Analysis of mathematical modeling in particular clustering process of mixed data

Xu Yuanyuan

(Linyi University,Yishui,Shandong,276400)

Keywords: mixed data; hierarchical difference; clustering algorithm;

Abstract: the analysis method of mathematical modeling in particular clustering process of mixed data is of great significance for improving the ability of data analysis. The traditional method for specific clustering process mathematics modeling of mixed data is based on K-Means clustering algorithm, it is easy to fall into local convergence, and clustering effect is poor. Therefore, the analysis method of mathematical modeling in particular clustering process of mixed data is proposed in the paper based on particle swarm density maximum distance concave function and boundary membership degree feature analysis. The mixed data clustering sample points are divided into k classes according to the degree of similarity to cluster centers, dimensionality reduction is performed for differentiation characteristics of primitive variable data, through searching particles in the space, each particle has the speed, position and fitness, and the optimal solution is found by iteration, preprocessing for data standardization is conducted, data preprocessing includes scale selection of number, type and characteristics, boundary membership feature analysis is processed to achieve mathematical modeling analysis for specific clustering process of mixed data. The simulation results show that, the algorithm has the superior clustering performance of mixed data, good convergence, and great application value.

1 Introduction

With the rapid development of information technology, mixed data time has come, mixed data exists in each application research field and play the function of information transmission, intelligent storage and expression, expression forms such as noise, signal, image and digital [1-3]. Mathematical modeling for specific clustering process of mixed data with massive information is built quickly and accurately, so as to improve the data mining and ability of intelligent analysis processing, has important significance and role in data mining[4-6].

2 Mathematical modeling analysis method principle of clustering process of mixed data

2.1 Description of the K-means algorithm

K-means data clustering method is a data classification method based on statistics[7,8]. The algorithm gives sample set $X=\{x_1,x_2,\cdots,\ x_n\}$ of N data points, find the k cluster centers $\{a_1,a_2,\ldots,\ a_k\}$, different cluster centers has a great influence on the clustering results [9,10]. Therefore, in the analysis process, the sample points according to the degree of similarity and cluster centers to divide into the k class $\{C_1,C_2,\ldots,\ C_k\}$. K-means algorithm for data clustering, dimensionality reduction is processed for differences characteristics of the original variable data, a grid unit is dense in d dimension space, it must also be dense in $d-1$ dimensional space, the following formula is used to normalize:

$$z_{jk}=\frac{y_{jk}-\overline{y_k}}{t_k}\in Z \qquad (1)$$

In the formula, $\overline{y_k}$ is the average of different dimensions vector of mixed data, t_k is corresponding standard deviation, $y_j\in Y,\ j=1,2,\cdots,N,k=1,2,\cdots,P$. For the data

set $X=\{x_1,x_2,\cdots,x_N\}$ in d dimensional data space $S=A_1\times A_2\times\ \cdots\times A_d$, which is mapped to d 1 dimensional space $A_i\left(i=1,2,\cdots,d\right)$, by using the grid limit density difference algorithm, the covariance matrix of different dimension vector is calculated:

$$T=\frac{1}{P}[Z-\overline{Zm})[Z-\overline{Zm}]^T \qquad (2)$$

The distribution of the data point in each A_i dimension is counted, there is always the interval $\left[s_i,s_i'\right]$ $\left(i=1,2,\cdots,d\right)$of relatively dense, the following equation is utilized to process difference decomposition for the relationship between characteristics value and corresponding U value of K-means cluster center:

$$(\mu\, m-T)V=0 \qquad (3)$$

The new clustering center is calculated, according to $c_j'\left(i=1,2,\cdots k\right)$, $c_j'=\frac{1}{\left|C_j\right|}\sum_{X\in C_j}X$, where $\left|C_j\right|$ is the number of data points in the class j. Through the above description to realize the data clustering design based on K-means.

2.2 The standardized processing of data sets

According to the function $J=\sum_{i=1}^{k}\sum_{x\in c_j}d\left(x-c_j\right)$ to process minimum iteration, until the clustering center have no changes or reach the maximum iteration number. The characteristic function described as:

$$\begin{cases}\max\sum_{k=1}^{m}\beta_k-\frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{m}z_jz_k\beta_j\beta_k(y_j\cdot y_k)\\ \qquad s.t.\sum_{k=1}^{m}z_k\beta_k=0 \qquad (4)\\ 0\le\beta_k\le v(y_k)D\ \ k=1,2,\cdots,m\end{cases}$$

The optimal solution of the above planning is described by the following formula:

$$\beta^*=(\beta_1^*,\beta_2^*,\cdots,\beta_m^*)^T \qquad (5)$$

Normalizing each dimension attribute of the samples, which is mapped to range[0,1], the following formula is utilized to express fuzzy optimal classification function:

$$g(y)=\mathrm{sgn}\{(x^*\cdot y)+c^*\} \qquad (6)$$

Effects of mixed data fluctuation of each attribute to the clustering accuracy is reduced to get:

$$m'=\frac{m-\min_m}{\max_m-\min_m} \qquad (7)$$

Mixed data clustering center must be according with the following constraints:

$$x^*=\sum_{k=1}^{m}\beta_k^*z_ky_k \qquad (8)$$

$$b^*=y_i-\sum_{j=1}^{l}y_j\alpha_j(x_j\cdot x_i) \qquad (9)$$

$$j\in\{j\left|0<\beta_j^*<v(y_j)D\right.\} \qquad (10)$$

With the following formula to obtain fuzzy classification optimal classification functions of specific clustering process for mixed data:

$$g(y)=\operatorname{sgn}\{\sum_{k=1}^{m}\beta_k^* z_k L(y,y_k)+c^*\} \quad (11)$$

2.3 Mathematical model optimization solution for specific clustering process of mixed data

Among mixed data, aiming at the defect of K-Means clustering algorithm, like sensitive initial clustering center and easy to fall into local convergence, mathematical model for specific clustering process of mixed data is optimized to solve, and improve clustering process performance for mixed data. Therefore, the analysis method of mathematical modeling in particular clustering process of mixed data is proposed in the paper based on particle swarm density maximum distance concave function and boundary membership degree feature analysis. Particle swarm algorithm through searching particles in the space, each particle has the speed, position and fitness, and the optimal solution is found by iteration, so as to achieve specific clustering of mixed data. Based on the above algorithm design to determine the cluster center. The following two formula are utilized to update the velocity and position of each particle:

$$\begin{aligned} v_t &= wv_{t-1}+c_1 rand_1()\cdot(pbest-x_{t-1}) \\ &\quad +c_2 rand_2()\cdot(gbest-x_{t-1}) \end{aligned} \quad (12)$$

$$x_t = x_{t-1}+v_t \quad (13)$$

Where: v_t is the current velocity of particle, x_t is the current position of a particle. c_1 and c_2 are acceleration constants, the dissimilarity matrix between *n* data points of data set is defined as **s**, it is $n\times n$ matrix:

$$D=\begin{bmatrix} 0 & d(1,2) & d(1,3) & \cdots & d(1,4) \\ d(2,1) & 0 & d(2,3) & \cdots & d(2,n) \\ d(3,1) & d(3,2) & 0 & \cdots & d(3,n) \\ \vdots & \vdots & \vdots & 0 & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \cdots & 0 \end{bmatrix} \quad (14)$$

$$d(i,j)=\sqrt{(\left|x_{i1}-x_{j1}\right|^2+\cdots+\left|x_{id}-x_{jd}\right|^2)} \quad (15)$$

For each particle, its fitness and the experienced best fitness are compared, the obtained two-dimensional space meets:

$$c^*=z_j-\sum_{k=1}^{m} z_k\beta_k L(y_k,y_j) \quad (16)$$

$$j\in\{j\,|\,0<\beta_j^*<v(y_j)D\} \quad (17)$$

The dissimilarity between arbitrary data points x_i and x_j is $s(i,j)$. When PSO enters the time of convergence state, the K-means calculation starts, thus local search is initiated, so as to achive solving for mathematical model, the steps are described :

(1) the population is initialized. The mixed data points are randomly divided into one class, as the initial clustering; particle swarm algorithm is used to optimize the K-means algorithm for specific clustering of mixed data, when PSO enters the convergence state time, the K-means calculation is initiated;

(2) for each $j(j=1,2,\cdots,d)$, 1 dimensional *Gaussian* type fuzzy function u_j is contructed, make

$$u_j(x)=\exp\left(-\frac{(x-\overline{v}_j)^2}{2s_j^{\,2}}\right), x\in A_j \quad (18)$$

(3) mixed data point set $X=\{x_1,x_2,\cdots,x_N\}$ is mapped to the partitioned space grid, based on particle swarm density maximum distance concave function and boundary membership feature analysis to scan each grid unit and count total number of mixed data points, so as to achieve mathematical modeling analysis of specific clustering process for mixed data.

3 simulation experiments and results analysis

In order to test the superior performance of this algorithm, simulation experiment is conducted. Computer simulation experiment platform configuration: Intel Core i5 processor, frequency 2.8 GHz; 4 G memory; Windows 7 Professional Edition 32 bit SP2 operating system, the simulation software is the Matlab of version 2013a. In parameters setting, HTR $h_{TR} = 1/6$, HGD $h_{GD} = 3$, HF $h_F = 2$ test data are from the Internet with CWT200G data binding mode, using the random sampling method to obtain more than 100,000 massive network online monitoring data, the clustering test of clustering center in experimental mixed data as shown below.

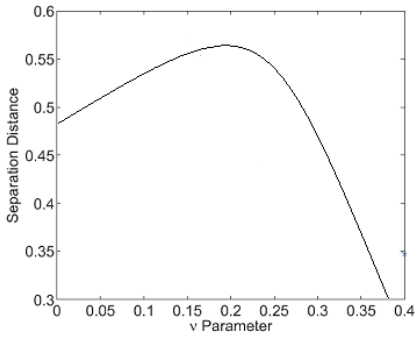

Figure 1 distance of interval

The proposed algorithm, can effectively achieve the specific clustering of mixed data, have high accuracy and good convergence performance. The clustering algorithm and cluster simulation environment are regarded as the overflow, applied to network in online fault detection, the error rate of network fault detection is shown as below.

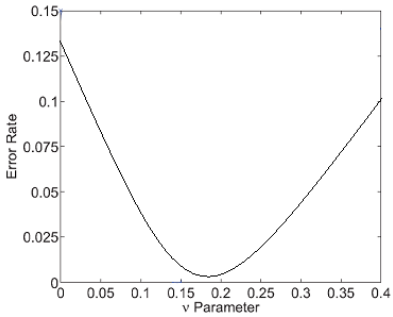

Fig. 2 Comparison of clustering detection error rate

It can be seen from the above results, with the algorithm proposed in this paper can efficiently realize rewriting and recovering for the fault detection data, through effective clustering algorithm to improve the efficiency of fault diagnosis.

4 Conclusion

The traditional method for specific clustering process mathematics modeling of mixed data is based on K-Means clustering algorithm, it is easy to fall into local convergence, and clustering effect is poor. Therefore, the analysis method of mathematical modeling in particular clustering process of mixed data is proposed in the paper based on particle swarm density maximum distance concave function and boundary membership degree feature analysis. The mixed data clustering sample points are divided into k classes according to the degree of similarity to cluster centers, dimensionality reduction is performed for differentiation characteristics of primitive variable data, through searching particles in the space, each particle has the speed, position and fitness, and the optimal solution is found by iteration, preprocessing for data standardization is conducted, data preprocessing includes scale selection of number, type and characteristics, boundary membership feature analysis is processed to achieve mathematical modeling analysis for specific clustering process of mixed data. The simulation results show that, the algorithm has the superior clustering performance of mixed data, good convergence.

References

[1] Cheng Ci, Chai Ruimin. Automatic determining of clustering number [J]. Science and technology information, 2008.14:143.
[2] Jin Wei, Chen Huiping. The K - means algorithm based on hierarchical clustering [J]. Journal of Hehai University (Changzhou Campus), 2007.1:7-10.
[3] Zhang Mingwei, Liu Ying, Zhang Bin, Zhu Zhiliang. A data clustering model based on concept [J]. Journal of software, 2009.9:2387-2396.
[4] Chen Yuanpu, Yin Jianwei, Dong Jinxiang. The clustering analysis based on possibility theory [J]. Computer engineering and applications, 2003.13:85-87.
[5] Qin Yongjun, Liu Xianfeng. The cluster analysis in data mining of [J]. Technology Consulting Herald, 2007.16:28-30.
[6] Bai Xu, Jin Zhijun. Simulation research on optimization model of K- center point cluster algorithm [J]. Computer simulation, 2011.1:218-221.
[7] Li Zhong, Liang Zhijian. An improved text clustering algorithm [J]. Journal of Shaanxi University of Science and Technology: Natural Science Edition, 2008.6:163-166.
[8] Chen Xuejin. Study on the cluster analysis in data mining [J]. Computer technology and development, 2006.9:44-45.
[9] Wang Wende, Gong Jianmin. Fuzzy cluster analysis using EXCEL [J]. Journal of Liaocheng Teachers College: Natural Science Edition, 2000.2:30-33.
[10] Jie Shuipinf. Cluster analysis of multidimensional scaling: problem and solution [J]. Statistics and decision, 2009.11:148-149.