International Journal of Database Theory and Application
Vol.9, No.1 (2016), pp.227-236
http://dx.doi.org/10.14257/ijdta.2016.9.1.20

A Fuzzy C-Means Clustering Algorithm Based on Improved Quantum Genetic Algorithm

An-Xin Ye[1] and Yong-Xian Jin[2]

[1]*Xingzhi College Zhejiang Normal University, Jinhua 321002, Zhejiang, China*
[2]*College of Mathematics; Physics and Information Engineering; Zhejiang Normal University, Jinhua 321002, Zhejiang, China*
yax@zjnu.cn, jyx@zjnu.cn

Abstract

Aiming at the problem of traditional fuzzy C-means clustering algorithm that it is sensitive to the initial clustering centers and easy to fall into the local optimization, an improved algorithm that combines Improved Quantum Genetic Optimization with FCM algorithm is proposed. In this study, chromosomes are comprised of quantum bits encoded by real number. Chromosomes are renovated by quantum rotating gates and mutated by quantum hadamard gate. The gradients of object function are utilized in adjusting the value of rotating angle by a dynamic strategy. Each chain of genes represents a optimization result, Therefore, a double searching space is acquired for the same number of chromosomes. Experimental results show that the proposed method improves the stability and the accuracy of classification.

***Keywords**: **F**uzzy clustering, Fuzzy c-means, Quantum, Genetic optimization*

1. Introduction

Clustering analysis is one of the important technology in the field of data mining, has been widely used in machine learning and data mining and pattern recognition [1]. Fuzzy c-means (FCM), proposed by Bezdek, Ehrlich, and Full (1984), is the most popular fuzzy clustering method. In FCM, the goal is to minimize the criterion function, taking into account the similarity of elements and cluster centers. It is more useful for data sets that have highly overlapping groups[2]. Since FCM is easily implemented and has obtained satisfactory results in many applications, it has become an important tool for pattern recognition [3, 4]. However, FCM has some shortcomings that have motivated the proposal of alternative approaches for fuzzy clustering, many of which are extensions of FCM. For instance, Zhang, Pedrycz, Lu, Liu, and Zhang (2014) proposed an FCM which uses a genetic heuristic strategy to search for interval weights for the data attributes, to model their different importance for the clustering performance. In another effort to improve clustering quality of FCM, Sabzekar and Naghibzadeh (2013) employed relaxed constraints support vector machines to solve the problem of multiple objects being assigned to clusters with low membership values [5].

Quantum Genetic Algorithm (QGA) is a kind of Genetic Algorithm based on the theory of Quantum computing [6, 7], it intergrate Quantum computation with Genetic Algorithm (GA), with some of the basic elements of Quantum computing, such as Quantum superposition, Quantum entanglement, Quantum interference, optimization Algorithm is applied to the Genetic operation, a probability. It has a good expression of population diversity, achieve rapid convergence in the global search ability [8, 9].

In this paper, the quantum genetic algorithm (QGA) is applied to clustering analysis, and the quantum genetic algorithm is improved, the quantum rotation Angle values by using dynamic change strategy, the chromosome variation by Hadamard gate transform,

ISSN: 2005-4270 IJDTA

the simulation experiments proved that the method has higher clustering accuracy and stability.

2. Fuzzy Clustering

Fuzzy clustering algorithms treat clusters as soft groups to which every data object has a membership degree [10, 11]. These degrees are valued between 0 and 1, with a high value representing a high similarity between the object and the group (Bezdek, *et al.*, 1984). The most well-known of these methods is fuzzy c-means [12, 13].

2.1. Fuzzy C-Means

Fuzzy c-means is a clustering method similar to K-means but the concept of fuzzy theory is incorporated to improve clustering results[14,15]. That is, fuzzy c-means allows that each data point belongs to more than one cluster according to their fuzzy memberships. Assume we are going to classify n data objects into c groups, the objective function used in FCM is defined below:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2 \quad (1)$$

Where $U=[u_{ik}]$ is a c* n matrix of membership degrees,u_{ik} is the membership degree of object k to group i,， which takes its value from the real interval [0, 1]. The higher the value of u_{ik}, the more k belongs to group i. $d_{ik}=\|x_k-v_i\|$ is Euclidean distance between data point k and cluster center i, m is the fuzziness index and its value falls in the range of [1, ∞].

The FCM algorithm is developed with the objective of obtaining a partition matrix $U=[u_{ik}]_{c*n}$ and a set $v=\{v_1,\ldots,v_c\}$ of cluster centers to minimize the objective function J(U,V). By Lagrange multiplier, the necessary conditions for the minimum of J(U,V) are the following updating equations:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}} \quad (2)$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \quad (3)$$

According to Eqs. (2) and (3), the FCM algorithm can be described as follows:

Step 1: Initialize the parameters. Set fix c, fix m, fix T (maximum number of iterations); randomly initialize u_{ik}（i =1,⋯ , c and k=1,⋯, n）of object k to group i, such that $u_{ik} \in [0,1]$, $\sum_{i=1}^{c} u_{ik} = 1$, t=0.

Step 2: Calculate the center vector of cluster i by Eq. (3).

Step 3: Update the fuzzy membership matrix U by Eq. (2).

Step 4: Calculate $J(U,V)$ by using Eq. (1) and check the stop criteria. If $|J(U,V)^{t+1} - J(U,V)^t| < \varepsilon$, then stop the execution; otherwise t=t+1 and return to step 2.

3. Quantum Optimization Algorithm

3.1. Quantum Bit

Quantum mechanics at the atomic level material internal atoms and their components are revealed the structure and properties of elementary particles. The smallest information unit in today's digital computers is a single bit, which at any given time is either in state "1" or "0". The corresponding analogue on a quantum computer is represented by a quantum bit or qubit. Similar to classical bits, a qubit may be in the basis state "0" or "1", but may also be in any superposition of both states. Additionally, the act of measuring (or observing) a qubit will project the quantum system onto one of its basis states.

A quantum bit state $|\varphi>$ can be represented as (4):

$$|\varphi>=\alpha|0>+\beta|1> \quad (4)$$

where α and β are complex numbers, and $|\alpha|^2$ and $|\beta|^2$ represent the probability that the qubit will be found in the "0" state and "1" state, respectively. The sum of $|\alpha|^2$ and $|\beta|^2$ will follow the rule of probability. and can be represented as:

$$|\alpha|^2 + |\beta|^2 = 1. \quad (5)$$

3.2. Quantum Chromosome Encoding

The conventional encoding with binary coding , real number coding and symbolic coding. In quantum optimization algorithm, the use of coding method based on quantum bits, a qubit quantum chromosome with m bit may be defined as:

$$q_j^t = \begin{bmatrix} \alpha_{j1}^t & \alpha_{j2}^t \ldots & \alpha_{jm}^t \\ \beta_{j1}^t & \beta_{j2}^t \ldots & \beta_{jm}^t \end{bmatrix} \quad (6)$$

Where j=1…n,n represents the population size,t represents genetic iteraitons. For example, a quantum chromosome with three qubits is as:

$$\begin{bmatrix} 1/\sqrt{3} & 1/2 & 1/\sqrt{2} \\ \sqrt{2}/\sqrt{3} & -\sqrt{3}/2 & -1/\sqrt{2} \end{bmatrix}$$

The system state is represent as:

$$\frac{1}{2\sqrt{6}}|000>-\frac{1}{2\sqrt{6}}|001>-\frac{1}{2\sqrt{2}}|010>+\frac{1}{2\sqrt{2}}|011>+\frac{1}{2\sqrt{3}}|100>-\frac{1}{2\sqrt{3}}|101>-\frac{1}{2}|110>+\frac{1}{2}|111>$$

This means that the system is in eight states $|000>$, $|001>$, $|010>$, $|011>$, $|100>$, $|101>$, $|110>$, $|111>$,the state probabilistically is 1/24, 1/24, 1/8, 1/8 , 1/12, 1/12, 1/4, 1/4. Only one Q-chromosome such as Eq. 7 is enough to represent eight states, QGA with Q-gene representation has a better characteristic of population diversity than other representations.

4. Fuzzy Clustering Algorithm based on Improved Quantum Genetic Optimization

This paper presents a improved quantum genetic algorithm (IQGA),IQGA uses the principles of GA such as selection, crossover and mutation .and a modified rotation Q-gate strategy is implemented. In the search point target function bigger change to slow down the search, the search point target function change rate is small, then speed up the search. To make the algorithm avoid falling into local optimal solution, mutation operation is realized by using quantum gate Hadamard transform.

4.1. Quantum Coding and the Solution Space Transformation

In this paper, the clustering centers $\vec{Z}_i$ are expressed by quantum bits. α and β compose the quantum bit. Data clustering number is k, dimension is d, the length of each quantum bit chromosome is m(m = k * d), every chromosome q_j^t in the population can be represented as:

$$q_j^t = \begin{bmatrix} \alpha_{j1}^t & \alpha_{j2}^t \ldots & \alpha_{jm}^t \\ \beta_{j1}^t & \beta_{j2}^t \ldots & \beta_{jm}^t \end{bmatrix} \tag{7}$$

Where ,t is number of iterations,j=1,2,…,n,n is population size. The 2m probability amplitudes can range from an m-dimensional space $I^m = [-1, 1]^m$ into the solution space of optimization problems $\Omega = [a_j, b_j]^m$. The corresponding solution space variables are：

$$z_{ji}^t = a_{ji} + (b_{ji} - a_{ji})(1 - \alpha_{ji}^t) \tag{8}$$

or

$$z_{ji}^t = a_{ji} + (b_{ji} - a_{ji})(1 - \beta_{ji}^t) \tag{9}$$

Where, a_{ji} and b_{ji} are minimum and maximum for j chromosome of the ith dimension data.

4.2. Quantum Revolve Gate

Quantum revolve gate realize the process of evolution. So, the design of revolve gate is the most important operation to QGA. Quantum revolve gate can be presented as:

$$\begin{bmatrix} \alpha_{ji}^{t\,'} \\ \beta_{ji}^{t\,'} \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_j) & -\sin(\Delta\theta_j) \\ \sin(\Delta\theta_j) & \cos(\Delta\theta_j) \end{bmatrix} \begin{bmatrix} \alpha_{ji}^t \\ \beta_{ji}^t \end{bmatrix} \tag{10}$$

$$\Delta\theta_j = \theta_{\min} + \frac{f_{\max} - f_x}{f_{\max}} * (\theta_{\max} - \theta_{\min}) \tag{11}$$

Where, $\theta_{\max}$ and $\theta_{\min}$ are respectively maximum and minimum of $\Delta\theta_j$. f_x is the current individual fitness. $f_{\max}$ is the optimum individual fitness. We can see that if the current individual is farther away from the current optimum individual, $\Delta\theta_j$ is relatively large. At the same time we can get a faster search speed. On the contrary, if the current individual is close to the current optimum individual, $\Delta\theta_j$ is relatively small.

4.3. Quantum Mutation

Adopting mutation strategy in genetic algorithm can increase the diversity of population, reduce the probability of premature convergence. In ordinary Quantum genetic algorithm using the Quantum gate (Quantum Not - gate) realize the chromosome variation, we found that the diversity of population increase is not obvious. So we adopt Hadamard gate to chromosome variation in this paper, variation of the results of the qubit Angle clockwise rotation, specific operation is as follows. The variation process is equivalent to clockwise to quantum bits of Angle, that is represented by:

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}\alpha_{ji}^{t}\\ \beta_{ji}^{t}\end{bmatrix}=\begin{bmatrix}\cos(\theta_j+\pi/4-2\theta_j)\\ \sin(\theta_j+\pi/4-2\theta_j)\end{bmatrix} \tag{12}$$

Where, $\theta_j = arctg(\frac{\alpha_{ji}^{t}}{\beta_{ji}^{t}})$ is the current qubit Angle.

4.4. Fitness Function

Data clustering aims to discover meaningful organization of data points in a data set in terms of their similarities and dissimilarities. A good clustering algorithm can classify a set of data points into several distinct clusters such that the members of a cluster are highly similar while the data points belonging to different clusters are dissimilar. So we adopt the fitness function as follows:

$$f=\frac{1}{1+J(U,V)} \tag{13}$$

4.5. Procedure of IQGA

Step 1 Determine the maximum number of iterations Tmax, according to the data sample space distribution, randomly generated K cluster center.

Step 2 Generate a certain number of quantum chromosome, and the initial set of α and β are $1/\sqrt{2}$..

Step 3 Rotate the quantum chromosome according to $\Delta\theta_j$.

Step 4 Perform chromosome mutation according to Hadamard gate with probability Pm.

Step 5 Transform the quantum chromosome space into the solution space of optimization problems.

Step 6 Calculate the fitness function of each chromosome,and choice the next generation of chromosomes with elite selection strategy.

Step 7 Transform the solution space of optimization problems into the quantum chromosome space, and produce a new generation of quantum chromosome.

Step8 Stop the execution if reaches the maximum number of iterations or$|J(U,V)^{t+1}-J(U,V)^{t}|\leq\varepsilon$, otherwise go back to step 3.

5. Experimental Simulations and Analysis

In order to evaluate the clustering performance of IQGA-FCM, it is compared with current algorithms, two clustering algorithms were selected form the literature. The first one is FCM [1] and the second one is GA-FCM [2]. The algorithms were coded under Matlab7.1, and all experiments were run on a personal computer with Pentium 4 (2.8 GHz) running Windows XP. The parameters of experiments were set as follows: the maximum number of iterations Tmax = 200, population size M = 80, crossover probability Pc = 0.45, the mutation probability Pm = 0.005, ε = 0.001.

5.1. Experimental Data Set

The four artificial data sets from UCI database (http://archive.ics.uci.edu/ml) were selected as test problems. The detailed information of these four data sets is listed in Table 1.

Table 1. Dataset Information

Dataset	Number of points	Number of attributes	Actual number of clusters
Iris	150	4	3
Glass	214	9	6
Vowel	990	10	11
Image segmentation	2310	19	7

The other part is composed of stochastic 577 two-dimensional data points data set [16].

5.2. Experimental Testing and Results Analysis

5.2.1. Experiment 1 Clustering Accuracy and Running Time: We run the program 50 times for these four data sets respectively with FCM, GA-FCM and the algorithm stated in this paper. The experimental results are summarized as follows shown in Table 2.

Table 2. Comparing the Performance of Three Algorithms

Algorithm	Dataset	Accuracy	Run time(ms)
IQGA-FCM	Iris	96.17%	298
	Glass	95.31%	450
	Vowel	89.23%	1569
	Image segmentation	85.78%	4560
FCM	Iris	89.16%	210
	Glass	71.25%	343
	Vowel	86.33%	1256
	Image segmentation	52.17%	3212
GA-FCM	Iris	92.56%	134
	Glass	81.27%	198
	Vowel	77.24%	954
	Image segmentation	67.41%	1896

From Table 2, we can see three algorithms has achieved high accuracy on low dimensional data set Iris, while IQGA-FCM is much more advantageous on the clustering effect of more complex data sets. IQGA-FCM adopts fixed rotation calculation to keep the continuity of rotation Angle, it can realize the global optimal solution in the search space. However, the algorithm needs to carry on the solution space transformation, rotation Angle of continuous adjustment, the Hadamard gate chromosome mutations, operation process is more tedious, it runs for a long time compared to other algorithms.

5.2.2. Experiment 2 Clustering Effect: The dataset is made up of 577 two-dimensional random data points. We run the program 10 times with FCM, GA-FCM and IQGA-FCM respectively, we take one of the highest accuracy, the effect shown in the figure below:

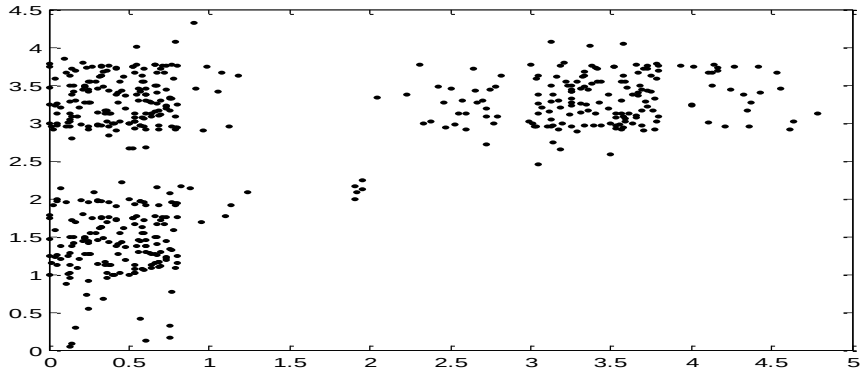

Figure 1. The Original Data Distribution

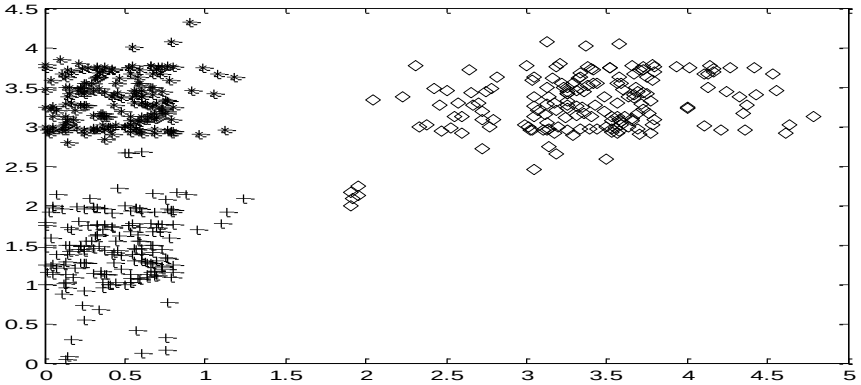

Figure 2. FCM Effect

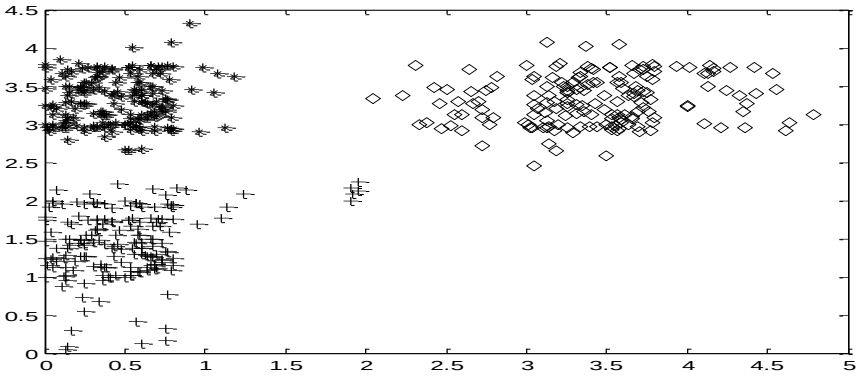

Figure 3. GA-FCM Effect

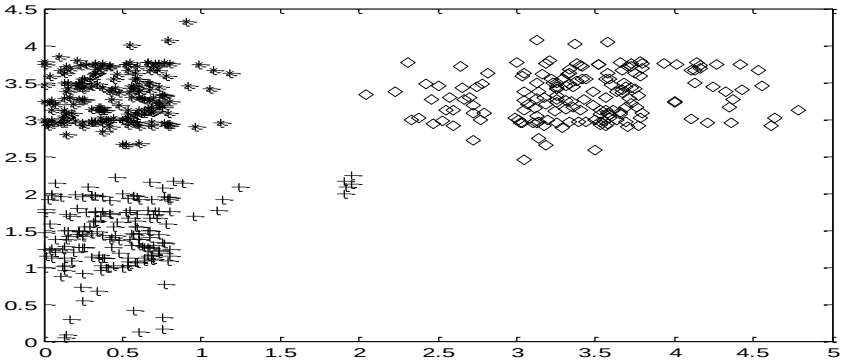

Figure 4. IQGA-FCM Effect

Figure 1. is a distribution of original data, the cluster boundary is not clear. From Figure 2, Figure 3 and Figure 4, we can see that there are two data clustering mistakes running FCM algorithm. This is due to the FCM algorithm traping into local optimum in the process of clustering. GA-FCM algorithm and IQGA-FCM algorithm obtain the same clustering results correctly, this is because the GA-FCM algorithm and IQGA-FCM algorithm adopts qubit chromosome, maintain the genetic diversity, the evolution has achieved the global optimal solution. We can find that IQGA-FCM algorithm in clustering data lower dimension is not obvious advantage compared with GA-FCM algorithm, but considering the experiment 1, the IQGA-FCM algorithm clustering accuracy is higher in dealing with high dimensional data clustering.

6. Conclusion

Due to general FCM clustering algorithm is sensitive to the initial data center distribution, easy to fall into local optimal solution, this paper proposes a clustering algorithm based on improved quantum genetic algorithm, by quantum rotation Angle of dynamic adjustment for continuous optimization of the global optimal solution, Hadamard gate chromosome mutation ensures the diversity of population, the experiment shows that IQGA-FCM algorithm compared with other similar algorithms existing defect at run time, but the accuracy of clustering has obvious advantages, especially in the multi-dimensional complex data clustering that can overcome the problems existing in general FCM clustering algorithm, its global optimization ability is superior clustering effect for complex data sets.

References

[1] S. Z. Selim and M. S. Kamel, “On the Mathematical and Numerical Properties of the Fuzzy C-means Algorithms”, Fuzzy Sets and Systems, vol. 49, **(1992)**, pp. 181-191.
[2] H. Hou and S. Liu, “An Improved Fuzzy C-means Algorithm Based on Genetic Algorithm”, Computer Engineering, vol. 31, **(2005)**, pp. 152-154.
[3] A. K. Jain, M. N. Murty and P. J. Flynn, “Data Clustering”, ACM computing Survey, vol. 31, no. 3, **(1999)**, pp. 264-323.
[4] R. Xu and S. Wun, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Network, vol. 16, no. 3, **(2005)**, pp. 645-678.
[5] D. Pollard, “Quantization and the method of k-means”, IEEE Trans. Inform. Theory, vol. 28, **(1982)**, pp. 199-205.
[6] L. Wang, H. Wu and F. Tang, “Hybrid quantum genetic algorithm and performance analysis”, Contol and Decision, vol. 20, no. 2, **(2005)**, pp. 156-158.
[7] J. Yang and Z. Zhang, “Actuality of research on quantum genetic algorithm”, Computer Science, vol. 30, no. 11, **(2003)**, pp. 13-15.

[8] K. H. Han and J. H. Kim, "Quantum-Inspired Evolutionary Algorithm for a Class of Combinatorial Optimization", IEEE Transactions on Evolutionary Computation, vol. 6, no. 6, **(2002)**, pp. 580-593.
[9] J. Yang and Z. Zhuang, "Research of Quantum Genetic Algorithm and its Application in Blind Source Separation", Journal of Electronics, vol. 20, no. 1, **(2003)**, pp. 62-68.
[10] L. Ertoz, M. Steinbach and V. Kumar, "Fiding clusters of different sizes, shapes, and densities in noise", high dimensional data, Technical Report, **(2002)**.
[11] S. Kisilevich, F. Mansmann and D. Keim, "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos", Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, ACM, **(2010)**.
[12] L. Xia and J. Jing, "A self-adaptive density-based clustering algorithm", Journal of the Graduate School of the Chinese Academy of Sciences, vol. 26, no. 4, **(2009)**, pp. 530-538.
[13] M. Ester, H. P. Kriegel and J. Sander, "A density-based algorithm for discovering cluster in large spatial databases with noise", Proceeding the 2nd International Conference on Knowledge Discovery and Data Mining KDD), Portland, **(1996)**, pp. 226-231.
[14] D. Zhai, J. Yu, F. Gao, L. Yu and F. Ding, "K-means text clustering algorithm based on initial cluster centers selection according to maximum distance", Application Research of Computer, vol. 31, no. 3, **(2014)**, pp. 713-715.
[15] J. Yu and M. Yang, "Optimality test for generalized FCM and its application to parameter selection", IEEE Trans. Fuzzy System, vol. 13, **(2005)**, pp. 164-176.
[16] M. Su and C. Chou, "A Modified Version of the K-means Algorithm with a Distance based on Cluster Symmetry", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 221, no. 6, **(2001)**, pp. 674-680.

Authors

An-Xin Ye, He was born in 1975, an associate professor of Xingzhi College Zhejiang Normal University in China. He got a Master's degree in Computer Science. He is mainly researching on intelligence algorithm, computer science education and data mining *etc*.

Yong-xian Jin, He was born in 1964, a professor of College of Mathematics; Physics and Information Engineering; Zhejiang Normal University in China. He is mainly researching on real-time system, computer science education.