Research and Simulation on effective classification model of nonlinear data

Tang Fenglin, Chen Min

Linyi University, Yishui Shandong 276400,China

Keywords: nonlinear data; classification; clustering;

Abstract: in the classification process of nonlinear data, due to large amount of the data of nonlinearity, the correlation between nonlinear data is reduced, resulting in the classification results of nonlinear data is not ideal. Therefore, this paper proposes a fast clustering algorithm based on improved quantum genetic evolutionary incentive, the algorithm is used to classify nonlinear data effectively. In this algorithm, firstly, high density partition and threshold parameters are utilized to process first cluster partition on nonlinear data sets, a number of clustering are generated; and then the clustering process of samples is regarded as dynamic optimization process of cluster centers, improved quantum genetic algorithm is adopted to search optimal clustering center of each cluster; adaptive mutation operator is introduced to improve the search ability of evolutionary algorithm, so as to enhance the global search capability of the algorithm. The experimental results show that, with the improved algorithm to conduct nonlinear data classification optimization processing, can improve the accuracy of classification, and achieve satisfactory results.

1 Introduction

With the continuous improvement of data processing ability, scale of nonlinear data is increasing [1]. Classification algorithm for nonlinear data classification has become an important method for data acquisition [2]. Classification method for nonlinear data is a core part of data extraction [3]. There are a lot of redundant data in nonlinear data, the data classification method can be utilized to classify a large number of redundant data reasonably, so as to realize the accurate extraction of the target data effectively [4]. The application value of the nonlinear classification method in the field of database, artificial intelligence and other related field is high, concerned by many experts [5]. Therefore, the nonlinear data classification algorithm has become a hot subject to study in data field [6]. The current stage, the nonlinear data classification algorithms mostly includes nonlinear data classification method based on support vector machine, nonlinear data classification method based on wavelet transform and nonlinear data classification method based on data gain algorithm [7]. Among them, the most commonly used is nonlinear data classification method based on wavelet transform algorithm [8]. Because the application scope of nonlinear data classification method is very extensive, it become the key topic to research for many experts, with broad space for development [9, 10].

2 establishment of nonlinear data classification model based on evolutionary incentive quantum genetic algorithm

2.1 improvement for quantum genetic algorithm

The existing quantum binary coding is fixed length coding, the nonlinear data classification problem is complex, the algorithm have slow convergence speed and low precision, when the gene length is short, may even have no solution. Thus, based on the calculation accuracy of adaptive value to calculate the adaptive value of coding length can improve the execution speed and calculation accuracy of the algorithm.

During the effective classification process of nonlinear data, assuming calculation accuracy of gene length independent variable is X_a, that is:

$$X_a = (\mathrm{Max} - \mathrm{Min})/2^N \quad (1)$$

Through the concept of particle distance, assuming in t-th iterations, the particle distance of the j-th particles in the i-th dimension is $d_{ij}(t)$, then

$$d_{ij}(t) = |x_{ij}(t) - g_{besti}(t)| \quad (2)$$

So the particle distance is $HamD_{ij}(t) = |x_{ij}(t), g_{besti}(t)|$, represents the number of different individuals at corresponding position in two chromosome gene position vector, then

$$d'_{ij}(t) = HamD_{ij}(t)/ChormLens \quad (3)$$

The particle in sequence of particle distance of random selection nonlinear data classification model is i', its characteristic response function value is J'_i, then

$$\overline{J'_i} = \frac{1}{m}\sum_{i=1}^{m} J'_{im} \quad (4)$$

Assuming $0 < \left|\overline{J'_i} - k * |J_i|\right| < X(d_m)$, $k = 1,2,\ldots,n$, where $|J_i|$ is mean of median of individual characteristic function values of particle distance. If the particle distance which meet the conditions is the n'-th particle distance in particle distance sequence, average information entropy of particle in the particle distance is:

$$\overline{H(i)} = -\sum_{j=1}^{n} p'_{ij} \ln p'_{ij} = -\sum_{j=1}^{n} \frac{|n'|}{|m|} \ln \frac{|n'|}{|m|} \quad (5)$$

For the particle inertia weight function of arbitrary iteration, then:

$$\begin{cases} H(i) > \overline{H(i)}, \Delta\theta_i = -(w_{start} - w_{end})\left(\frac{t}{t_{max}}\right)^2 + w_{start} \\ H(i) \le \overline{H(i)}, \Delta\theta_i = (w_{start} - w_{end})\left(\frac{t}{t_{max}}\right)^2 + (w_{end} - w_{start})\left(\frac{2t}{t_{max}}\right) + w_{start} \end{cases} \quad (6)$$

Among them, $w_{start} = 0.95$, $w_{end} = 0.15$, $t_{max} = 1000$。

The calculation for incentive value of mutation operator of nonlinear data classification model:

$$r_i(t) = \exp\left(\frac{prog_i(t)}{\sum_{j=1}^{N} prog_j(t)}\alpha + \frac{S_i}{M_i}(1-\alpha)\right) + c_i p_i(t) - 1 \quad (7)$$

S_i is the number of individual of good offspring fitness value after mutation operator i mutated in nonlinear data classification model, $p_i(t)$ is the selection probability of mutation operator i in the t-th propagation, α is the random value weight between $(0,1)$, N is the number of mutation operators, c_i is the penalty factor for mutation operators, that is:

$$c_i = \begin{cases} 0.9 & \text{if } S_i = 0 \text{ and } p_i(t) = \max_{j=1}^{N}\left(p_j(t)\right) \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

The mutation operator in classification model of nonlinear data is assigned a better incentive value, which make local optimal value of variation.

The probability of offspring for mutation operator selection:

$$p_i(t+1) = \frac{r_i(t)}{\sum_{j=1}^{N} r_j(t)}(1 - N\gamma) + \gamma \quad (9)$$

Where γ is the minimum probability of selection, the value is 0.01.

2.2 implementation of nonlinear data clustering process based on evolutionary incentive fast clustering algorithm

During the clustering process of nonlinear data, evolutionary algorithms have common characteristic that the optimal position of individual in the search space is not necessarily unified with the actual position of the treated object sample points, thus, nonlinear data sample point position which is closest to the optimal individual position is utilized for calculating clustering center, and updating the center of cluster of individuals, then:

$$m_i(i, j, t) = \arg\min_{x_k}\{d(xk, p_{best}(i,t))\} \quad (10)$$

The $m_i(i, j, t)$ represents the j-th cluster center point corresponding to i-th nonlinear data

individual in the t-th iteration process, $p_{best}(i,t)$ is the history optimal sample corresponding to i-th nonlinear data individual in t-th iteration, x_k is sample data of nonlinear data.

In the effective classification process of nonlinear data, nonlinear data individual represents sample points of nonlinear data of each cluster center, in order to process iterative optimization for the clustering center point, the objective function f is used to estimate the clustering precision, like:

In the effective classification process nonlinear data, nonlinear data sample points of nonlinear data individual represents each cluster center, in order to the clustering center point for iterative optimization, need f to estimate the clustering precision by using the objective function, is:

$$f = \sum_{j=1}^{k} \sum_{x_k \in c_i} \left| x_k - m_j \right|^2 \qquad (11)$$

Where c_i is the data collection formed by sample x_k with center m_j.

Then the fitness function is:

$$f_s = \frac{1}{1+f} \qquad (12)$$

3 experimental results and analysis

In order to verify the effectiveness of improved algorithms, there is the need for an experiment. To construct the experiment environment by using simulation software matlab7.1. The number of nonlinear data is setting as 1000. Nonlinear data characteristics species number is 20.

10 nonlinear data of different characteristics were selected from the above database randomly, the specific situation of selected nonlinear data can be described by the following table:

Table 1 nonlinear data table of different features

Data attribute number	Data attribute characteristic	The amount of data (million)
1	Property data	1.1
2	Endowment data	1.3
3	Medical data	1.2
4	Education data	1.4
5	Industry reform data	1.2
6	Environment data	1.1
7	Travel data	1.3
8	Beauty data	1.4
9	Fitness data	1.2
10	Infrastructure construction data	1.1

The distribution of the nonlinear data in the above table were arranged, the following Figure was obtained:

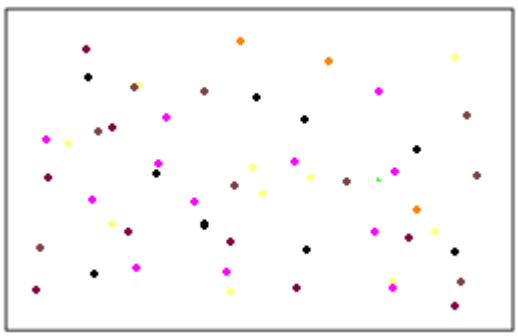

Figure 1 the data distribution map of different attribute

In the situation of many nonlinear data types, 10 times nonlinear data classification were carried out with different algorithms, after analyzing the data, the experimental results is shown in the table below:

Table 2 experimental results of different algorithms of various nonlinear data types

The number of experiments	The classification accuracy of support vector machine algorithm (%)	The classification accuracy of wavelet transform algorithm (%)	The classification accuracy of data gain algorithm (%)	The classification accuracy of improved algorithm (%)
1	79	85	84	96
2	77	82	81	95
3	78	88	86	96
4	77	85	89	96
5	75	88	87	94
6	77	84	88	96
7	76	86	90	98
8	74	87	88	97
9	78	88	92	95
10	77	82	91	96

Through the above experiments it can be learnt that, with the improved algorithm for nonlinear data classification, can improve the efficiency of nonlinear data classification.

4 Conclusion

Aiming at the problems occurred in the classification process of nonlinear data, due to large amount of the data of nonlinearity, such as, the correlation between nonlinear data is reduced, resulting in the classification results of nonlinear data is not ideal. This paper proposes a fast clustering algorithm based on improved quantum genetic evolutionary incentive, the algorithm is used to classify nonlinear data effectively. In this algorithm, firstly, high density partition and threshold parameters are utilized to process first cluster partition on nonlinear data sets, a number of clustering are generated; and then the clustering process of samples is regarded as dynamic optimization process of cluster centers, improved quantum genetic algorithm is adopted to search optimal clustering center of each cluster; adaptive mutation operator is introduced to improve the search ability of evolutionary algorithm, so as to enhance the global search capability of the algorithm. The experimental results show that, with the improved algorithm to conduct nonlinear data classification optimization processing, can improve the accuracy of classification, and achieve satisfactory results.

References

[1] Peng Jing, Tang Changjie, Yuan Chang'an, Li Chuan, Hu Jianjun. A data classification method based on concept similarity [J]. Journal of software, 2007.2:311-322.
[2] Fan Bo, Li Haigang, Guo Qiong. Spatial data classification research for customer segmentation [J]. Journal of systems engineering, 2008.1:87-95.
[3] Wu Fei, Classification method of network data based on real and anonymous address [J]. Journal of Changjiang University natural sciences: Polytechnic volume, 2008.1:244-246.
[4] Liu Hongyan, Chen Jian, Chen Guoqing. Review on data classification algorithm in data mining [J]. Journal of Tsinghua University: Natural Science Edition, 2002.6:727-730.
[5] Longqian, Guo Jichi. The investigation and analysis of 985 University Library self-built database. [J]. The study of Library Science, 2010.18:27-31.
[6] Li Wenjie. Optimal design scheme of large ORACLE database [J]. Technology wind, 2011.19:145.

[7] Ren Li’an, He Qing. A new classification method of mass data [J]. Computer engineering and applications, 2002.14:58-60.
[8] Jiang Mei. Comparison of SciFinder Scholar database and CA on CD database [J]. Modern information, 2006.3:83-86.
[9] Wu Guangchao, Chen Qigang. The combined classification algorithm of imbalance dataset [J]. Computer engineering and design, 2007.23: 5687-5689.
[10] Deng Naiyang, Tian Yingjie. A new method for data mining: support vector machine [M]. Beijing: Science Press, 2004.6.36-39