Research on Intrusion Detection Model of Heterogeneous Attributes Clustering

Linquan Xie
School of Science, Jiangxi University of Science and Technology, 341000 Ganzhou, China
Email: lq_xie@163.com

Ying Wang[1], Fei Yu[2,] Chen Xu[2, 3] and GuangXue Yue[4]
[1] School of Science, Jiangxi University of Science and Technology, 341000 Ganzhou, China
[2] Jiangsu Provincial Key Laboratory for Computer Information Processing Technology,
Soochow University, 215000 Soochow, China
[3] School of Information Science and Engineering, Hunan University, 416000 Changsha, China
[4] Department of Computer Science and Technology, Huaihua University, Huaihua, China
Email: hunanyufei@126.com

***Abstract*—A fuzzy clustering algorithm for intrusion detection based on heterogeneous attributes is proposed in this paper. Firstly, the algorithm modifies the comparability measurement for the categorical attributes according to the formula of Hemingway; then, for the shortages of fuzzy C-means clustering algorithm: initialize sensitively and easy to get into the local optimum, the presented new algorithm is optimized by GuoTao approach. We simulate our algorithm with the KDDCUP99 data set, and the results show that the convergence rate of the new algorithm is faster than the original fuzzy C-means clustering algorithm and the performance of our algorithm is more stable.**

***Index Terms*—Intrusion Detection, Heterogeneous Attributes, Fuzzy Clusterin.**

I. Introduction

The computer network has developed at full speed with the internet as the representative. It provides a convenience and efficient method for information open access spreading and sharing. At the same time, the network faces with kinds of security issues which are getting more serious. Intrusion detection is an important part in the field of network security research, and how to make models for intrusion detection so that it can detect intrusions fast and precisely is the key point in this area at present.

Intrusion detection can be considered as a classification problem which classifies the given datasets: what kind of data is normal and what kind of data is abnormal [1]. Cluster as an unsupervised anomaly detection algorithm, it can classify the large data sets independent of the pre-defined data types and the training set of labeled data, avoiding the high cost of marking the data.

The research of how to continuously improve the detection efficiency of intrusion detection system has always been a hot. The fuzzy C-means clustering algorithm has a preferable scalability efficiency and expandability performance when it is used in processing large data sets. But the algorithm can only process the continuous data, and it is helplessness to the discrete data. However, in fact, the KDDCUP99 dataset which is used in the simulation below is consist of continuous data and discrete data. If the research only focuses on the continuous data or numerical alternative of the discrete data simply, it may affect the efficiency of the intrusion detection. In this paper, both the continuous data and the discrete data are considered, and the similarity measure formula of the discrete attribute is improved so that detection efficiency can be enhanced. We also propose an intrusion detection algorithm of fuzzy clustering based on heterogeneous attributes; then for the shortages of the fuzzy C-means clustering algorithm: initialize sensitively and easy to get into the local optimum, optimize the presented new algorithm combining with GuoTao algorithm.

In paper, the first section will introduce the intrusion detection algorithm of fuzzy clustering based on heterogeneous attributes; the second section will recommend the Intrusion Detect System of the algorithm; the third section will describe the simulation and the performance analyses will be given; the last part of this paper will draw the conclusions.

II. Fuzzy Clustering Algorithm for Intrusion Detection based on Heterogeneous Attributes

A. Fuzzy C-means Clustering Algorithm

The detail of the fuzzy C-means clustering algorithm is to divide a dataset X which contains n instances into K categories, $(1 < K < N)$, $X = \{X_i / X_i \in R(i = 1,2,\cdots,n)\}$, according to the

Manuscript received June 11, 2011; revised June 29, 2011; accepted August 27, 2011.

doi:10.4304/jsw.7.12.2823-2831

minimization principle of the sum of squares in sub-groups, which category does the data belong to determined by using the membership, and calculate each clustering center so that the objective function is minimum[2]. The matrix of classification express as $U = \left(\mathrm{u}_{ij} / i = 1,2,\cdots n;\ j = 1,2,\cdots,k\right)$, u_{ij} mean the membership between i and j , i express an extent instance, j express an extent category, and it must satisfy the conditions as follow:

$$\sum_{j=1}^{k} u_{ij} = 1, \forall i = 1,\cdots,n \tag{1}$$

The membership between 0, 1 of each data objects can determine the cluster which their belong to after fuzzy partition of using Fuzzy C-means Clustering. The interval of elements of the matrix U is $[0,1]$ .We defined the objective function as follow[2]:

$$J_{\mathrm{m}}(U,C) = \sum_{\mathrm{i}=1}^{N}\sum_{j=1}^{k} \mathrm{u}_{ij}^{\mathrm{m}} d_{ij}^{2}\left(X_i, C_j\right) \tag{2}$$

In the function, J_m is seen as the sum of squares of the distance between each instance and cluster center; $C_j \in I$ mean clustering centers, and $C = \left\{C_j / C_j \in I, j = 1,2,\cdots,k\right\}$; $X_i \in I$ is the data set of instances; u_{ij} indicate the membership between instance i and clustering center j , its interval is $[0,1]$, $U = \left\{u_{ij}\right\}$ is a matrix with $n \times k$, and $C = \left[C_1, C_2, \cdots, C_k\right]$ is a matrix with $s \times k$; $X_i \in R^{\mathrm{p}}$ indicate the data objects; $d_{ij}\left(X_i, C_j\right)$ mean the distance of the instance i and the clustering center j ; m $(1 \le m < \infty)$ is the fuzzy coefficient; k is the number of categories which was given in advance, and determined by the initial clustering. The necessary condition of minimizing J_m using the Lagrange multiplier method is [2]:

$$\mathrm{u}_{ij} = 1 / \sum_{i=1}^{k} (d_{ij} / d_{i1})^{2/(m-1)}, \forall i \tag{3}$$

$$c_{ij} = (\sum_{i=1}^{m} u_{ij}^{m} x_j) / (\sum_{i=1}^{m} u_{ij}), \forall j \tag{4}$$

The fuzzy coefficient m is a scalar used to control the fuzzy clustering algorithm in formulas, it can measure the blur length of the membership matrix U , the greater the value m is, the algorithm represents more blurred. As $m = 1$, the fuzzy C-means clustering algorithm reduces to the traditional C-means clustering algorithm, if we want to make the objective function to minimize, we need to calculate iteratively for the fuzzy C-means clustering algorithm.

B. Heterogeneous Attributes of Fuzzy Clustering

Number citations consecutively in square brackets [1]. No punctuation follows the bracket [2]. Use “Ref. [3]” or “Reference [3]” at the beginning of a sentence:

Give all authors’ names; use “et al.” if there are six authors or more. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. In a paper title, capitalize the first word and all other words except for conjunctions, prepositions less than seven letters, and prepositional phrases.

For papers published in translated journals, first give the English citation, then the original foreign-language citation [6].

For on-line references a URL and time accessed must be given.

At the end of each reference, give the DOI (Digital Object Identifier) number as long as available, in the format as “doi:10.1518/hfes.2006.27224”

C. Footnotes

Research on the heterogeneous attributes of the sample data, the distance can only be suit to numerical data, so we solve the problem with a new approach according to the paper [3]. The description of the distance between x_i and x_j of the categorical attributes is:

$$d\left(x_{ik}, x_{jk}\right) = \begin{cases} a, x_{ik} \ne x_{jk} \\ 0, x_{ik} = x_{jk} \end{cases} \tag{5}$$

In (5), a indicates the distance of x_i and x_j when k has the dissimilar values. Assume the number of continuous attributes and categorical attributes respectively indicate p , q , the distance between objects can be expressed as:

$$d\left(x_i, x_j\right) = d_n\left(x_i^{'}, x_j^{'}\right) + d_c\left(x_i, x_j\right) \tag{6}$$

In (6), $d_n\left(x_i^{'}, x_j^{'}\right)$ means the distance between objects of numeric attributes after standardizing, $d_c\left(x_{ik}, x_{jk}\right)$ means the distance between objects of categorical attributes.

The objective function of the heterogeneous attributes datasets can be modified from the formula(2), it can be expressed as[3]:

$$J_m(U,C) = \sum_{i=1}^{n}\sum_{j=1}^{k} \mu_{ij}^{m} \left\{ \sum_{l=1}^{p} \left(x_{il}^{'} - x_{jl}^{'}\right)^2 + \sum_{l=p+1}^{p+q} d_c\left(x_{il}, x_{jl}\right) \right\} \tag{7}$$

In (7), $m > 1$ is the fuzzy coefficient, it is used to control the blur length of the membership matrix U .

Suppose

$$C_i^n = \sum_{j=1}^{k} u_{ij}^{m} \sum_{l=1}^{p} \left(x_{ik}^{'} - x_{jl}^{'}\right)^2 \tag{8}$$

$$C_i^c = \lambda \sum_{j=1}^{k} u_{ij}^{m} \sum_{l=p+1}^{p+q} d_c\left(x_{ik}, x_{jk}\right) \tag{9}$$

Because of C_i^n and C_i^c are non-negative, we can minimize $J_m(U,C)$ by respectively minimizing C_i^n and C_i^c. Meanwhile, the expression can be described as follow by using the Lagrangian multiplier method [3].

$$u_{ij} = \left\{ \sum_{l=1}^{k} \left[\frac{d(x_i, x_j)}{d(x_l, x_j)} \right]^{\frac{2}{m-1}} \right\}^{-1}, \forall i \tag{10}$$

The formula of cluster center can be corrected as follows:

$$C_{ij} = \begin{cases} C_{il}^{'} = \dfrac{1}{\sum_{i=1}^{n} u_{ij}^{m}} \sum_{i=1}^{n} u_{ij}^{m} x_{il}; l = 1,2,\cdots,p \\ C_{il} = C_l^{\max}; l = p+1,\cdots,p+q \end{cases} \tag{11}$$

Because of $m > 1$, we can prove the algorithm is convergent.

D. Fuzzy Clustering Algorithm of Heterogeneous Attributes

The optimization process of fuzzy clustering algorithm can be summarized as follows:

Step1: initialize the membership matrix U with values between 0 and 1, so that it will satisfy the constraints: $\sum_{i=1}^{n} u_{ij} = 1, \forall j = 1,\cdots,n$.

Step2: for the different attributes of data, calculating the cluster centers as formula (11).

Step3: calculate the new membership matrix U with the formula (10).

Step4: calculate the objective function by formula (7). When the value of the objective function is less than a certain threshold, or the value of the change is less than a certain threshold, the algorithm will stop, while the result of clustering will output. Otherwise, it will get to Step2 for iterating.

With the above method, we not only consider the continuous attributes of the sample dataset, but also considered the categorical attributes. It analysis the data in the round, so that reducing the rate of fault, while combined the method with the optimization method of fuzzy clustering algorithm can further improve the detection efficiency.

E. GuoTao Algorithm

The problem for processing large data sets of fuzzy c-means algorithm as fallow: the time it takes a lot, sensitive to the initialization and it is easy to fall into the local minimum. There are many improved methods to save time, but it can't solve the problem of initialization sensitive, such as neural networks; however, genetic algorithm is a global initialization algorithm which can overcome the initialization sensitive problem of fuzzy C-means clustering algorithm.

Genetic algorithm is built on the basis of biological evolution algorithm; it is a search algorithm which is based on natural selection and genetic mechanism. Genetic algorithm needn't build models and have complex computation for the complex problems, it can find out the optimal value when only use the genetic operators.

GuoTao algorithm is a kind of random search algorithm which is proposed based on genetic algorithm, and it is a improved algorithm of genetic algorithm. GuoTao algorithm was proposed in 1999 by Guo Tao[4], it combined the sub-space search method with group climbing method, and it suitable for solving the function with inequality constraints. GuoTao algorithm is conducive to get the global optimal in search space; the application of random search strategy in subspace reflects the non-convexity of random search in sub-space, which is expressed as:

$$X^{'} = \sum_{i=1}^{m} a_i X_i^{'}, \sum_{i=1}^{m} a_i = 1, \quad -0.5 \le a_i \le 1.5$$

(12)

Assuming the search space is $V = \left\{ \begin{array}{l} X \mid X = (x_1, x_2, \cdots, x_d)^T \wedge x_{\min} \le x_i \le x_{\max}, \\ i = 1,2,\cdots,d \end{array} \right\}$, the dimension is expressed as d, the objective function $f(X)$ means the minimization function, suppose $X_j^{'} = (x_{j1}^{'}, x_{j2}^{'}, \cdots, x_{jd}^{'})^T, j = 1,2,\cdots,m$ is the different points in V, write the subspace as $V = \left\{ X \in V \mid X = \sum_{i=1}^{m} a_i X_i^{'} \right\}$, and a_i must satisfy the conditions of $\sum_{i=1}^{m} a_i = 1$ and $-0.5 \le a_i \le 1.5$. When $l_i(X) = \begin{cases} 0, if\, g_i(X) \le 0 \\ g_i(X), else \end{cases}$ and $L(X) = \sum_{i=1}^{m} l_i(X)$, the logic function Better can be defined as [6]:

$$Better(X_1, X_2) = \begin{cases} 1, if\, \mathrm{L}(\mathrm{X}_1) \le \mathrm{L}(\mathrm{X}_2) or (\mathrm{L}(\mathrm{X}_1) = \mathrm{L}(\mathrm{X}_2) \wedge f(X_1) \le f(X_2)) \\ 0, if \mathrm{L}(\mathrm{X}_1) > \mathrm{L}(\mathrm{X}_2) or (\mathrm{L}(\mathrm{X}_1) = \mathrm{L}(\mathrm{X}_2) \wedge f(X_1) > f(X_2)) \end{cases}$$

(13) $Better(X_1, X_2)$ means X_1 is better than X_2.

The advantages of GuoTao algorithm summarized in the following five aspects[5]: 1) the algorithm just have less than one hundred language program of C to implement, so it's a simple algorithm; 2) the algorithm is versatile, it can be used to solve the optimization problems of complex function; 3) the algorithm does not need to modify the parameters, it is as long as to input different function expressions for different problems; 4) as usual, the algorithm can be obtained the global optimal in a relatively short period of time; 5) when the optimal is not unique, the algorithm may also find more than one optimal.

F. Algorithm Optimization of Heterogeneous Attributes of the Fuzzy C-means Clustering

For the analysis of the original fuzzy C-means clustering algorithm, the shortage was mainly expressed as: first, the algorithm performance is not stable enough, the reason is mainly caused by the initialization sensitive; second, the algorithm is easy to fall into the local optimum; third, when the value of the function attained the minimum, the highest detection rate is not, it is in contradiction to the original fuzzy C-means clustering algorithm that the value of the function is minimum and the rate of detection is highest.

To avoid the algorithm may be confronted with these shortcomings, first of all, we choose GuoTao algorithm to solve the problem of being easy to the local optimum, described as above.

Combining with GuoTao algorithm, it need to initialize the population first, and the population with multiple individuals. Running the heterogeneous attributes of the fuzzy C-means clustering algorithm to iterative, it must successively update the cluster center and the membership of each individual of population. As each update is required to be initialized and the values of random initialization may be different, the performance of the algorithm will be instable. In the modified algorithm, it use the parallel method to deal with the problem, it use the FOR loop to update the cluster centers of all the individuals before selecting the best and the worst individual.

Owing to the application of the group search strategy of Evolutionary Computation, GuoTao algorithm ensure the global of the search space, and it is conducive to obtain the optimal set in global scope; at the same time, the algorithm only eliminate the individual with the worst fitness, the pressure is quite minimum so that ensure the diversity of population and make sure the individual with best fitness can be retained. For the problem that the value of the function attained the minimum, the highest detection rate is not, we introduced a crossover probability p , and the crossover probability p is defined as follow：

$$p=\frac{\min J_m^{(n)}}{avgJ_m^{(n)}} \qquad (14)$$

In (14), n means the number of individuals of the initial population, using the parallel method to iterate, it will have n function values, and the minimum is $\min J_m^{(n)}$, $avgJ_m^{(n)}$ is the average of all the values. The introduction of the crossover probability p can make the algorithm only run by p , but not in each iteration of each individual.

There will generate a random probability before the cross operation of each iteration, compare the random probability with the crossover probability p , when the random probability is less than p , then do the cross-operation. Select m individuals from n individuals and compose a new individual, if the fitness of the new one is worse than the worst individual in the original population, then composed a new individual to do the crossover operation.

Use the GuoTao algorithm to optimize the fuzzy C-means clustering algorithm with heterogeneous attributes, the process can be described as:

Step1: initialize the population $P=\{X_1,X_2,\cdots,X_n\}, X_i\in V^n$, and $generation=1$， $X_{worst}=0$，$X_{best}=0$， ε , the maximum number of cycles $MAXgen$.

Step2: decoding, obtain the cluster center from the genotype of each individual chromosome.

Step3: use the fuzzy C-means clustering algorithm with heterogeneous attributes; calculate the membership μ_{ij} for each cluster center according to equation (10), which $i=1,2,\cdots,n;k=1,2,\cdots,c$, then calculate the value of objective function by equation (7).

Step4: calculate the fitness $F(X)=\frac{1}{1+J_m}$ of each individual X_i; compare the fitness of X_i with the fitness of X_{worst} , if the fitness of X_i is greater than the fitness of X_{worst}, replace X_{worst} with X_i; while if the fitness of X_i is less than the fitness of X_{best}, replace X_{best} with X_i.

Step5: when the number of iterations $generation$ is less than the maximum number of cycles $MAXgen$, update the cluster centers of all the individuals.

Step6: select (X_{best},X_{worst}) to satisfy $Better(X_{best},X)=1$ and $Better(X_{worst},X)=0$； respectively calculate the fitness of X_{worst} and X_{best} , if the difference of them is greater than ε, the value of the logic function is 1, otherwise the value is 0.

Step7: when $Better(X_{best},X_{worst})=1$, calculate the crossover probability p by the formula (14), if the random probability is less than p , generate the subspace $V^{'}=\{X_1,X_2,\cdots,X_m\}, X_i\in V^n$; randomly select $X^{'}, X^{'}\in V^{'}$.

Step8: obtain X_{best} : calculate the fitness of $X^{'}$; if $Better(X^{'},X_{worst})=1$, then $X_{worst}=X^{'}$, and replace the fitness of X_{worst} with the fitness of $X^{'}$, if $Better\left(X^{'}, X_{best}\right)=1$, then X_{worst} will be assigned to X_{best}.

Step9: obtain X_{worst}: assume $X_{worst}=0$, calculate the fitness of X_{worst} and the fitness of each individual, if $Better\left(X_{worst}, i\right)=1$, that is to say, X_{worst} is better than i, then replace X_{worst} with i, loop the process until all the individuals are in the comparison complete.

Step10: chose $\left(X_{best}, X_{worst}\right)$ once again, make sure $Better\left(X_{best}, X\right)=1$ and $Better\left(X_{worst}, X\right)=0$: calculate the fitness of X_{worst} and X_{best}, if the difference of them is greater than ε, the value of the logic function is 1, otherwise the value is 0.

Step11: calculate the fitness of the final obtained X_{best}, at the same time $generation+1$.

Step12: judge the evolution conditions, if $generation$ is less than or equal to $MAXgen$, return to Step1, otherwise, exit the evolutionary loop.

III. Intrusion Detection System of clustering with Heterogeneous Attributes

In anomaly detection model of fuzzy C-means algorithm, it mainly composed of three parts [7]: data pre-processor、classifier of fuzzy C-means clustering and anomaly detection system, shown in Figure 1.

When input the network data, the pre-processor will select the attributes of the data, for the data preprocessing, it include data standardization, normalization, etc; the classifier of fuzzy C-means clustering is used to cluster the preprocessed data, and afford the obtained cluster centers to the anomaly detection system; the anomaly detection system is used to determine the data are normal or abnormal in test data.

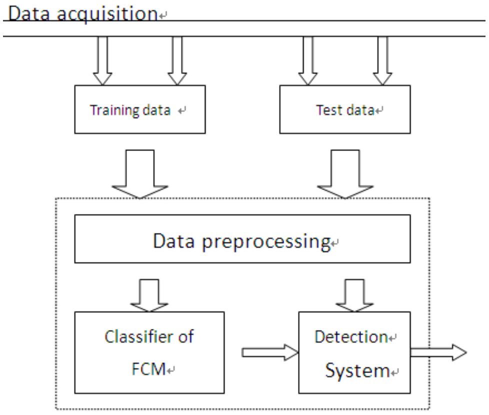

Figure 1 anomaly detection system of fuzzy C-means clustering algorithm

Unsupervised clustering based anomaly detection algorithms have one in common: all of them built on the basis of a hypothesis, that the number of normal data is much larger than the number of abnormal data. When the assumption established, the set can be judged is normal or abnormal according to the size of the cluster. Data in the larger cluster can be judged as normal, and the smaller cluster is often the abnormal data.

IV. Experiments and Analysis of Performance

A. Data Selection

In the research of intrusion detection system, we generally choose to use the network packet of intrusion detection called KDDCUP99, especially the packet of kddcup_data_10percent, it was formed from 10% of kddcup_data packet (about 4.9 million data records)[8]. In the later experiments, we choose 5000 records as a sample set from the 10% test set. In order to satisfy the needs of the two assumptions, that is, 1) the number of the normal data in the practical application is much more than the number of the abnormal; 2) intrusions and normal behavior was really different. Select 5000 records from the entire test as s sample set, 1000 of which are invasion data, it is consistent with the first requirement. The data type and number of the selected sample was shown in Table 1.

TABLE 1
Selection of experimental data

Identification type	number
normal	4000
dos	450
probing	250
r2l	250
u2r	50

The sample as follow:

1)0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,
1.00,0.00,0.11,0.00,0.00,0.00,0.00,0.00,normal.

2)0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,
8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,19,
19,1.00,0.00,0.05,0.00,0.00,0.00,0.00,0.00,normal.

3)0,tcp,http,SF,235,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,29,
29,1.00,0.00,0.03,0.00,0.00,0.00,0.00,0.00,normal.

4)0,tcp,http,SF,219,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
0,6,6,0.00,0.00,0.00,0.00,1.00,0.00,0.00,39,
39,1.00,0.00,0.03,0.00,0.00,0.00,0.00,0.00,normal.

……………………………

1000)0,tcp,http,SF,211,5693,0,0,0,0,0,1,0,0,0,0,0,0,0,
0,0,0,5,5,0.00,0.00,0.00,0.00,1.00,0.00,0.00,
5,255,1.00,0.00,0.20,0.07,0.00,0.00,0.00,0.00,normal.

1001)0,tcp,http,SF,211,2271,0,0,0,0,0,1,0,0,0,0,0,0,0,
0,0,0,15,15,0.00,0.00,0.00,0.00,1.00,0.00,
0.00,15,255,1.00,0.00,0.07,0.07,0.00,0.00,0.00,0.00,n
ormal.

1002)0,tcp,http,SF,511,238,0,0,0,0,0,1,0,0,0,0,0,0,0,0,
0,0,1,3,0.00,0.00,0.00,0.00,1.00,0.00,0.67,1,
255,1.00,0.00,1.00,0.07,0.00,0.00,0.00,0.00,normal.

……………………………

5000)47,tcp,telnet,SF,2402,3816,0,0,0,3,0,1,2,1,0,0,0,
0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,
0.00,10,10,1.00,0.00,0.10,0.00,0.00,0.00,0.10,0.10,buffer
_overflow.

B. Data Standardization

In the process of the data analysis, we must first standardize the data, the common methods are the minimum-maximum standardization, calibration by fractional, and Z-score standardization [9].

In general, clustering algorithm use the method of calculating the distance to cluster the data, however, there are two types of attributes exit: discrete and continuous [10]. For the discrete attributes, we encode the attributes of the data to make them into continuous values, as in this paper, it mainly modified the measure way of similarity of the discrete attributes, there will not standardize and normalize the discrete data; for the continuous attributes, the measure method are not the same, in order to avoid the phenomenon" large number eat the smaller", we generally need to standardize for the values of attributes before clustering to the data set.

Suppose the number of the network connection record of test data set is n, that is, the number of the data objects is n, the attributes vector of each data is written as $X_{ij}(1 \le i \le n, 1 \le j \le 41)$. Standardize the X_{ij} [8]:

$$X_{ij}^{'} = \frac{X_{ij} - \frac{1}{n}(X_{1j} + X_{2j} + \cdots + X_{nj})}{\frac{1}{n}\left(\left|X_{1j} - \frac{1}{n}(X_{1j} + \cdots + X_{nj})\right| + \cdots + \left|X_{nj} - \frac{1}{n}(X_{1j} + \cdots + X_{nj})\right|\right)} \quad (15)$$

Among them, $X_{ij}^{'}$ mean the values after standardizing of X_{ij}.

C. Data Normalization

The process that we need to make the data objects limit in a certain called data normalization. It is convenient for data processing after normalizing, and it ensure to improve the rate of convergence.

We will choose the method of linear function to normalize the data objects X_{ij}, and ensure the values after normalizing to the interval $[0,1]$. Suppose $X_{ij}^{''}$ is the value after normalizing of $X_{ij}^{'}$, its normalization process as follow [8]:

$$X_{ij}^{''} = \frac{X_{ij}^{'} - \min\{X_{ij}^{'}\}}{\max\{X_{ij}^{'}\} - \min\{X_{ij}^{'}\}}$$
$$1 \le i \le n, 1 \le j \le 41 \quad (16)$$

D. Analysis of Performance

The analysis of performance of fuzzy C-means clustering algorithm with heterogeneous attributes

In the fuzzy C-means clustering algorithm, the parameters value as follow: the number of clusters is 2, that is, cluster the data into two categories, normal data and abnormal data; the fuzzy coefficient value is 4; the distance a of two of the data objects which have the difference discrete attributes values is 0.28. Now we will analyze the performance of fuzzy C-means clustering algorithm with heterogeneous attributes, and compare with the original fuzzy C-means clustering algorithm, Table 2 indicate the performances of the two algorithms while the parameters C and m respectively have the same values.

TABLE 2
THE PERFORMANCE COMPARISON OF THE TWO ALGORITHMS

Performance index Algorithms	number of iteration	rate of detection	rate of error detection	values of function
FCM with heterogeneous attributes	79	0.973	0.01925	605.4851
FCM	511	0.975	0.0195	740.0417

Seen from the above table, the rate of detection and the rate of error detection of the two kinds of algorithms are respectively approximately equal, and the number of iterations of fuzzy C-means clustering algorithm with heterogeneous attributes is far less than the number of the original algorithm. It shows that consider the discrete attributes and continuous attributes of the data will improve the speed of convergence and more efficient of detection without affecting the rate of detection and the rate of error detection.

Do the further experiment for the fuzzy C-means clustering algorithm with heterogeneous attributes, Table 3 shows the results of the experiment while m =4 , a =0.28, and it runs 10 times, the simulation results were shown in Figure 2.

TABLE 3
THE PERFORMANCE ANALYSIS OF FUZZY C-MEANS CLUSTERING ALGORITHM WITH HETEROGENEOUS ATTRIBUTES

Performance index Run times	Number of iteration	rate of detection	rate of error detection	values of function
1	77	0.973	0.01925	605.4851
2	56	0.612	0.42275	571.5549
3	84	0.973	0.01925	605.4851
4	78	0.973	0.01925	605.4851
5	75	0.973	0.01925	605.4851
6	100	0.973	0.01925	605.4851
7	77	0.973	0.01925	605.4851
8	57	0.612	0.42275	571.5549
9	75	0.973	0.01925	605.4851
10	86	0.973	0.01925	605.4851
average	77	0.901	0.09995	598.6991

From the above table and the figure, it can be seen that the performance of the algorithm is not stable enough, its number of iterations, the rate of detection, the rate of error detection and the value of function all have a certain volatility, so it indicate that the fuzzy C-means clustering algorithm with heterogeneous attributes is sensitive to initialize and easy to falling into the local optimum.

In order to avoid the above problems, we will combine with global optimization algorithms to optimize the fuzzy C-means clustering algorithm with heterogeneous attributes, and the following experiments will analysis the optimized algorithm.

Experiment II: The performance comparison between the fuzzy C-means clustering algorithm with heterogeneous attributes and the optimal algorithm

For the three problems of fuzzy C-means clustering algorithm with heterogeneous attributes: the initialization sensitive cause the performance of algorithm is not stable enough, easily falling into the local optimum, the value of function is minimum but the rate of detection is not the highest. We will consider the rate of detection, the rate of error detection and the value of function to analysis the optimal algorithm whether it can solve the problems or not.

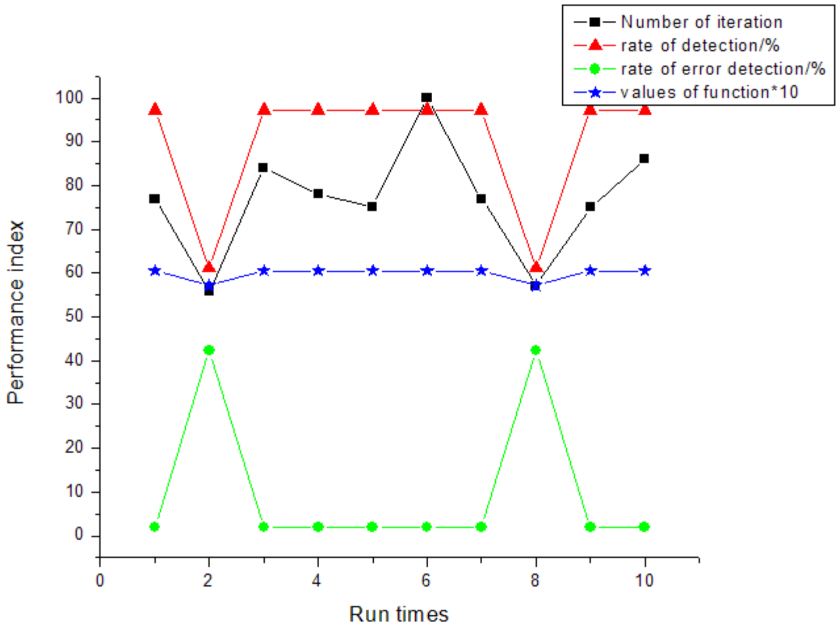

Figure 2 variation trend of performance of fcm

Suppose the datasets is clustered into two categories, that is $C = 2$; the fuzzy coefficient m =4; a =0.28; and the number of iterations is 200. Now, respectively run 10 times of fuzzy C-means clustering algorithm with heterogeneous attributes and the optimal algorithm, the results was respectively shown in Table 4 and Table 5.

From the above tables, the results illustrate the performance of fuzzy C-means clustering algorithm with heterogeneous attributes is not stable enough, that is to say, the robust of the algorithm is not strong. By combining with the GuoTao algorithm and introducing the crossover probability, the performance of algorithm is provided with stability. In the results of running 10 times, the rate of detection and the rate of error detection remained unchanged, with strong stability. For the values of function, as it is operated on all the individuals of the population each time, we use the average of the values of function to measure. It is more clear to show that the instability of the original algorithm and the stability of the optimized algorithm in Figure 3 and Figure 4.

TABLE 4
PERFORMANCE OF FUZZY C-MEANS CLUSTERING ALGORITHM WITH HETEROGENEOUS ATTRIBUTES

Performance index Run times	rate of detection	rate of error detection	values of function
1	0.973	0.01925	605.4874
2	0.973	0.01925	605.4874
3	0.973	0.01925	605.4874
4	0.612	0.42275	571.5569
5	0.973	0.01925	605.4874
6	0.973	0.01925	605.4874
7	0.973	0.01925	605.4874
8	0.973	0.01925	605.4874
9	0.973	0.01925	605.4874
10	0.612	0.42275	571.5569
average	0.901	0.09995	598.7013

TABLE 5
PERFORMANCE OF OPTIMAL ALGORITHM

Performance index Run times	rate of detection	rate of error detection	Average values of function
1	0.973	0.01925	605.4874
2	0.973	0.01925	598.7013
3	0.973	0.01925	591.9152
4	0.973	0.01925	595.3082
5	0.973	0.01925	602.0943
6	0.973	0.01925	605.4874
7	0.973	0.01925	602.0943
8	0.973	0.01925	605.4874
9	0.973	0.01925	595.3082
10	0.973	0.01925	602.0943
average	0.973	0.01925	600.3978

By the performance comparison of original algorithm and the optimized algorithm, we can obtain that the algorithm which use the parallel method have the stable performance; from the experimental data, the algorithm combing with GuoTao algorithm does not fall into the local optimum, the rate of detection was 97.3%, and the rate of error detection was 1.925%, its detection efficiency is stable; in the results of the experiment, the minimum average value of function was 591.9152, and the corresponding rate of detection was 97.3%, it was the highest, the rate of error detection was 1.925%, so the optimized algorithm can solve the problem of the value of function is minimum but the rate of detection is not the highest. It improved the availability of the optimized algorithm.

Get the number of iterations from 10 to 100 for 10 times, then respectively obtain the rate of detection and the rate of error detection, as shown in Table 6. The experimental data in the table are all the average of the obtained results.

TABLE 6
THE RATE OF DETECTION AND THE RATE OF ERROR DETECTION OF FCM WITH HETEROGENEOUS ATTRIBUTES AND THE OPTIMIZED ALGORITHM

Performance index number of iteration	rate of detection of the original algorithm	rate of detection of the optimized algorithm	rate of error detection of the original algorithm	rate of error detection of the optimized algorithm
10	0.9054	0.979	0.10135	0.021
20	0.8288	0.974	0.18145	0.02
30	0.8286	0.973	0.18075	0.0195
40	0.973	0.973	0.01925	0.01925
50	0.8286	0.973	0.18065	0.01925
60	0.8768	0.973	0.09995	0.01925
70	0.973	0.973	0.01925	0.01925
80	0.8286	0.973	0.18065	0.01925
90	0.8768	0.973	0.09995	0.01925
100	0.8768	0.973	0.09995	0.01925

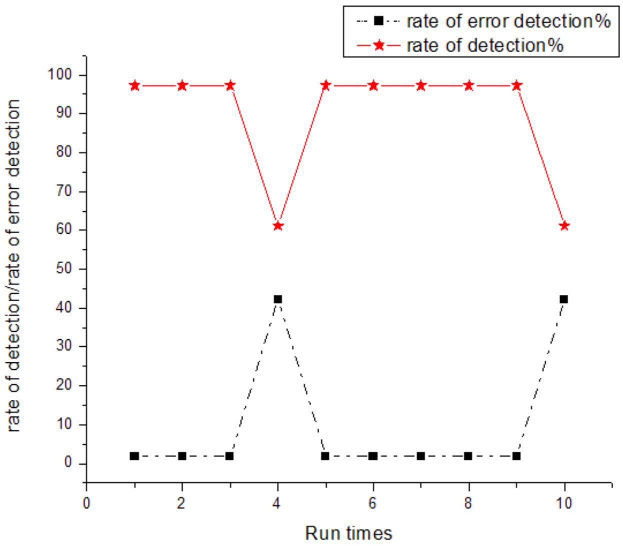

Figure 3 the rate of detection and the rate of error detection of FCM with heterogeneous attributes

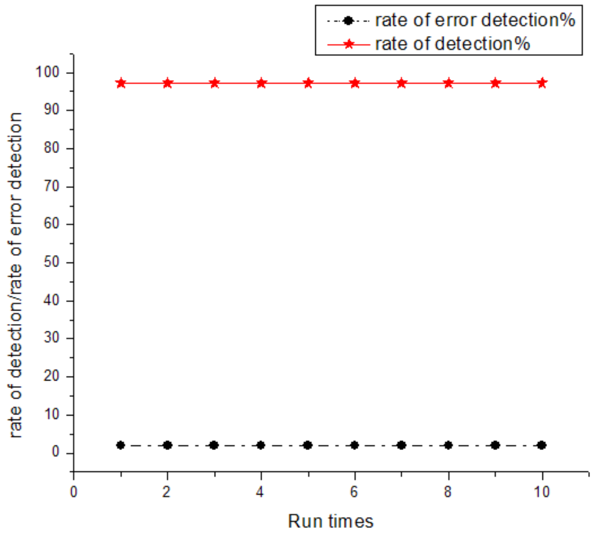

Figure 4 the rate of detection and the rate of error detection of the optimized algorithm

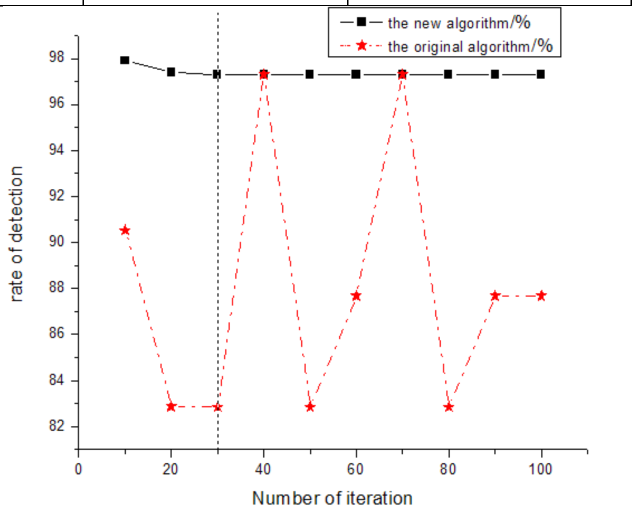

Figure 5 the rate of detection comparison between the original algorithm and the optimized algorithm

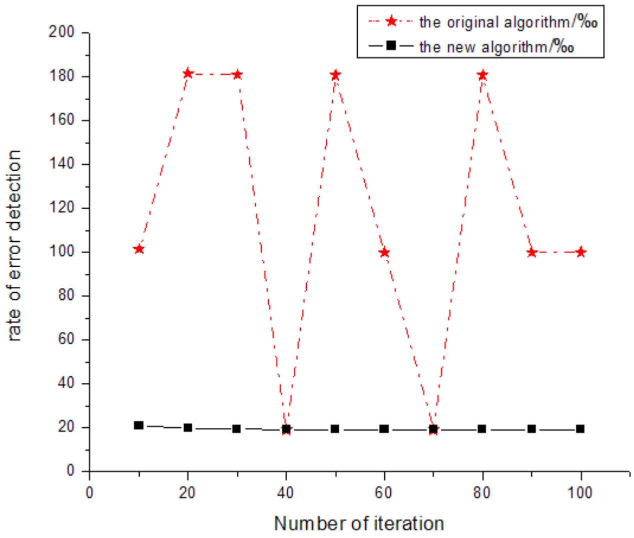

Figure 6 the rate of error detection comparison between the original algorithm and the optimized algorithm

From the above table, when the number of iterations is the same of the fuzzy clustering algorithm and the optimized algorithm, the rate of detection of former is slightly lower than the rate of detection of the latter. With the number of iterations increasing, as the performance of the former algorithm is not stable enough, the average of the rate of detection and the error detection have no the obvious trend, but the optimized algorithm have a certain stability, and when the number of iterations is 30, the algorithm began to converge. The compare of the rate of detection and the rate of error detection between them were shown in Figure 5 and Figure 6

Form the above analysis can determine that the algorithm combined with GuoTao algorithm and introduce the crossover probability can improve the speed of convergence.

V. Conclusion

In order to solve the problem of the experimental data with heterogeneous attributes which are commonly used in intrusion detection, an intrusion detection method based on fuzzy clustering algorithm with heterogeneous attributes is proposed. And we combined with GuoTao algorithm to optimize the model. By the simulation of the data set, the convergence of algorithm is more quickly, the detection efficiency was improved, and the optimized algorithm is more stable, more robust, and can solve the problem of the original algorithm.

Acknowledgment

This paper is sponsored by the Foundation of Jiangxi Educational Committee (GJJ10478), Project supported by the Key Laboratory for Computer Information Processing Technology, Jiangsu Provincial, China (2008-03), and supported by the Hunan Provincial Natural Science Foundation of China (07JJ6113).

References

[1] An JY, Yue GX, Yu F, et al(2006).Intrusion Detection Based on Fuzzy Neural Networks. Lecture Notes in Computer Science Vol.3973,pp. 231-239

[2] Yang DG (2005). Research of the Network Intrusion Detection Based on Fuzzy Clustering: Computer Science, Vol.32 (1), pp86-87

[3] Li J, Gao XB, Jiao LC (2004). A GA-Based Clustering Algorithm for Large Data Sets with Mixed Numerical and Categorical Values:Journal of Electronics and Information Technology, Vol.26(8),pp1203-1209

[4] Guo T (1999). Evolutionary Computation and Optimization. Ph.D. Thesis State Key Laboratory of Software Engineering of Wuhan University, China

[5] Li Y, Kang Z, Liu P(2000). Guo’s Algorithm and Its Application. Journal of WUHan Automotive Polytechnic University, Vol.22(3),pp101-104

[6] He YC, Zhang CJ, Wang PC , et al. (2007).Comparison between Particle Swarm Optimization and GuoTao algorithm on function optimization problems. Computer Engineering and Applications, Vol.43(11),pp100-103

[7] Xiao LZ, Shao ZQ, Ma HH, et al (2008). An Algorithm for Automatic Clustering Number Determination in Networks Intrusion Detection. Journal of Software, Vol.19 (8), pp2140-2148

[8] Wang SJ, Zhang XF (2008). Analysis and Preprocessing of KDDCup99 network data of Intrusion Detection. Science and Technology Information, Vol.15 (9), 407-408

[9] Han JW, Micheline Kamber. Data Mining Concepts and Techniques (2007). China Machine Press

[10] Chen ST, Chen GL, Guo WZ, et al (2010). Feature Selection of the Intrusion Detection Data Based on Particle Swarm Optimization and Neighborhood Reduction. Journal of Computer Research and Development, Vol.47(7),pp1261-1267