (2) to calculate the distance between sample points and the average distance.

Step 2: According to the formula (3) (4) calculate the density and the average density of samples.

Step 3: According to the formula (5) will comply with the requirements for outlier points in a set of N, the rest of the collection points in the U. The sample of greater than the average density into the set S.

Step 4: $For\ k = k_{min}\ to\ k_{max}$

In the set S to calculate the distance between any two samples, Find meet dis ($d_{1,2}$) ≥dis ($d_{i,j}$), (i, j = 1, 2, …, m) two sample points d_1 and $\boldsymbol{d_2}$, them as two initial clustering centers.

In the rest of the sample points, the selection to meet dis ($d_{1,3}$) ×dis ($d_{2,3}$) ≥dis ($d_{1,i}$) × dis ($d_{2,i}$), d_i is in addition to d_1, d_2sample points outside the sample points, d_3 as the third initial clustering centers.

And so on, until they get the k initial clustering center.

Using the original K - means clustering algorithm to the rest of the sample points according to the principle of minimum distance in K class respectively.

On the basis of the formula to calculate average Silhouette index clustering results, and record; the maximum of the cluster is expected the best clustering.

IV. Test Results and Analysis

A. Test Data Source

Experimental data from the KDD99 data set, the selection of training data set about 2700 network traffic connection records each sample data contains a total of 42 properties, including 41 as attributes, while one for decision attribute. The data set includes four invasion types, namely, the Probe (scanning and detection), DoS (denial of service attacks), U2R (illegal access to the local super user) and R2L (remote access of unauthorized)

TABLE I. SAMPLE DISTRIBUTION TABLE

Type	Training sample	Test sample
Normal	800	700
DoS	1000	200
Probe	600	400
U2L	300	200
U2R	30	20

B. The Evaluation Index

The experiment by Detection Rate (DR) and the False Detecting Rate (FDR) to measure; the performance of intrusion detection system, which are defined as follows:

$$DR = \frac{Intrusion\ data\ sample\ were\ detected}{Total\ number\ of\ \ intrusion\ \ sample} \tag{8}$$

$$FDR = \frac{\begin{array}{c}Be\ mistakenly\ identified\ as\ the\ invasion\ of\\ the\ normal\ sample\end{array}}{Total\ number\ of\ \ normal\ sample} \tag{9}$$

C. The Test Results

Through the experimental results, get the conclusion of figure I and figure II. Relative to the K - means algorithm, in this paper, the improved algorithm can improve the network intrusion detection rate, greatly reduces the false detecting rate. This algorithm can well reflect the distribution of data. Adopt the method based on density value of each data point, which can well reflect the general distribution of the data; And under the premise of this paper, based on the density, the Maximum distance between sample and using the maximum distance between two points is greater than the average density of sample selecting initial cluster centers, very representative, makes the selection of the initial clustering center and clustering center actual error is very small, shows the improved algorithm has obtained the good effect, the method of detecting rate is higher, the detection error rate is low.

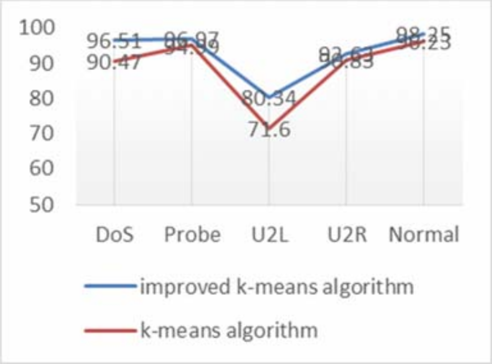

FIGURE I. THE DETECTION RATE OF DISTRIBUTION

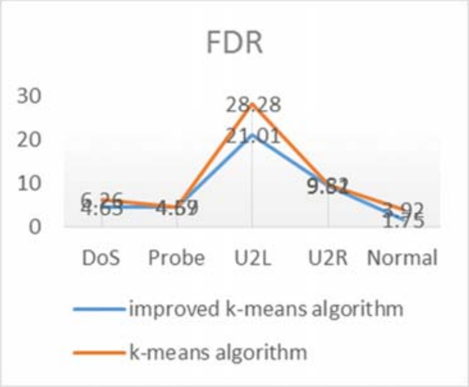

FIGURE II. THE ERROR DETECTION RATE DISTRIBUTION

V. SUMMARY

In order to effectively achieve the network attack prevention and detection, we design a k - means algorithm based on density and maximum distance, similarity algorithm using markov distance said sample, on the sample data to carry on the clustering, each sample is obtained corresponding attack category, and the clustering results used in intrusion detection, the experiments proved that the proposed method has higher detection rate and lower false alarm rate, is an effective intrusion detection method, has certain theoretical and practical reference value.

ACKNOWLEDGMENT

This research was financially supported by the Scientific Search project of Zhejiang Yuexiu University of Foreign Languages (2016QDA035). The natural science foundation of Heilongjiang Province (QC2013C067).

REFERENCES

[1] Zhao Jianhua，Li Weihua． Intrusion detection based on improved SOM with optimized GA［J］． Journal of Computers，2013，8(6)：1456-1463．

[2] Cui Wenke. Design and implementation of intrusion detection system based on clustering algorithm[D].ChenDu: School of information and software engineering，2015：31-32.

[3] Zhai Donghai etc. Maximum distance method to select the initial clustering center of the K-means of text clustering algorithm research [J]. Journal of research in computer application, 2014(3):714-719.

[4] Chu Zenan etc. Based mahalanobis distance k-means and HMM network intrusion detection method design [J]. Computer measurement and control,2014.22(10) :3406-3408.

[5] Xing long march, GuHao. Based on the average density of the k - means algorithm to optimize the initial clustering center [J]. Computer engineering and application,2016.50(20):135-137.

[6] Yang Shanlin, Li yongsen, Hu xiaoxuan, Pan Ruoyu. K - means algorithm of K value optimization study [J], systems engineering theory and practice,,2006(2):98-101

[7] Dudoit S, Fridlyand J.A prediction-based resampling method for estimating the number of clusters in a dataset[J].Genome Biology,2002,3(7)：1-21.