A Fuzzy Clustering Method based on Genetic Algorithm

Jung Bok Jo*, Kyeong Hoon Do*, Linhu Zhao**, Mitsuo Gen***

*: Dept. of Computer Engg., Dongseo University, Pusan, Korea

**: Department of Mechanical Engineering and Materials Science, Yokohama National University, 79-5 Tokiwadai, Hodogaya-ku, Yokohama, 240-8501, Japan

***: Dept. of Industrial and Information System Engg., Ashikaga Institute of Technology, Ashikaga, Japan

Tel: +82-51-320-1720, Fax: +82-51-327-8955

E-mail : jobok@dongseo.ac.kr

Abstract: In this paper, we apply to a genetic algorithm for fuzzy clustering. We propose initialization procedure and genetic operators such as selection, crossover and mutation, which are suitable for solving the problems. To illustrate the effectiveness of the proposed algorithm, we solve the manufacturing cell formation problem and present computational comparisons to generalized Fuzzy c-Means algorithm.

1. Introduction

Cluster analysis is one of the major techniques in pattern recognition. It is an approach to divide a data set into some categorically homogenous subsets called "clusters". The ordinary crisp clustering methods restrict that each point of the data set belongs to exactly one cluster, however there may exist points whose lineage are much less evident. Here, the fuzzy set theory provides a means for a more accurate presentation of uncertain or inexact information. Fuzzy clustering has already been widely studied and applied in various areas of pattern search such as, character recognition, image processing, texture classification, medical diagnosis and so on[1][2].

The three main categories of fuzzy clustering, fuzzy clustering based on fuzzy relation, fuzzy clustering based on objective function, and the fuzzy generalized k-nearest neighbor rule are summarized by Yang[2]. The objective function-based methods allow the most precise formulation of the clustering criterion. Many different types of objective function are suggested, and several algorithms are proposed, the Fuzzy c-Means(FCM) algorithm is perhaps the best known one. Typically, local extrema of these objective functions are defined as optimal clustering, and the performance of these methods will mostly rely on the given initial solutions.

On the other hand, genetic algorithms(GAs) based on the mechanism of natural selection and genetics have been widely used for various optimization problems[3][4][9]. Because GAs use population wide search instead of a point search, and the transition rules of GAs are stochastic instead of deterministic, the probability of reaching a false peak in GAs is much less than one in other conventional optimization methods. Although GAs can not guarantee to attain the global optimum in theory, but non-inferior solutions can be obtained at least and sometimes it is possible to attain the global optimum.

Because of the above reason, we create a GA for fuzzy clustering in this paper. We will consider and derive representation structure, initialization procedure, and genetic operators such as selection, crossover and mutation, which are suitable for solving the problems. To illustrate the effectiveness of the proposed algorithm, we solve the manufacturing cell formation problem and present computational comparisons to generalized Fuzzy c-Means algorithm.

2. Fuzzy Clustering

Let the data set $X=\{x_1, x_2, \cdots, x_n\} \in R^p$ be a sub set of the real p-dimensional vector space R^p, and $2 \le c < n$ is an integer. The matrix $\tilde{U}=[\mu_{ik}]$ is

called a fuzzy c-partition if it satisfies the following conditions:

$$\mu_{ik} \in [0,1] \qquad 1 \le i \le c, \quad 1 \le k \le n \qquad (1)$$

$$\sum_{i=1}^{c} = 1, \qquad 1 \le k \le n \qquad (2)$$

$$0 < \sum_{k=1}^{n} \mu_{ik} < n, \qquad 1 \le i \le c \qquad (3)$$

By contrast to the crisp c-partition, μ_{ik} are not restricted to integer value of 0 or 1, they can be fractional. Conditions (2) and (3) ensure that the "total membership" of a point is normalized to 1 and that it cannot belong to more clusters than there exist. The set of all matrixes that satisfy these conditions is called M_c.

Since the number of possible $\tilde{U}$ matrixes that satisfies conditions (1) to (3) is infinite, it is necessary to define an objective criterion for improving an initial partition at first. One of the frequently used criterion is the variance criterion. This criterion measures the dissimilarity between the points in a cluster and its cluster center by the Euclidean distance. Then this distance d_{ik} is as follows[1].

$$d_{ik} = d(x_k, v_i) = \|x_k - v_i\| = \left[\sum_{j=1}^{p} (x_{kj} - v_{ij})^2 \right]^{1/2} \qquad (4)$$

and v_i representing the i-th cluster center is,

$$v_i = \frac{1}{\sum_{k=1}^{n} \mu_{ik}} \sum_{k=1}^{n} (\mu_{ik})^m x_k, \qquad m > 1 \qquad (5)$$

Here v_i is the mean of the x_k m-weighted by their degrees of membership. That means that the x_k with high degrees of membership have higher influence on v_i than those with low degrees of membership.

According to the concept of fuzzy c-partition and the variance criterion, fuzzy clustering amounts to solving the following optimization problem:

$$\min \quad Z_m(\tilde{U}, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m (d_{ik})^2 \qquad (6)$$

Though this is a combinatorial optimization problem which is hard to solve even for rather small values of c and n, it has the advantage that using differential calculus one can determine necessary conditions for local optima. There exist several algorithms which iteratively solves the necessary conditions for local optima, and converges to a local optimum starting from a given initial matrix $\tilde{U}^{(0)}$. One of the best known algorithms for fuzzy clustering is the FCM algorithm. The details of FCM algorithm can be found in [1].

3. Genetic Algorithm

In this section, we will introduce the major components of the GA for fuzzy clustering in turn.

3.1 Representation Structure

A matrix representation perhaps is the most natural representation of a solution for fuzzy clustering. In this representation, a $c \times n$ matrix $\tilde{U} = \mu_{ik}$ representing a fuzzy c-partition should satisfy conditions (1) to (3). For example, a fuzzy 3-partition of 4 points is as follows:

$$\tilde{U}_{3\times 4} = \begin{bmatrix} 0.5 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.3 & 0.2 & 0.4 \\ 0.3 & 0.4 & 0.7 & 0.2 \end{bmatrix}$$

3.2 Initialization Procedure

It is fundamental component for the program to generate a population size of feasible initial chromosomes which satisfy conditions (1) to (3). Define an integer *pop_size* as the size of population. *pop_size* initial chromosomes will be randomly generated through the following steps:

[Step1] Generate c random numbers $r_{1k}, r_{2k}, \cdots, r_{ck}$ in the interval of [0,1] for the k-th point of a chromosome.

[Step2] Calculate $\mu_{ik} = r_{ik} / (r_{1k} + r_{2k} + \cdots + r_{ck})$, for $i = 1, 2, \cdots, c$. Obviously the obtained μ_{ik} will satisfy conditions (1) and (2).

[Step3] Repeat Steps 1 and 2 for $k = 1, 2, \cdots, n$, and produce a chromosome.

[Step4] If the produced chromosome satisfies condition (3) then go to Step 5, else return to Step 1.

[Step5] Repeat above Steps 1 to 4 pop_size times and produce *pop_size* initial chromosomes.

3.3 Evaluation Function and Selection

In this paper, we employ the well-known rank-based evaluation function, and the selection process is based on spinning the roulette wheel[3][4] *pop_size* times and each time one selects a chromosome for the next population.

3.4 Crossover Operation

Define a parameter P_c as the probability of crossover, then $P_c \cdot pop\_size$ chromosomes will undergo the crossover operation in the following way:

[Step1] Generate a random real number r_c in the interval of [0,1] for the given i-th chromosome.

[Step2] Select the given i-th chromosome for crossover if $r_c < P_c$.

[Step3] Repeat Steps 1 and 2 for $i = 1,2,\cdots$, pop_size, and produce $P_c \cdot pop\_size$ parents, averagely.

[Step4] For each pair of parents(matrices $\widetilde{U}_1$ and $\widetilde{U}_2$), the crossover operator on $\widetilde{U}_1$ and $\widetilde{U}_2$ will produce two children $\widetilde{U}_3$ and $\widetilde{U}_4$ as follows:

$$\widetilde{U}_3 = \lambda_1 \widetilde{U}_1 + \lambda_2 \widetilde{U}_2 \;\&\; \widetilde{U}_4 = \lambda_2 \widetilde{U}_1 + \lambda_1 \widetilde{U}_2 \quad (7)$$

where $\lambda_1 + \lambda_2 = 1$ and λ_1 is a random real number generated in the interval of [0,1].

This arithmetical crossover operation ensures that both children are feasible if both parents are.

3.5 Mutation Operation

Mutation is usually defined as a change in a single bit in a solution vector. This would correspond to a change of one element μ_{ik} of a chromosome $\widetilde{U}_i$ for our problem, but this in turn would trigger a series of changes in the elements of the same column in order to maintain the condition (2). To solve the problem, the mutation operation can be proceed in column-wise. Define P_m as the probability of mutation, then $P_m \cdot pop\_size \cdot n$ columns would be selected for mutation. For these selected columns, we perform the mutation operation by using the Steps 1 and 2 of initialization procedure(see 3.2).

4. Manufacturing Cell Formation Problem

Cellular Manufacturing(CM) is an application of Group Technology(GT) in which parts with similar processing requirements and/or geometrical features are classified into part families. The equipment requirements for each part family are determined subsequently or simultaneously and grouped into manufacturing cells[5].

A fundamental issue in CM is the determination of part families and machine cells. This is known as the "cell formation" problem. During the past decade, a number of approaches have been proposed for cell formation. Extensive reviews of the technique are available in the literature[6].

Most of the approaches to cell formation implicitly assume that the information about processing cost, processing time, part demand, etc. are precise. It is also assumed that each part can only belong to one part family. However, there may exist parts whose lineages are much less evident.

Fuzzy clustering provides a solution to such problems. Only a few papers which address the issues of vagueness in cell formation problem[7][8]. Chu and Hayya[7], and Xu and Wang[8] formulated the part family-machine cell formation problem using the generalized fuzzy c-means algorithm, and the major difference between these two paper is that the former uses manufacturing routing data while the latter uses design features in the part families formation.

To demonstrate the effectiveness of the proposed evolution progrm comparatively, let us consider an example from Chu and Hayya[7] that consists of nine machines and nine parts. The data are given in the form of part-machine matrix (8). An entry of 1 for x_{kj} indicates that part k visits machine j.

Parameters selected for this example are as follows: The degree of fuzziness m=2, the maximum generation max_gen=2000; population size pop_size=50; the probability of crossover P_c=0.7; the probability of mutation P_m=0.1; the parameter in rank-based evaluation function a=0.08.

The best solution was obtained in the 1859-th generation, with 3.86 of the value of objective function which is better than 4.36 presented by Chu and Hayya[7]. Tables 1 and 2 list the final membership functions and cluster centers means for machines. According to Table 1 and Table 2, the rearranging rows and columns in part-machine matrix (8) results in matrix (9).

$$x_{kj} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

Table 1. The membership function Options of part.

k	μ_{ik}			Max	Options of part family	
	1	2	3	μ_{ik}	1st choice	2nd choice
1	0.12	0.68	0.20	0.68	2	3
2	0.13	0.19	0.68	0.68	3	2
3	0.92	0.05	0.04	0.92	1	2
4	0.35	0.51	0.14	0.51	2	1
5	0.07	0.88	0.05	0.88	2	1
6	0.01	0.03	0.96	0.96	3	-
7	0.77	0.14	0.09	0.77	1	2
8	0.83	0.12	0.05	0.83	1	2
9	0.02	0.03	0.95	0.95	3	-

$$x_{kj} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad (9)$$

Table 2. The cluster centers means for machines.

j	v_{ij}			Max	Options of machine cell	
	1	2	3	(v_{ij})	1st choice	2nd choice
1	0.02	0.81	0.21	0.81	2	1
2	0.01	0.32	0.96	0.96	3	2
3	0.93	0.02	0.00	0.93	1	2
4	0.73	0.20	0.21	0.73	1	3
5	0.36	0.96	0.02	0.96	2	1
6	0.00	0.02	0.97	0.97	3	-
7	0.93	0.02	0.00	0.93	1	2
8	0.62	0.32	0.06	0.62	1	2
9	0.00	0.02	0.97	0.97	3	-

5. Conclusions

In this paper, we presented a genetic algorithm for fuzzy clustering. Fuzzy clustering is one approach for a more accurate presentation of clustering problems based on uncertain or inexact real-data structures. Generally, fuzzy clustering based on objective function can be formulated a combinatorial optimization problem, and there exist several iterative algorithms. Comparatively, genetic algorithm for this type of combinatorial optimization problem has more chance to attain better solutions because of its way of search, population-wide and stochastic. Our program may be a very effective tool for a number of real-world clustering problems, such as the manufacturing cell formation problem.

Reference

[1] H. J. Zimmermann, "Fuzzy Set Theory and Its Applications", Kluwer Academic Publishers 1991.

[2] M. S. Yang, "A survey of fuzzy clustering", Mathematical and Comput. Modelling Vol.11, No.14, pp.1-16, 1993.

[3] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison -Wesley, 1989.

[4] Z. Michalewicz, "Genetic Algorithm + Data Structures = Evolution Programs", 2nd Ed. Spring-Verlag, New York, 1994.

[5] K. A. Gajendra, R. Divakar, and S. Doug, "A mathematical model for cell formation considering investment and operational costs", E.J.O.R. Vol.69, pp.330-341, 1993.

[6] N. Singh, "Design of cellular manufacturing systems: An invited review", E.J.O.R. Vol.69, pp.284-291, 1993.

[7] C. H. Chu, and J. C. Hayya, "A fuzzy clustering approach to manufacturing cell formation", Int. J. Production Research, Vol.29, pp.1475-1487, 1991.

[8] H. P. Xu, and H. P. Wang, "Part family formation for GT application based on fuzzy mathematics", Int. J. Production Research, Vol. 27, pp.1637-1651, 1989.