ICIC Express Letters
Volume **2**, Number **4**, December **2008**
 ISSN 1881-803X
pp. 331–337

FUZZY CLUSTERING ANALYSIS BY USING GENETIC ALGORITHM

Yingjie Wang

College of Information Engineering
Dalian University
Dalian 116622, P. R. China
yingjiehappy@hotmail.com

Received April 2008; accepted August 2008

Abstract. *The fuzzy c-means clustering algorithm (FCM) is applied extensively. However, it can easily be trapped in a local optimum, and also strongly depends on initialization. Therefore, a method of fuzzy clustering by using genetic algorithm is proposed in this paper. Genetic algorithm refers to choose the number of cluster centers and the data that are cluster centers firstly, and clustering analysis is processed by FCM consequently. Experiment results show that the method can search global optimum partly to make the clustering analysis more rational.*
Keywords: Fuzzy c-means clustering (FCM), Clustering analysis, Genetic algorithm (GA)

1. **Introduction.** In the past several years, it has been made a great progress related the research of artificial intelligence. However there are still many important problems which have not been resolved befittingly, in which the fuzzy c-means clustering algorithm [1,2] is applied extensively, but it is sensitive to initial value and easily able to be trapped in local optimum. Therefore, this paper proposes to study fuzzy clustering analysis and then presents a kind of fuzzy clustering method based on the genetic algorithm.

2. **Clustering Analysis.** Clustering analysis is classifying samples according to their similarity by means of unsupervised training. It makes the samples, which have greater similarity, as a class, and occupies the partial area of feature space. The clustering center of each partial area is respectively acting as a representative of the corresponding type. Clustering analysis is not only a kind of impactful method for information compression and extraction, but also it is always the basis of pattern recognition [3]. In spite of clustering may utilize the different algorithm processes, generally, they can be classified three methods: system clustering, progressive dissociation and discrimination clustering.

2.1. **System clustering.** In the beginning of analysis, it is supposed that the number of data points is n and every one becomes a class itself, and then proceeds to cluster them in phase. During this clustering process, the amount of classification decreases step by step until the number of classification amount comes to a proper value. This clustering process is called system clustering. The concrete process of the system clustering realization is: firstly, to choose the nearest two-subgroups combined as a new subgroup, then the number of subgroups becomes $n-1$; secondly, the number of clustering statistic of each $n-1$ subgroups is calculated again, and then the process mentioned above proceeds again, the number of subgroups becomes $n-2$. Similarly, all subgroups are merged finally. According to system clustering, only the distance between the merged subgroup and other subgroup needs to be calculated repetitively, the one between other subgroups does not need to be calculated again. So the computing time is much small by this method, and also it is easy to be implemented. Presently, this kind of method has developed maturely and been applied broadly.

2.2. **Progressive decomposition.** It is supposed that the number of data points is n and all the data points combine into a same class in the beginning of clustering, and then disassemble one by one. In this process, the number of classes will be more and more by disassembling until they reach an appropriate amount which is called progressive decomposition. When using this method, more problems need to be considered, and also this method is not mature enough. So in fact, it was rarely applied on the whole.

2.3. **Discriminating clustering.** It is discriminating clustering which is to determine umpteen clustering centers firstly, and then compares discrete data points in order to decide which class they belong to. The key point of discriminating clustering is the determination of center point of subgroups. There are a lot of methods about the determination of center point, and many people are still researching and improving them, which makes the discriminating clustering is gradually accepted and applied widely.

Above, the system clustering and the progressive decomposition are built on a local consideration. During the realization process, in both methods the system clustering and the progressive decomposition, keeping the global distribution characteristic of groups is out of consideration totally [4,5]. Although, there will be a better result using the discriminating clustering when it is corresponding to the global distribution of clustering centers, which depends on the selection of clustering centers in great measure. Therefore, we need to seek a way, which can considerably keep the global distribution characteristic. Nevertheless, looking for the clustering centers of global distribution characteristic is a good shortcut.

3. **Characteristic of Genetic Algorithm.** American scholar, J. Holland, firstly raised Genetic Algorithm (GA) concept in 1975. It is based on "survival of the fittest" in Darwin's theory of evolution. The basic genetic operations, which are repetitively utilized for the groups possibly containing solution, make the new groups generated then make them evolved constantly. At the same time, the optimization individuals in optimized groups are searched based on the global parallel search technique so as to obtain the global optimum solution fulfilled demands. Comparing with other search algorithms such as random seek, gradient descent and simulated annealing, the major advantage of GA is simple with strong robustness. Since the global parallel search has been done during GA process, the search space is large and it can be constantly adjusted to the direction containing optimization solution, so global optimum or quasi global optimum can be easily found out. Comparing with other approach, the more complicated the problems pending solution are and the more indefinite objects are in this process, the more obvious advantage of GA is. In recent years, it presented the prospect and potentiality that GA has been applied to the fields of combination optimization solving, machine learning and artificial life. Therefore, GA has already become a popular research subject all over the world [6].

GA is a search algorithm based on the genetic mechanisms of natural selection. There is a great deal of difference between GA and traditional search methods: (1) GA does work with a coding modality from parameter set, but not calculate the parameters themselves. (2) Each step in GA is searching from a solution group to another group in solution space rather than from solution to another solution. (3) GA utilizes the probability transition instead of the certainty rule. (4) GA only utilizes the function information of object but not the derivational process and other auxiliary information. GA is provided with the operation parallelism, and it can appraise several data-points at the same time in a complicated search space, which result is propitious to search the global optimum solution in multi-value solution space. It is cared for individual quality of the groups evolved each time in GA process, namely the solution quality of problem, which is different from many optimization algorithms which required recursive information or all information of the

problem such as structure and parameters. Consequently, GA is especially suited to the solution of indefinite problems or nonlinear complex problems.

4. **Fuzzy Clustering Analysis Based on Genetic Algorithm.** Combining respective characteristic of the genetic algorithm and the clustering discrimination algorithm, it is not difficult to be found that the key of the clustering discrimination is how to determine the clustering centers but GA is provided with characteristic of global optimum search. Therefore, firstly, the clustering centers which are keeping with the global characteristic are automatically selected by utilizing GA, and then, other data points are distinguished by the clustering discrimination algorithm. The space clustering analysis result which is corresponding to the global distribution characteristic is produced.

Design of the fuzzy clustering based on GA to resolve the following six problems.

4.1. **Produce initial group.** The initial population consists of initial individuals randomly produced whose number is popsize. A chromosome delegates a data point, namely it contains the location of each data in clustering space. If the number of popsize is too small, the situation will be out of diversity; if the number of popsize is too large, the clustering will spend much time. By experiences, it is shown that the number of popsize should be to range from 30 to 75 [7].

4.2. **Determine coding.** The coding is the primary problem which needs to be solved when utilizing GA and the critical step of GA design. The coding method determines not only the permutation form of individual chromosome (the concrete value of a pair of code), but also the decoding method of the individual from the genotype transform of search space to phenotype of solution space. Also coding method affects the operation of GA, such as crossover operator, mutation operator and so on. Looking for clustering centers in a group of the data points requiring clustering analysis, each data point may be the cluster center or not. It will be faced the problem if each selection needs to be denoted by the coding that: if the coding structure is complex, the calculation will be more complex greatly when the number of data points is large. Accordingly, on design of coding method, it is very convenient to choose coding, decoding and crossover operations and then process binary coding.

The concrete method of coding is that: the length of coding is equal to the number of centers in the group for data points and the value selected from chromosome allele shows the case that whether relative position is selected as the cluster center: 1 shows selected and 0 shows unselected. For example, if the data group is $P\{p_1, p_2, p_3, \cdots, p_{10}\}$ and its chromosome is 0111011010, it does mean that the data points $p_2, p_3, p_4, p_6, p_7, p_9$ are selected as the clustering centers.

For the experiment in this paper, the initial population number popsize is set 50, and the number n of the data point which will be clustered is 50. Each chromosome is composed of 0/1 character string of the binary coding. And then, by utilizing selection operator, crossover operator and mutation operator of GA, it will be processed to calculate optimum number of clustering centers and to search the clustering centers.

4.3. **To determine the fitness function.** It reflects that how strong the capability of individual fitting for circumstance is by the fitness function. According to the result, the probability of individual survival can be commendably controlled, furthermore, to present the law of nature survival of the fittest. Generally, there are different definitions of the fitness function for different problems. For the experiment of this paper, the fitness function is designed as the following.

In the beginning, it will be processed to decode a chromosome, and then to calculate distances of among all data points in initial group.

$$f(p) = \sum_{k=1}^{n} \max\{D_{1k}, D_{2k}, \cdots, D_{nk}\} \tag{1}$$

The parameter D_{nk} shows the distance between data point n and data point k. By this fitness function, it is shown that, the longer the distance from the data point to others is, the higher the adaptability is; and it does easily become clustering center and be propagated into next generation.

4.4. **To determine the operation methods of GA.** In GA operation, the methods which need to be determined are chiefly selection method, crossover method and mutation method.

4.4.1. *Selection operator.* In the course of the biologic inheritance and the natural evolution, the species which possess higher adaptability for life environment, will obtain more opportunity to propagate to next generation, while the lower ones will obtain the less. Imitating this course, GA makes the individuals from group processed "survival of the fittest" by utilizing the selection operator.

How to select operator from the genetic operator is multiplex, but the methods, such as proportionality selection, championship selection and so on, are usually applied. After analyzing this experiment here, 10 percent excellent individuals of group propagate to next generation directly, and the remainders are selected by roulette wheel selection. In this way, it can be ensure that the optimum individuals of next generation could not be worse than the optimum individuals of this generation.

4.4.2. *Crossover operator.* In this experiment, the crossover operator adopts two-point crossing method, for concrete process, see Figure 1. The crossover probability p_c should not be too small because the crossover operation is global search, maybe from 90% to 100%.

A = 0 1 0 0 1 | 0 0 1 1 1 0 0 1 | 0 1 0
B = 1 1 0 1 0 | 0 1 0 1 0 0 1 0 | 1 1 0

(a) The Father String pending two-point crossover operation

A' = 0 1 0 0 1 | 0 1 0 1 0 0 1 0 | 0 1 0
B' = 1 1 0 1 0 | 0 0 1 1 1 0 0 1 | 1 1 0

(b) The Sub String completed two-point crossover operation

FIGURE 1. Two-point crossover operation of GA

4.4.3. *Mutation operator.* In the mutation operation, the position of every gene mutates such as 0-1 or 1-0 with the mutation probability p_m as well as the gene obtain another reasonable value. After the chromosome coding with binary, the mutation is to set the non-value for every bit. The mutation probability has controlled the proportion of the new genes entering into group. Although the mutation probability does only effect on local search ability, if the mutation probability is much small, some beneficial and excellent genes will not be selected; contrarily, if the mutation probability is so large and random variation is so much, the offspring will probably lose excellent character from its parents,

therefore it will lead to losing the capability learning from former search. The experience shows that p_m can be estimated by the following formula [8]

$$p_m \approx 1.75 \Big/ (popsize \times \sqrt{bits}) \tag{2}$$

where, the value "*bits*" is the length of a chromosome.

4.5. **To determine the terminating condition.** The terminating condition of algorithm can be controlled by the convergence degree of solution, and the inheritance can be controlled by the evolution algebra. By this algorithm provided, it is the terminating condition that, to judge the combination between the maximum adaptability and mean adaptability, if the difference between maximum adaptability and mean adaptability is within the allowable range, the algorithm will be terminated [9].

4.6. **To cluster by utilizing fuzzy c-means algorithm [10].** During above processes, the number c of clustering centers and the coordinate of every clustering center are obtained. Filling them into FCM algorithm to cluster for data points of non clustering centers as the following, see formula (3)

$$J(\boldsymbol{U}, \boldsymbol{V}) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ki}^{m} d_{ki}^{2} \tag{3}$$

in which

$$d_{ki}^{2} = \|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T \boldsymbol{A} (\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{4}$$

where, $\boldsymbol{A}$ means a positive matrix; the parameter m $(1 < m < \infty)$ means a fuzzy weighting exponent; the set $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k, \cdots \boldsymbol{x}_n\}$ means data set; the set $\boldsymbol{x}_k \in R^p$, R^p means p-dimensional space; n means the number of data points; c means the number of clusters; and $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_i \cdots, \boldsymbol{v}_c\}$ means clustering centers set. Let $\boldsymbol{U} \in R_{n \times c}$ be a $n \times c$ matrix of fuzzy partition for given training, and $\mu_{ki} \in \boldsymbol{U}$ be a fitness function value from k vector $\boldsymbol{x}_k$ to i clustering center vector, and $\mu_{ki} \in [0, 1]$ and $\sum_{i=1}^{c} \mu_{ki} = 1$.

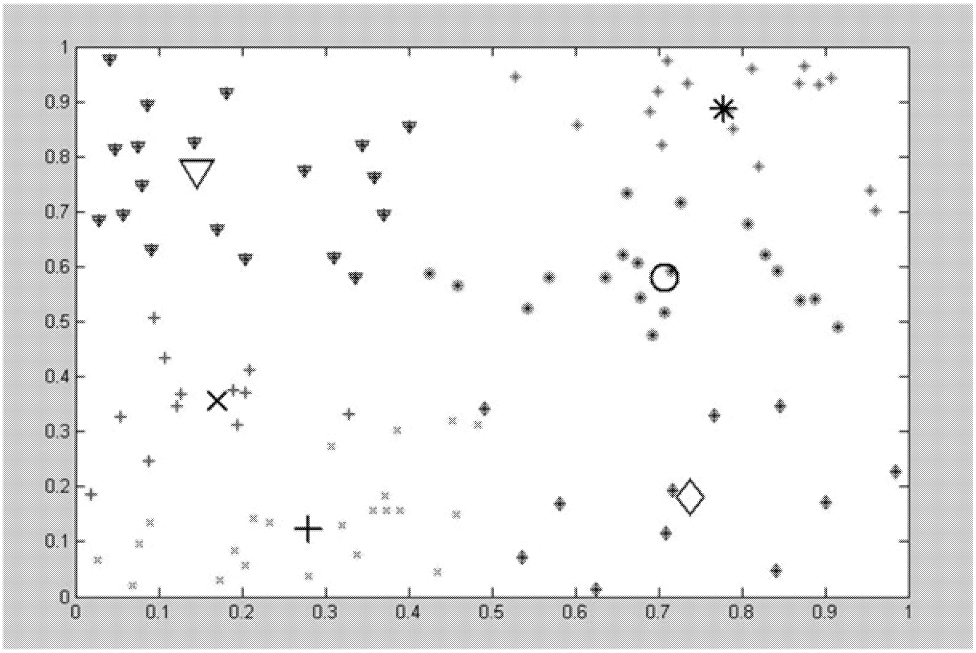

FIGURE 2. The clustering result by FCM based on GA

5. **Experiment Result.** GA is applied 500 times from different groups as the beginning, to obtain the number of the optimum clustering centers and the initial value of clustering centers, and then process clustering analysis with FCM. In the result of discriminated clustering, 80 percent are better than that using FCM alone. Choosing a group of the result as samples to compare, and sample data is processed with clustering object for 100 two-dimensional by two different methods.

Method One: to cluster by utilized FCM based on GA, to find out the clustering centers keeping the global distribution characteristic by utilizing the global optimum search of GA [11].

According to the data provided in this experiment, after iterated the data 125 times, the number of clustering centers is hardly alterative any more. Therefore, the number of optimum clustering centers $c = 6$ and their coordinates are found. The clustering result can be worked out with FCM, as shown in Figure 2.

Method Two: fuzzy c-means method, supposed cluster centers $c = 3$, the clustering result, as shown in Figure 3.

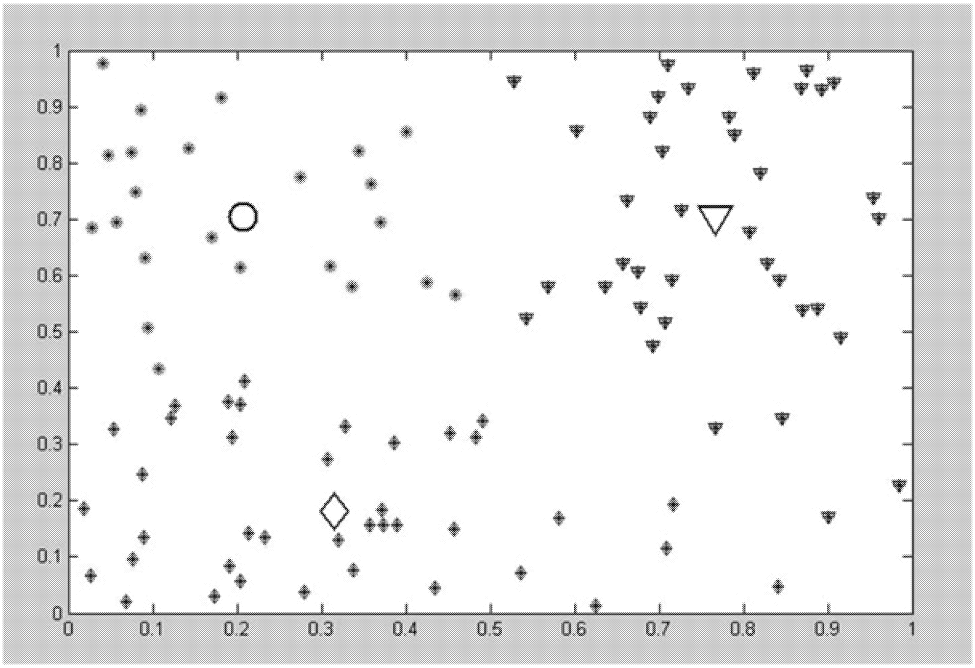

FIGURE 3. The clustering result by FCM alone

By the result of this experiment, it was shown that the space clustering centers keeping the global characteristic can be found by utilizing GA. Combining GA with FCM, with the clustering centers obtained by utilizing GA as initial value of FCM, the global optimum solution can be doubtless obtained after continuing to local search. This method can obtain rational result, not only overcome that FCM algorithm is easily able to be trapped in a local optimum solution and also strongly depend on initial value, but also solve the problem that GA can merely find out approximate optimum solution.

6. **Conclusion.** In this paper, an approach of the fuzzy clustering based on genetic algorithm is proposed, to a certain degree, which overcomes the defects that FCM is sensitive to initial value and it is easily able to be trapped in a local optimum. The practicability of this approach is analyzed in principle, and its practical effect is confirmed by experiment in technical. The experiment results show that the global distribution characteristic of the space clustering centers which are found during the process of the fuzzy clustering analysis by utilizing GA is properly kept, so clustering effect is more rational.

REFERENCES

[1] Y. Shi and M. Mizumoto, An improvement of neuro-fuzzy learning algorithm for tuning fuzzy rules, *Fuzzy Sets and Systems*, vol.118, no.2, pp.339-350, 2001.

[2] L. O. Hall and I. B. Ozyurt, Clustering with a genetically optimized approach, *IEEE Trans*, vol.7, no.3, pp.103-112, 1999.

[3] J. Li, X. Gao and L. Jiao, A new feature weighted fuzzy clustering algorithm, *Acta Electronica Sinica*, vol.34, no.1, pp.89-92, 2006.

[4] Y. Lu and X. Fan, Fuzzy weighting distance and its rationality discussing, *Journal of Northern Transportation University*, Beijing, 1990.

[5] Y. Jiang and Z. Guan, A similarity-based soft clustering algorithm for web documents, *Computer Engineering*, no.2, pp.59-61, 2006.

[6] M. Zhou and S. Sun, *Genetic Algorithm Principle and Application*, National Defence Industry Press, Beijing, 1999.

[7] T. Back, *Evolutionary Algorithms in Theory and Practice*, Oxford University Press, New York, 1996.

[8] Y. Ma, H. Lu and L. He, A genetic encoding scheme for fuzzy clustering, *Journal of Air Force Radar Academy*, vol.16, no.1, pp.40-41, 2002.

[9] X. Dai and M. Li, A dynamic clustering method based on genetic algorithms, *Systems Engineering Theory and Practice*, no.10, pp.108-116, 1999.

[10] T. J. Ross, *Fuzzy Logic with Engineering Applications*, Electronics Industry Press, 2003.

[11] X. Wu and Z. Lin, *Matlab Auxiliary Fuzzy System Designs*, Xidian University Press, 2002.

[12] W. Pedrycz, Collaborative and knowledge-based fuzzy clustering, *International Journal of Innovative Computing, Information and Control*, vol.3, no.1, pp.1-12, 2007.

[13] M. Sato-Ilic, General class of weighted fuzzy cluster loading models, *International Journal of Innovative Computing, Information and Control*, vol.4, no.5, pp.1023-1032, 2008.

[14] K. Zou, J. Hu and X. Kong, The structure optimized fuzzy clustering neural network model and its application, *International Journal of Innovative Computing, Information and Control*, vol.4, no.7, pp.1627-1634, 2008.