Improved Genetic Fuzzy Clustering Algorithm Based on Serial Number Coding

Yong Zhou[1] Shixiong Xia[1] Dunwei Gong [2] Liangjie Guo[1]

[1]School of Computer Science & Technology, China University of Mining & Technology, Xuzhou 221008, P. R. China

[2]School of Information & Electrical Engineering, China University of Mining & Technology, Xuzhou 221008, P. R. China

Abstract

In illustration to high computational complexity of the genetic algorithm-based FCM clustering algorithm, combining with traditional genetic algorithms and FCM algorithm, an improved GFCA algorithm with serial number coding is proposed in this paper. The simulations of two standard data sets for the proposed algorithm show that the algorithm can reduce the computational complexity under the precondition of no loss in accuracy.

Keywords: Genetic algorithms, FCM, Coding

1. Introduction

Fuzzy c-means clustering algorithm (Fuzzy C - Means FCM) is one of the principal ways to resolve the clustering problem, which is not only used for fuzzy engineering study, but also widely used in other branches of disciplines. FCM have strong local search capability, but is sensitive to initial value. If choosing the initial value improperly, FCM will plunge into local extreme point. Genetic Algorithm (Genetic Algorithm GA) has strong global search capability. In the search process, it reoriented the search space to keep the algorithm to find global optimal solutions or prospective global optimal solution easily. So, many literatures proposed the FCM clustering algorithm based on genetic algorithms and they can get better clustering solutions.

Genetic Algorithm Clustering (GAC) is the algorithm that achieves data clustering by genetic algorithm [1]-[3]. GAC usually encode cluster center, division matrix or encode cluster centers together with the division matrix. The code is the binary code, float code or binary code together with float code. GAC using global search capability of genetic algorithms to overcome the defects that FCM algorithm is sensitive to initial conditions and easy to fall into the local mean point obtaining better result. However, the GAC encoding complex, the search space is huge, the convergence of the algorithm is slow, and need more evolutionary generations, so the algorithm have higher computational complexity. GA has strong global search capability, but their local search capability is poor. FCM's global search ability is poor, but it has strong local search capability. Thus, literature [4] and [5] proposed an algorithm combing GA with FCM. In this paper we call the algorithm Genetic Fuzzy Clustering Algorithm (GFCA). The different between GFCA and GAC is that the cluster center or the division matrix gotten from GAC will be the initial conditions of FCM in GFCA, and then get the result of clustering. So GFCA have two parts, one is GA, the other is FCM. GFCA combines the advantages of GA and FCM to overcome the defects that FCM is sensitive to initial conditions and easy to fall into local extreme point. And FCM enhance the local search capabilities of GA. So it made a good clustering effect. But in GFCA, the part of GA is same to GAC so it also has the shortcomings of having high computational complexity and time cost being big.

Literature [6] proposed a plan to improve the FCM based on genetic algorithm, we call it Improved Genetic Fuzzy C-Means Algorithm (IGFCMA). In IGFCMA, the algorithm optimizes each individual of the group by FCM in every generation of genetic evolution. Some time the optimization happen before genetic manipulation, and some time the optimization happen after genetic manipulation. Then IGFCMA uses the optimized chromosome to replace the original chromosome and goes to the next generation of genetic operation. In this way, GA and FCM combine more closely, and the advantages of GA and FCM can be used more sufficiently. So, the clustering effect is further enhanced, and local extreme optimization is also eliminated. And because every generation has to go through the FCM algorithm optimization, greatly reducing the number of generations when the algorithm converges. For FCM optimization must be taken in every generation, a time expense of every generation is greatly increased. Therefore, the

evolutionary requirements lower generations, but the time cost is high.

The existed FCM clustering analysis algorithms based on genetic algorithms adopted binary or real number coding. These coding methods have big search space and the search space is infinite in real number coding, which can get more accurate clustering results, but evolutionary generations are increased and convergence speed is suspended. So, these coding methods increase the complexity of operation and decoder, and increase the time cost. In GFCA, the function of genetic algorithm is to find suitable initial conditions of FCM algorithm, so as to avoid local extreme solution. This algorithm requires a lower accuracy of genetic algorithm. But the actual cluster center is close to the individual of data. Hence, we can take the appropriate individual combination in data objects that is similar to actual cluster centers as the initial cluster centers for FCM clustering analysis. Comparing with the binary and real number coding, although the accuracy of genetic algorithm is lower, the search space is significantly reduced, and the convergence generation is also reduces. So, the initial cluster centers of FCM can be gotten and avoid local extreme optimal solution. Based on the methodology mentioned above, we encode the cluster centers with the individual serial number, which reduces the search space of genetic algorithm, the number of evolutionary convergence generations. Therefore the genetic algorithms provide suitable initial cluster centers for the FCM algorithm.

2. Improved genetic fuzzy clustering algorithm based on serial number coding

In allusion to the disadvantages of high computational complexity and large time spending in GAC, GFCA and IGFCMA, the improved genetic fuzzy clustering algorithm based on serial number coding is proposed in this paper, which greatly simplifies the coding method, reduces the search space and the number of evolutionary generations and accelerates the convergence rate. But it not increases the computational complexity and greatly reduces the overall time cost.

2.1. Coding method

The coding method in this paper is cluster center coding, and the code is a serial number for the clustering objects. Each individual of clustering objects is sorted in serial number firstly. If there are N individuals in Clustering objects, each individual is named an integer between 1 to *N* as its serial number. So that each individual has only one number, and each number corresponds to only one individual.

clustering objects	Individual 1	Individual 2	……	Individual *n*
Serial number	1	2	……	*N*

Table 1: Serial number for individuals of clustering objects

The gene of chromosome is a serial number. If a clustering object with *n* individuals is aggregated to c clusters, then its chromosome code is as following:

$$a_1, a_2, \cdots a_i, \cdots, a_c \quad a_i \in \{1, 2 \cdots, n\}$$

where a_i is the serial number, c is clusters and *n* is the number of individuals in cluster object..

2.2. Fitness function

In an FCM clustering algorithm, the results of clustering is measured by the objective function. And it is small when the result is good. The objective function of clustering is:

$$\begin{cases} J_m(U,P) = \sum_{i=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m (d_{ik})^2 & m \in [1,\infty] \\ s.t \;\; U \in M_{fc} \end{cases} \quad (1)$$

Where, *U* is fuzzy division matrix, *P* is clustering center, μ_{ik} is the membership of number *K* individual belonging to number *i* class, *m* is The weighted index and d_{ik} is the distance between number *k* individual to the center of number *i* class.

In genetic algorithm, if the value of individual's fitness is greater then the individual is finer and has a greater probability to survive. So, the reciprocal of clustering index is taken as the fitness function to evaluate evolutionary individual is excellent or inferior. The formulation is as the following.

$$f = \frac{1}{1 + J_m(U,P)} \quad (2)$$

2.3. Design genetic operator

2.3.1 Selection operator

The most commonly used method during a selection operation is ratio selection method. That is the probability of individual being selected is proportional to its fitness value. However, in order to avoid the current best individual being destroyed by crossover or mutation operation, the optimal preservation strategy and the ratio selection method is taken in this paper.

2.3.2 Crossover operator

This paper adopts single-point crossover operator. And the crossover probability is p_c. A random number k belonging to [1, c-1] is generated as crossover point, and two new chromosomes are generated after crossover. The crossover process is shown as Fig.1.

$$\begin{cases} a_1, a_2, \cdots a_k \cdots a_c, f_a \\ b_1, b_2, \cdots b_k \cdots b_c, f_b \end{cases} \rightarrow \begin{cases} a_1, a_2, \cdots a_k, b_{k+1} \cdots b_c, f_b \\ b_1, b_2, \cdots b_k, a_{k+1} \cdots a_c, f_a \end{cases}$$

Fig. 1: Crossover operator

2.3.3 Mutation Operator

Each individual generates a random number r between [0, 1], if $r < p_m$ (p_m is a given mutation probability), then the individual mutates. Firstly, a integer k between [1,c] is generated, then a random serial number is used to replace the k^{th} gene. The new serial should differ from the former, if it is the same, and then regenerate a new serial number until it differs from the former. The mutation process is shown as Fig.2.

$$a_1, a_2, \cdots a_k \cdots a_c \rightarrow a_1, a_2, \cdots x \cdots a_c$$
$$x \in \{1,2 \cdots n\} \ \ and \ \ x \neq a_k$$

Fig. 2: Mutation operator

2.4. Steps of algorithm

The steps of the new algorithm GAFCM based on serial number coding are as the following.

Step 1: Set the parameters.

Set the number of clusters c, population size N evolutionary algebra T, crossover probability p_c and mutation probability p_m.

Step 2: Set the serial numbers to the individuals of cluster object.

In practical application, clustering objects are stored in a one dimension array, thus the row number of the array can be regard as the serial number of the objects. Set the serial number from 1 to n.

Step 3: Population initialization.

Randomly generated $c \times n$ serial numbers, and make a chromosome with these c serial numbers, which forms the initialization population.

Step 4: Genetic operation.

The individuals of c serial numbers corresponding to cluster data set are taken as the cluster center and calculate the fitness function with formula (2). Then the selection, crossover and mutation operator are preceded.

Step 5: Optimal preservation.

Pick up the worst individual after mutation operation, which has smallest value of fitness to compare with the best individual of selection operation. If the latter's fitness value is better than the former, then the worst individual is replaced by the best one.

Step 6: Judge the terminating condition of evolution.

Judge the terminating condition of evolution, if it is satisfied then the evolution stops, otherwise go into step 3.In this paper the terminating condition is the number of evolutionary generations. Because it adopts a serial number based coding method, the algorithmic convergence speed is accelerated. The evolutionary generations is set less than 100 in this paper.

Step 7 : Decoding

The best individual generated by genetic algorithm is decoded and the cluster centers are gotten.

Step 8 : FCM algorithm

Set the cluster center generated by genetic algorithm as the initial value of FCM algorithm. And then get the final cluster results by FCM.

3. Simulation

Simulation program is achieved by Matlab 7.0. The operating environment is Windows XP. System hardware configurations are as following: the CPU is the Celeron M, the Frequency is 1.5G and the RAM is 256M.

3.1. Experimental data

This paper uses two kinds of data sets in simulation.

Group I : IRIS data [7]. Iris is Iris plants, Iris data set comes from the data records of the renowned British statisticians R. A. Fisher. Each data include Iris flower four attributes: sepal length, sepal width, petal length and petal width. Three different flowers have 50 sets of data respectively. So the total data set is 150.

Group II : literature [2] used the iron to a certain area actinolite rock samples. The samples contain 19 samples, each samples include 11 indicators that were attached to the five categories. The five categories is (1, 3, 4, 6, 8, 11, 15), (2, 7, 16, 17), (9, 10, 12, 13), (5, 18), (14, 19).

3.2. Compare with the existing algorithms

Experiment one: GAC, IGFCMA, improved GFCA, convergence rate comparisons.

To IRIS data, we carry through cluster analysis using GAC, IGFCMA and improved GFCA respectively. The evolutionary generation is set to 200. Population size is set as N = 50. Mutation probability is 0.9 and crossover probability is 0.1. Observing the convergence of the algorithm, the convergence of the

GAC is shown in Fig. 3, the convergence of IGFCMA is shown in Fig.4, and the convergence of improved GFCA is shown in Fig. 5. The time that is used by each algorithm which evolves 200 generation is listed in Table 2.

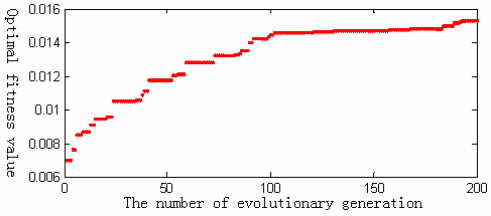

Fig. 3: The convergence curve of GAC

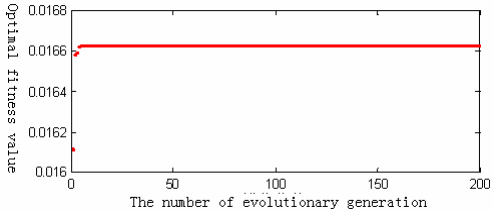

Fig. 4: The convergence curve of IGFCMA

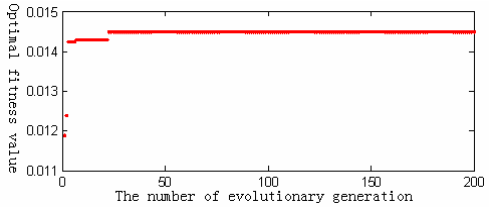

Fig. 5: The convergence curve of improved GFCA

	GAC	IGFCMA	Improved GFCA
Times	44.3440	110.7500	50.2500

Table 2: The time spending of evolving 200 generations

From the experimental results shown above, we can see that IGFCMA and improved GFCA converge fast, and they converge in less than 30 generations. GAC does not converge in 200 generations, but it uses the shortest time in evolving 200 generations. Improved GFCA uses a little more time than GAC. And IGFCMA uses the longest time, it is about 2.5 times of GAC.

It is shown from the experiment that convergence generation of the improved GFCA is greatly reduced comparing with GAC. Improved GFCA needs slightly more generations than IGFCMA. But the computational complexity of IGFCMA is higher than the improved GFCA. So that the time of evolving 200 generation in improved GFCA is less than half of that in IGFCMA. So the improved GFCA converges faster than IGFCMA.

Experimental results show that the GAC needs much more evolutional generations to get a better clustering result, but the improved GFCA only needs dozens of generations. This can significantly reduce time spending. IGFCMA also only needs dozens of generations, but it has high computational complexity, so the overall time spending is higher than improved GFCA.

Experiment two: In this experiment, we use two data set to compare the clustering accuracy and time spending of GAC, GFCA, IGFCMA and improved GFCA.

In this experiment we use GAC, GFCA, IGFCMA and improved GFCA to cluster two data sets 100 times respectively. The number of wrong clustering and the average time are listed in the table 3.

	Group I			Group II		
	The number of generation	The number of errors	average time expense (s)	The number of generation	The number of errors	average time expense (s)
GAC	200	9	45.3717	200	13	14.5377
	500	6	111.4603	500	8	36.1103
	1000	4	221.5923	1000	5	72.0497
	2000	0	443.0023	2000	3	144.1012
GFCA	200	3	48.6223	200	4	15.1067
	500	2	113.3261	500	3	36.9022
	1000	1	222.9938	1000	2	72.8024
IGFCMA	30	0	16.5274	30	0	4.3750
	50	0	27.6115	50	0	7.2040
	100	0	55.2301	100	0	14.2802
Improved GFCA	30	4	11.6856	30	0	2.5079
	50	3	15.0113	50	0	4.0233
	100	3	28.7421	100	0	7.8056

Table 3: The accuracy and average time expense of four clustering analysis algorithms

The experimental results indicate that in the first data set when the number of errors of GAC and improved GFCA is both 4, the average time expense of improved GFCA is 209.9067 seconds less than GAC.

When the number of errors is 3, the average time cost of improved GFCA is 33.6110 seconds less than GFCA. So the experiments show that in the same accuracy rate, improved GFCA can greatly reduce the time cost.

In the second group of data, when the number of errors of GAC and improved GFCA are 2 the average time cost of improved GFCA is 70.2945 seconds less than GFCA. In the same cluster accuracy of GFCA the average time spending of the improved is 140 seconds less than GAC.

In the first data set, IGFCMA has better performance and clustering can eliminate mistakes and time cost is relatively small. Evolutionary time cost of improved GFCA is less than IGFCMA in the same number of generation. But the number of errors is a few more then IGFCMA. In the second group of data, improved GFCA and IGFCMA can eliminate mistakes. But the average time cost of improved GFCA is less than IGFCMA. Improved GFCA shows better performance. This is because the second group data set has less individual, which will highlight the advantages of the serial number coding.

4. Conclusions

In this paper, we proposed improved genetic fuzzy clustering algorithm based on serial number coding, which uses serial number coding on GFCA, reduces the number of evolutionary generation and time expense. Although the accuracy is a little lower, it can provide suitable initial cluster centers to FCM avoiding local optima. Simulation results show that improved GFCA can reduce time expense without loss clustering accuracy rate. Compare with IGFCMA improved GFCA in small group of data can achieve better performance.

Acknowledgement

This work is partially supported by National Natural Science Foundation of China (Grant No. 50674086), National Research Foundation for the Doctoral Program of Higher Education of China (Grant No. 20060290508) and Youth Scientific Research Foundation of China University of Mining and Technology (Grant No. 2006A047).

References

[1] Y. Ou, W. Cheng, Han Ferng-ching, Based on genetic algorithm fuzzy c-means clustering algorithm. *Journal of Chongqing University*, 27(6): 89-92, 2004.

[2] W. Zhang, F.Z. Pan, A genetic algorithm based fuzzy clustering. *Journal of Hubei University (Natural Science)*, 24(2): 101-104, 2002.

[3] H.X. Guo, K.J. Zhu, Based on fuzzy c-means algorithm and the new genetic algorithm clustering method. *South China University of Technology Journal (Natural Science)*, 32(10): 93-96, 2004.

[4] P. Wang, X.l. Zhao, L.H. Wan, The rock structure of the hybrid clustering method based on GA and FCM. *Beijing keji University Journal*, 26 (3): 227-231, 2004.

[5] S.Q. Bai, C.K. Hui, X.J. Wu, A genetic algorithm based fuzzy clustering algorithm and its FCM combination. *East China Shipbuilding Institute Journal (Natural Science)*, 15 (6): 40-43, 2001.

[6] Y. Dong, Y.J. Zhang, C.L. Chang, Improved genetic fuzzy clustering algorithm. *Fuzzy Systems and Mathematics*, 19 (2):123-133, 2005.

[7] http://mlearn.ics.uci.edu/databases/iris/