Soft Comput (2013) 17:1787–1796
DOI 10.1007/s00500-013-0997-7

METHODOLOGIES AND APPLICATION

A hybrid genetic algorithm–fuzzy *c*-means approach for incomplete data clustering based on nearest-neighbor intervals

Dan Li · Hong Gu · Liyong Zhang

Published online: 24 February 2013

Abstract Incomplete data are often encountered in data sets used in clustering problems, and inappropriate treatment of incomplete data can significantly degrade the clustering performance. In view of the uncertainty of missing attributes, we put forward an interval representation of missing attributes based on nearest-neighbor information, named nearest-neighbor interval, and a hybrid approach utilizing genetic algorithm and fuzzy *c*-means is presented for incomplete data clustering. The overall algorithm is within the genetic algorithm framework, which searches for appropriate imputations of missing attributes in corresponding nearest-neighbor intervals to recover the incomplete data set, and hybridizes fuzzy *c*-means to perform clustering analysis and provide fitness metric for genetic optimization simultaneously. Several experimental results on a set of real-life data sets are presented to demonstrate the better clustering performance of our hybrid approach over the compared methods.

Keywords Fuzzy clustering · Hybrid approach · Incomplete data · Nearest-neighbor interval

1 Introduction

As an important data processing technique, fuzzy clustering partitions the data set into overlapping groups to describe an underlying structure within the data (Hoppner et al. 1999), and in the literature of fuzzy clustering, the fuzzy *c*-means (FCM) algorithm (Bezdek 1981) is a milestone and widely used method (Wei and Fahn 2002; Bandyopadhyay 2005). Like most fuzzy clustering techniques, FCM is designed for handling complete data with their class memberships using the idea of fuzzy set theory. In practice, however, it is not unusual to encounter situations where a data set contains vectors that are missing one or more of the attributes, as a result of failures in data collection, measurement errors, random noise, missing observations, etc. Therefore, some strategies should be employed to handle incomplete data so that FCM is applicable to such incomplete data sets.

In order to reduce the effects of the presence of missing values for clustering, many approaches have been proposed to deal with this problem in pattern recognition. The expectation–maximization (EM) algorithm (Dempster et al. 1977) was a useful approach to modeling and estimation of missing attributes, and was used in probabilistic clustering (Mclachlan and Basford 1988). Subsequently, several methods were proposed for handling missing values in FCM (Miyamoto et al. 1998). One basic strategy is to substitute the missing values by the weighted averages of the corresponding attributes, while another approach is to ignore the missing values and calculate the distances from the non-missing data records. Therefore, these are the two main ideas used to partition incomplete data sets later: imputation, which replaces missing values with estimates that are obtained based on non-missing data, and discarding/ignoring, which is a non-recovery method that ignores incomplete data or only missing attributes. The latter method is applicable only when a small amount of data is missing, and the elimination brings a loss of information. And since in many cases data sets contain relatively large amount of missing data, it is more constructive to consider

Communicated by T. P. Hong.

D. Li (✉) · H. Gu · L. Zhang
School of Control Science and Engineering,
Dalian University of Technology,
Dalian 116024, China
e-mail: ldan@dlut.edu.cn

imputation (Farhangfar et al. 2007). In 2001, Hathaway and Bezdek (2001) proposed four strategies to continue the FCM clustering of incomplete data, in which whole data strategy (WDS) and partial distance strategy (PDS) are discarding/ignoring methods, and optimal completion strategy (OCS) and nearest prototype strategy (NPS) belong to the imputation methods. Besides, based on the Gath and Geva algorithm, Timm et al. (2004) proposed a fuzzy clustering algorithm by taking into account the reasons why attributes were missing. Hathaway and Bezdek (2002) developed an approach for clustering incomplete relational data on the basis of incomplete dissimilarity, in which the data were completed using triangle inequality-based approximation schemes. Honda and Ichihashi (2004) partitioned the incomplete data sets into linear fuzzy clusters by extracting local principal components, and the methods needed no preprocessing of data such as imputation or elimination of incomplete data. Moreover, neural network was another technique that can train incomplete data for clustering (Lim et al. 2005), and statistical representation of missing attributes was studied (Li et al. 2010b). In view of the uncertainty of missing attributes, interval representation of missing attributes based on nearest neighbor information had been proposed and combined into fuzzy clustering in our previous research (Li et al. 2010a), in which the incomplete data set was transformed into an interval-valued one so that the FCM clustering algorithm for interval-valued data could be employed to solve the incomplete data clustering problem. The convex hyper-polyhedrons formed by interval prototypes could present to some extent the shape of clusters and sample distribution of the data set; however, the algorithm performance was sensitive to the upper and lower bounds of the interval representation of missing attributes.

In this paper, we continue to focus on the interval representation of missing attributes, and a hybrid genetic algorithm–fuzzy c-means approach (IGA–FCM) for incomplete data clustering is proposed. Firstly, through the partial distance recommended by Dixon (1979) and used in PDS–FCM (Hathaway and Bezdek 2001), nearest-neighbor information of incomplete data can be obtained, and missing attributes are represented by intervals, named nearest-neighbor interval. Secondly, based on the interval representation of missing attributes, an imputation-based algorithm for incomplete data clustering is proposed. The algorithm involves genetic algorithm (GA) (Davis 1991) which searches for optimal imputations of missing attributes in the corresponding nearest-neighbor intervals to recover the incomplete data set, whereas FCM obtains compact clusters and provides fitness metric for the genetic search. The excellent optimization ability of genetic algorithm can decrease the algorithm sensitivity to the upper and lower bounds of the interval representation, and optimized imputations of missing attributes can be obtained. While the interval representation, as a suitable way to represent the uncertainty of missing attributes, can reduce the search space of GA to subsets that contain nearest neighbors of incomplete data so as to avoid that the improper information misleads the genetic search. Therefore, more satisfying clustering results are likely to be gotten on the basis of the appropriate imputations of missing attributes.

The rest of the paper is organized as follows: The next section presents a short description of the FCM algorithm based on clustering objective function minimization. Section 3 provides the nearest-neighbor interval representation of missing attributes and the hybrid IGA–FCM algorithm, whereas Sect. 4 presents clustering results of several UCI data sets and a comparative study of our hybrid algorithm with other methods for handling missing values in FCM. Finally, conclusions are drawn in Sect. 5.

2 Fuzzy c-means algorithm

The fuzzy c-means (FCM) algorithm partitions a set of complete data $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^s$ into c-(fuzzy) clusters that are characterized by prototypes $V = [v_1, v_2, \ldots, v_c] \in \mathbb{R}^{s\times c}$. The algorithm performs clustering by minimizing the following objective function

$$J(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\|x_k - v_i\|_2^2, \tag{1}$$

taking the constraint

$$\sum_{i=1}^{c} u_{ik} = 1, \quad \text{for } k = 1,2,\ldots,n, \tag{2}$$

into account. Here, $x_k = [x_{1k}, x_{2k}, \ldots, x_{sk}]^{\mathrm{T}}$ is an object datum, and x_{jk} is the jth attribute value of x_k; v_i is the ith cluster prototype, $v_i \in \mathbb{R}^s$; u_{ik} represents the degree of x_k in the ith cluster, $\forall i,k : u_{ik} \in [0,1]$, and let the partition matrix $U = [u_{ik}] \in \mathbb{R}^{c\times n}$; the parameter m influences the fuzziness of the partition, $m \in (1,\infty)$; and $\|\cdot\|_2$ stands for the Euclidean norm.

The necessary conditions for minimizing (1) with the constraint of (2) are the update equations as follows (Bezdek 1981):

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^{m}\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}^{m}}, \quad \text{for } i = 1,2,\ldots,c \tag{3}$$

and

$$u_{ik} = \left[\sum_{t=1}^{c}\left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2}{\|\boldsymbol{x}_k - \boldsymbol{v}_t\|_2^2}\right)^{\frac{1}{m-1}}\right]^{-1}, \tag{4}$$

for $i = 1,2,\ldots,c$ and $k = 1,2,\ldots,n$.

The procedure of FCM is to optimize the clustering objective function (1) by alternating optimization (AO), that is, the minimization steps (3) and (4) are repeated until the change in memberships and/or prototypes drops below a certain threshold ε.

3 Hybrid IGA–FCM approach for incomplete data clustering

3.1 Nearest-neighbor interval determination

As an important issue for incomplete data clustering, missing attribute handling has great effects on clustering performance. Recently, nearest-neighbor (NN) based techniques have been used to impute missing values in pattern recognition. A simple NN imputation was to substitute missing attribute by the corresponding attribute of the nearest neighbor (Stade 1996). And in the widely used *k*-nearest-neighbor imputation (Acuna and Rodriguez 2004), missing attribute was filled by the mean value of the attributes in the *k* nearest neighbors. Subsequently, other than the traditional Euclidean distance, the pseudo-similarity between data was introduced in searching for nearest neighbors, and the effect of pseudo-nearest-neighbor substitution on Gaussian distributed data sets was studied (Huang and Zhu 2002). All the nearest-neighbor based approaches mentioned above can solve incomplete data clustering problem well, however, the numerical imputations developed are unsuitable to represent the uncertainty of missing attributes.

In this paper, we present a nearest-neighbor interval representation of missing attributes by introducing the partial distance (Dixon 1979) between incomplete data and other samples in data set. Let $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ be an s-dimensional incomplete data set which contains at least one incomplete datum with some (but not all) missing attributes, the distance between incomplete datum $\tilde{x}_b$ and instance $\tilde{x}_l$ (incomplete or complete) is given by

$$D_{bl} = \sqrt{\frac{s}{\sum_{j=1}^{s} I_j} \sum_{j=1}^{s} (\tilde{x}_{jb} - \tilde{x}_{jl})^2 I_j}, \qquad (5)$$

where $\tilde{x}_{jb}$ and $\tilde{x}_{jl}$ are the jth attribute of $\tilde{x}_b$ and $\tilde{x}_l$ respectively, and

$$I_j = \begin{cases} 1, & \text{if both } \tilde{x}_{jb} \text{ and } \tilde{x}_{jl} \text{ are nonmissing} \\ 0, & \text{otherwise} \end{cases}, \qquad (6)$$

for $l, b = 1, 2, \ldots, n,$ and $j = 1, 2, \ldots, s$.

And in the extreme case that $\tilde{x}_b$ and $\tilde{x}_l$ have nonmissing values only in different attributes, for example, $\tilde{x}_b = [*, 2, *, 4]^{\mathrm{T}}$ and $\tilde{x}_l = [3, *, 5, *]^{\mathrm{T}}$, where $\tilde{x}_{1b}$, $\tilde{x}_{3b}$ and $\tilde{x}_{2l}$, $\tilde{x}_{4l}$ are missing, the distance between the two data will be set to infinity, and this is helpful to ensure the rationality of nearest neighbor searching. Thus, by using the partial distance (5), the nearest neighbors can be found in a way that uses all available information, including both complete data and non-missing attributes of incomplete data.

In this paper, we consider the case that attributes are missing completely at random (MCAR). In view of the uncertainty of missing attributes, for an incomplete datum $\tilde{x}_b$, we search for its q nearest neighbors to form the interval representation of its missing attribute $\tilde{x}_{jb}$. Let x_{jb}^- and x_{jb}^+ be the minimum and maximum of the neighbors' jth attribute values respectively, therefore, missing attribute $\tilde{x}_{jb}$ can get its interval representation as $[x_{jb}^-, x_{jb}^+]$.

In MCAR problems, it is likely that some attribute j $(j = 1, 2, \ldots, s)$ misses a relatively large proportion of its values. For an incomplete datum $\tilde{x}_b$ who losses its attribute $\tilde{x}_{jb}$, let us consider an extreme case that none of the attribute j in the q nearest neighbors of $\tilde{x}_b$ is non-missing. Then, in this case, the nearest-neighbor interval of $\tilde{x}_{jb}$ will be [0,1] (the data set is normalized before clustering), that is, the interval range is maximum, which is equivalent to take no nearest-neighbor information into account. So, to avoid the case mentioned above and involve nearest neighbor information into missing attribute representation, we search for q nearest neighbors of $\tilde{x}_b$ whose attribute j are non-missing to make possible an estimate of the interval representation of $\tilde{x}_{jb}$.

Clearly, the proposed interval representation integrates informative neighboring relationship into the missing attribute representation, and can introduce pattern similarities in the incomplete data set into the subsequent missing attributes estimation and clustering analysis.

3.2 Hybrid IGA–FCM approach

Genetic algorithms (GAs) (Goldberg 1989; Davis 1991; Deb 2001) are popular search and optimization strategies inspired on the Darwinian theory of evolution. And recently, various hybrid GAs have been proposed to solve optimization or classification problems. In 2006, a simple GA was hybridized with the Wang and Mendel (WM) model to evolve the fuzzy rule base (Chang and Liao 2006). And GAs could also be integrated with *K*-means clustering algorithm and fuzzy decision tree to forecast the future sales (Chang et al. 2009) and construct a decision-making system for data classification (Chang et al. 2010). Besides, Bandyopadhyay and Saha (2008) presented a symmetry-based genetic clustering algorithm that could automatically evolve the number of clusters as well as the proper partition, and Mukhopadhyay et al. (2009) proposed a multiobjective genetic algorithm-based fuzzy clustering algorithm for clustering categorical data sets. And in this paper, we hybridize GA with fuzzy c-means algorithm to solve the problem of incomplete data clustering.

As proposed by Hathaway and Bezdek (2001), when solving the incomplete data clustering problem, missing attributes can be imputed in the way that leaded to the smallest possible value of the clustering objective function. And this is the basic idea of optimal completion strategy fuzzy c-means algorithm (OCS–FCM) mentioned above, which optimizes missing attributes in the entire attribute space by using gradient optimization. Inspired by this work and based on the aforementioned interval representation of missing attributes, we propose a hybrid IGA–FCM algorithm for incomplete data clustering, which is an imputation method. With the interval representation of missing attributes, the imputations of missing attributes can be limited to appropriate ranges, that is, the subsets that contain nearest neighbors of incomplete data rather than the entire attribute space. And this characteristic makes evolutionary algorithms, such as genetic algorithm, good candidates for estimating the appropriate imputations of missing attributes. Thus, in this paper, missing attributes are viewed as variables to recover the incomplete data set, and our hybrid framework combines FCM that performs clustering analysis on the recovered data set and provides fitness metric for genetic optimization, and genetic algorithm that guides the search of missing attributes in the corresponding nearest-neighbor intervals to make the clustering objective function achieve its minimum. Finally, clustering results of FCM based on the optimized imputations of missing attributes can be obtained.

In the following, we will describe the design of our genetic algorithm, as well as the procedure of the proposed hybrid IGA–FCM approach for incomplete data clustering.

3.2.1 Genetic representation

In the genetic clustering applications, binary and real parameter representations are commonly used (Liu et al. 2004; Mukhopadhyay et al. 2009). Compared with the binary-coded GAs, real representations are believed more practical due to their consistency with the real world's number system and, thus, are convenient for further processing (Su et al. 2009). In this paper, the chromosome is composed of a sequence of real valued numbers that represent the imputations of missing attributes in corresponding nearest-neighbor intervals.

Let E be the population and M be the population size, and for a set of s-dimensional incomplete data $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ with h missing attributes, the pth individual chromosome of the population at generation t has h components, i.e.,

$$E_p(t) = \left[e_{p,1}, e_{p,2}, \ldots, e_{p,h}\right], \tag{7}$$

where $e_{p,g} (1 \le p \le M, 1 \le g \le h)$ is the gth missing attribute imputation in the pth individual. Note that the coding scheme only encodes the missing attributes, which is helpful to reduce the computational cost of genetic algorithm. For convenience, we sort and renumber the nearest-neighbor intervals of missing attributes by their appearance order in the data set, and Fig. 1 gives an example on a four-dimensional data set, in which missing attributes are denoted by *. Therefore, in the genetic process, each element $e_{p,g}$ $(1 \le p \le M, 1 \le g \le h)$ in chromosome $E_p(t)$ should satisfy its interval constraint, that is, $e_{p,g} \in [e_g^-, e_g^+]$.

The initial population of solutions can be generated randomly in the corresponding nearest-neighbor intervals, that is

$$\begin{aligned} E_p(1) &= [e_{p,1}, e_{p,2}, \ldots, e_{p,h}] \\ &= \left[\text{rand}(e_1^-, e_1^+), \text{rand}(e_2^-, e_2^+), \ldots, \text{rand}(e_h^-, e_h^+)\right], \\ &\text{for } p = 1, 2, \ldots, M. \end{aligned} \tag{8}$$

And this initial population of solutions is allowed to evolve to achieve optimized individual using a set of genetically motivated operations.

3.2.2 Fitness function

For a set of incomplete data $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ with h missing attributes, using each of the chromosome $E_p(t) = [e_{p,1}, e_{p,2}, \ldots, e_{p,h}] (1 \le p \le M)$, we can then obtain a recovered complete data set $X = \{x_1, x_2, \ldots, x_n\}$, therefore, FCM can be directly applicable.

In the proposed hybrid algorithm framework, the use of FCM can perform clustering analysis on the recovered data sets and provide fitness metric for genetic optimization simultaneously. Here, we use the reciprocal of clustering objective function as fitness function to evaluate the optimality of each chromosome $E_p(t)$ $(1 \le p \le M)$ at generation t:

$$\text{fitness}(E_p(t)) = \frac{1}{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|x_k - v_i\|_2^2}. \tag{9}$$

It is easy to see that the chromosome is evaluated according to clustering objective function (1), and with the guide of these fitness values, the genetic mechanism will search for optimized missing attribute imputations in the corresponding nearest-neighbor intervals to make the clustering objective function achieve its minimum.

3.2.3 Genetic operators

In genetic algorithm, selection, crossover, and mutation are the basic operators that provide an effective search

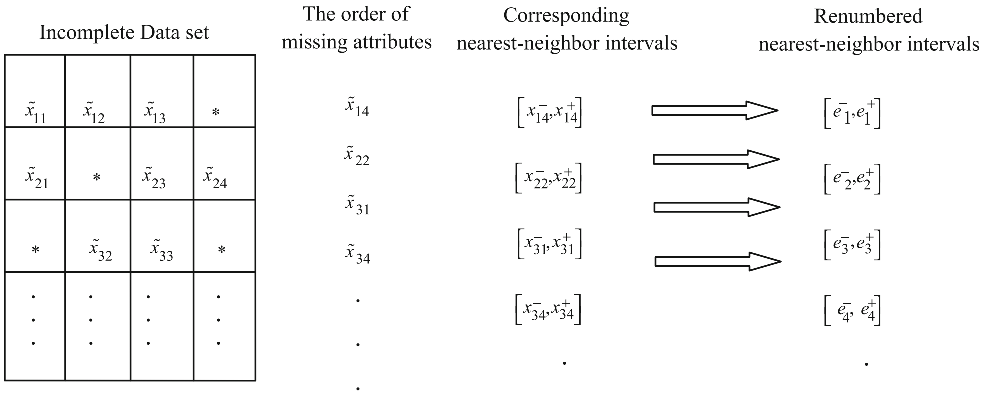
Incomplete Data set
$\tilde{x}_{11}$ $\tilde{x}_{12}$ $\tilde{x}_{13}$ *
$\tilde{x}_{21}$ * $\tilde{x}_{23}$ $\tilde{x}_{24}$
* $\tilde{x}_{32}$ $\tilde{x}_{33}$ *
The order of missing attributes
$\tilde{x}_{14}$
$\tilde{x}_{22}$
$\tilde{x}_{31}$
$\tilde{x}_{34}$
Corresponding nearest-neighbor intervals
$[x_{14}^-, x_{14}^+]$
$[x_{22}^-, x_{22}^+]$
$[x_{31}^-, x_{31}^+]$
$[x_{34}^-, x_{34}^+]$
Renumbered nearest-neighbor intervals
$[e_1^-, e_1^+]$
$[e_2^-, e_2^+]$
$[e_3^-, e_3^+]$
$[e_4^-, e_4^+]$

Fig. 1 An example of renumbered nearest-neighbor intervals on a four-dimensional data set

technique and improve a population of potential solutions iteratively. And the genetic operators adopted here are as follows:

(i) Selection: This operation is a mechanism related to individual fitness, in which chromosomes from the parent population are selected according to their selection probability to replicate and form offspring chromosomes. Generally speaking, common selection schemes include roulette wheel selection (Michalewicz 1994), tournament selection and rank-based selection (Blickle and Thiele 1996). And the most used selection mechanism is the roulette wheel selection, by which the individuals are selected by spinning a roulette wheel with its slots sized according to their fitness values (Silva et al. 2000). Our GA employs roulette wheel strategy for implementing the selection scheme, and after ascending sorting the individuals according to their fitness values, the selection probability of each individual is defined as follows (Zhu et al. 2004)

$$P_{\text{selection}}(E_p(t)) = \frac{2p}{M(M+1)}, \quad \text{for } p = 1, 2, \ldots, M. \tag{10}$$

(ii) Crossover: The mechanics of the crossover operation is to change the genetic materials of the individuals by swapping some information between a pair of chromosomes. In natural evolution, a pair of parent chromosomes may generate several offspring, and there also exists competition among the offspring. Inspired by this phenomenon, a crossover operator based on competition and optimal selection (Leung et al. 2003; Ren and San 2007) is used here.

Let the parent chromosomes be $E_p(t) = [e_{p,1}, e_{p,2}, \ldots, e_{p,h}]$ and $E_f(t) = [e_{f,1}, e_{f,2}, \ldots, e_{f,h}]$ $(1 \le p, f \le M, p \ne f)$ at generation t, and four offspring as follows are generated at first:

$$offsp_1 = \frac{E_p(t) + E_f(t)}{2}, \tag{11}$$

$$offsp_2 = \frac{(E_{\max} + E_{\min})(1-w) + (E_p(t) + E_f(t))w}{2}, \tag{12}$$

$$offsp_3 = E_{\max}(1-w) + \max(E_p(t), E_f(t))w, \tag{13}$$

$$offsp_4 = E_{\min}(1-w) + \min\left(E_p(t), E_f(t)\right)w, \tag{14}$$

where crossover factor $w \in [0,1]$ denotes the weight to be determined by users; $E_{\max} = [e_1^+, e_2^+, \ldots, e_h^+]$, $E_{\min} = [e_1^-, e_2^-, \ldots, e_h^-]$; and vectors $\max(E_p(t), E_f(t))$ and $\min(E_p(t), E_f(t))$ are formed by the maximum and minimum of corresponding elements in $E_p(t)$ and $E_f(t)$ respectively. Subsequently the two offspring with higher fitness values can be chosen to substitute the parent chromosomes in new population. Obviously, the value of w has no effect on $offsp_1$, which always generates offspring at the center of the parent chromosomes $E_p(t)$ and $E_f(t)$. As for the other three offsprings, when $w \to 1$, $offsp_2 \to offsp_1$, while (13) and (14) result in searching around the maximum and minimum genes of $E_p(t)$ and $E_f(t)$; and when $w \to 0$, the crossover operation tends to develop offsprings at the center and boundary of nearest-neighbor intervals E_{min} and $E_{\max}$. Thus, the smaller the value of w is, the more important the E_{min} and $E_{\max}$ are to the generation of offsprings. And when $w = 0.5$, the importance degree of the boundary of nearest-neighbor intervals ($E_{\min}$ and $E_{\max}$) and parent chromosomes ($E_p(t)$ and $E_f(t)$) are equal. So, the solutions may spread all over the nearest-neighbor intervals, and the above crossover operator can generate superior offspring than arithmetic crossover or heuristic crossover (Leung et al. 2003).

(iii) Mutation: The operation randomly alters some individuals with a small probability, which provides a means to increase the population diversity. And the simple *uniform mutation* is used here, that is, a randomly selected chromosome $E_p(t)$ $(1 \le p \le M)$ is replaced by $E_p(t+1)$, in which each element $e_{p,j}$ $(1 \le j \le s)$ of the vector is a random number in the corresponding nearest-neighbor interval $\left[e_j^-, e_j^+\right]$.

(iv) Elitist strategy: A common selection operator is the fitness-proportional selection, which does not guarantee the selection of any particular individual, including the fittest (Bai et al. 2009). To overcome this drawback, the elitist strategy proposed by Bai et al. (2009) is employed here, which requires that the best two individuals will be selected and a copy of them will not be disrupted by crossover or mutation. And this elitist strategy can effectively avoid the loss of the best solutions.

3.2.4 Termination condition

In general, the termination condition of GA is often specified as a maximal number of generations, or as a given value of the fitness function that is deemed to be sufficient. In our implementation, we employ the former criteria.

3.2.5 Algorithm procedure of hybrid IGA–FCM algorithm

For a set of s-dimensional incomplete data $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ with h missing attributes, the procedure of the hybrid IGA–FCM algorithm for incomplete data clustering can be described as follows:

Step 1 For each incomplete instance $\tilde{x}_b$ $(1 \le b \le n)$ whose attribute $\tilde{x}_{jb}$ $(1 \le j \le s)$ is missing, find its q nearest neighbors with non-missing attribute j according to the partial distance (5), and determine the interval representation $\left[x_{jb}^-, x_{jb}^+\right]$ of $\tilde{x}_{jb}$. Renumber the nearest-neighbor intervals by their appearance order and get $\left[e_g^-, e_g^+\right]$ $(1 \le g \le h)$.

Step 2 Choose m, c and ε for clustering, where $\varepsilon > 0$ is a small positive constant; Set the genetic population size M, maximal number of generations G, and crossover probability P_c, mutation probability P_m. Initialize the genetic population by (8).

Step 3 When the genetic generation index is t $(t = 1, 2, \ldots, G)$, recover the incomplete data set $\tilde{X}$ using each chromosome $E_p(t)$ $(1 \le p \le M)$ and get complete data set X, perform FCM on X.

Step 4 Calculate the fitness value of each chromosome using (9), and ascending sort the individuals by their fitness values. Save the best two individuals.

Step 5 Perform *roulette wheel selection* according to the selection probability defined as (10). Select $M - 2$ individuals and preserve the best two ones.

Step 6 Except for the best two individuals, perform crossover based on competition and optimal selection according to the crossover probability p_c, and generate four offspring by (11)–(14), and then choose the two offspring with higher fitness values to substitute the parent chromosomes.

Step 7 Except for the best two individuals, perform *uniform mutation* according to the mutation probability p_m.

Step 8 If genetic generation index $t = G$, then stop and get the optimized imputations of missing attributes (the best individual) and the corresponding clustering results; otherwise set $t = t + 1$ and return to Step 3.

To give a brief example of the process we went through, let us consider a simple two-dimensional data set shown in Fig. 2a. In this example, the incomplete data set contains 18 complete data, which are depicted as points in the figure, and the incomplete data $\tilde{x}_g = [\tilde{x}_{g1}, ?]^{\mathrm{T}}$ and $\tilde{x}_f = [?, \tilde{x}_{f2}]^{\mathrm{T}}$ are represented as a vertical dashed line with horizontal component $\tilde{x}_{g1}$ and a horizontal dashed line with vertical component $\tilde{x}_{f2}$ respectively. Since the incomplete data set only contains 20 data, the number of nearest neighbors is set to $q = 3$. Thus, according to the partial distance (5), the nearest-neighbor intervals of missing values $\tilde{x}_{g2}$ and $\tilde{x}_{f1}$ can be obtained and renumbered as $\left[e_1^-, e_1^+\right] = [0.2424, 0.3182]$ and $\left[e_2^-, e_2^+\right] = [0.5556, 0.7778]$ respectively, which contains the imputations of the missing values in appropriate ranges rather than the whole lines. Then, let the numbers of clusters $c = 2$, the genetic population size $M = 30$, iteration number $G = 50$, crossover factor $w = 0.3$, crossover probability $P_c = 0.6$, mutation probability $P_m = 0.1$, and initialize the genetic population by (8). Thus, aiming at minimizing the clustering objective function, the optimized imputations of $\tilde{x}_{g2}$, $\tilde{x}_{f1}$ in the constraint of nearest-neighbor intervals (as shown in Fig. 2b) and the corresponding clustering results can be obtained through the genetic evolution presented in this section.

In Fig. 2b, the two solid lines represents the nearest-neighbor intervals of the two missing values, and the imputations obtained by the imputation-based IGA–FCM ($\tilde{x}_{g2} = 0.2446$, $\tilde{x}_{f1} = 0.7633$) and OCS–FCM ($\tilde{x}_{g2} = 0.1703$, $\tilde{x}_{f1} = 0.8648$), NPS–FCM ($\tilde{x}_{g2} = 0.1702$, $\tilde{x}_{f1} = 0.8649$) are represented by ○, △ and □ respectively. And it is quite noticeable that the imputations gotten by OCS–FCM and NPS–FCM algorithms are out of the range of nearest-neighbor intervals, and the imputations obtained by the proposed IGA–FCM algorithms are more rational from the nearest-neighbor perspective, and this is naturally helpful to improve the clustering performance.

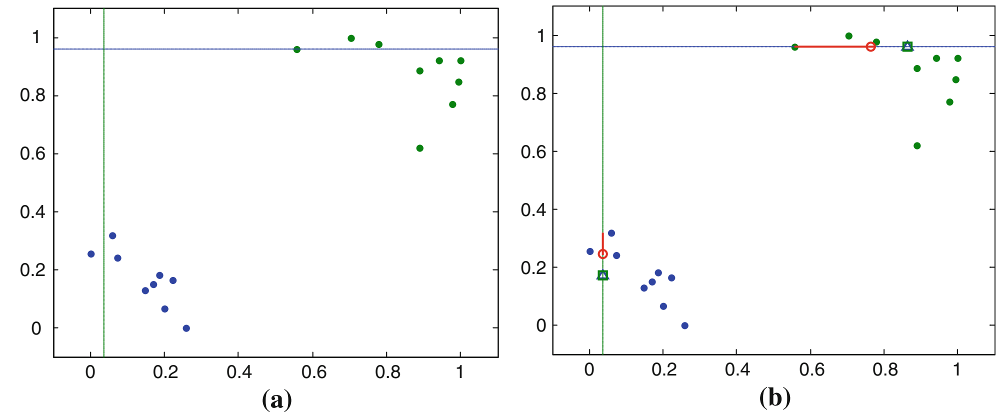

Fig. 2 **a** The incomplete two-dimensional data set, **b** the nearest-neighbor intervals and imputations obtained by imputation-based algorithms

4 Numerical experiments

4.1 Data sets

In the experiments presented below, we tested the performance of hybrid IGA–FCM algorithm on three well-known data sets: IRIS, Wine and New-Thyroid, which are taken from the UCI machine repository (Blake and Merz 1998), and often used as standard databases to test the performance of clustering algorithms.

The IRIS data contains 150 four-dimensional attribute vectors, depicting four attributes of iris flowers, which include petal length, petal width, sepal length and sepal width. The three IRIS classes involved are Setosa, Versicolor and Virginica, each containing 50 vectors. Setosa is well separated from the others, while Versicolor and Virginica are not easily separable due to the overlapping of their vectors. Hathaway and Bezdek (1995) presented the actual cluster prototypes of the IRIS data:

$$V^* = \begin{bmatrix} 5.00 & 5.93 & 6.58 \\ 3.42 & 2.77 & 2.97 \\ 1.46 & 4.26 & 5.55 \\ 0.24 & 1.32 & 2.02 \end{bmatrix} \tag{15}$$

The Wine data set is the results of a chemical analysis of wines grown in the same region but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. It contains 178 data points.

The New-Thyroid data set comprises 215 patients from the same hospital, and for each of the samples, there are five attributes. The individuals are divided into three groups based on diagnosis results where there are 150 healthy individuals, 35 patients suffering from hyperthyroidism, and 30 from hypothyroidism.

In this paper, the scheme for artificially generating an incomplete data set $\tilde{X}$ is to randomly select a specified percentage of components and designate them as missing, thus the data missingness can be considered as MCAR. The random selection of missing attribute values is constrained so that (Hathaway and Bezdek 2001)

1. each original attribute vector $\tilde{x}_k$ retains at least one component;
2. each attribute has at least one value present in the incomplete data set $\tilde{X}$.

4.2 Experimental results

To test the clustering performance, the clustering results of IGA–FCM and those of whole data strategy (WDS), partial distance strategy (PDS), optimal completion strategy (OCS), and nearest prototype strategy (NPS) versions of FCM (Hathaway and Bezdek 2001) are compared. As in our previous research (Li et al. 2010a), for the last four versions of FCM as well as standard FCM adopted in the hybrid IGA–FCM algorithm, the initialization of these algorithms is partition matrix $U^{(0)}$ that satisfies (2), and the corresponding stopping criterion is $\left\| U^{(l)} - U^{(l-1)} \right\| < \varepsilon$. In addition, the missing attributes are randomly initialized in OCS–FCM and NPS–FCM.

For the three data sets, choose fuzzification parameter $m = 2$, the numbers of clusters $c = 3$, convergence threshold $\varepsilon = 10^{-5}$, and number of nearest neighbors $q = 5$. And set the genetic population size $M = 30$, iteration number $G = 50$, crossover probability $P_c = 0.6$, mutation probability $P_m = 0.1$, crossover factor $w = 0.3$ to emphasize the effect of the boundary of nearest-neighbor intervals.

To eliminate the variation in the results from trial to trial, Tables 1, 2, and 3 present the mean number of misclassifications obtained over ten trials on incomplete IRIS, Wine and New-Thyroid data sets, and the same incomplete data set is used in each trial for each of the five approaches, so that the results can be correctly compared. In Tables 1, 2, and 3, the optimal solutions in each row are highlighted in bold, and the suboptimal solutions are underlined.

The imputation error can be calculated by

$$\left\|X_M^* - \tilde{X}_M\right\|_F^2 = \sum_{k=1}^{n} \sum_{j=1}^{s} \left|x_{jk}^* - \tilde{x}_{jk}\right|^2 \tag{16}$$

where $\tilde{X}_M = \{\tilde{x}_{jk} | \text{the value of } \tilde{x}_{jk} \text{ is missing}\}$, $\tilde{x}_{jk} \in \tilde{X}_M$ is the imputation gotten by imputation-based algorithms, and x_{jk}^* is the actual attribute value of $\tilde{x}_{jk}$ in complete data sets, $X_M^* = \left\{x_{jk}^*\right\}$. In addition, as the WDS–FCM and PDS–FCM algorithms cannot provide imputations of missing values, their imputation errors are unavailable.

However the actual cluster prototypes of the IRIS data are already known, Fig. 2 shows the mean prototype error calculated by (Hathaway and Bezdek 2001)

$$\|V - V^*\|_F^2 = \sum_{j=1}^{s} \sum_{i=1}^{c} (v_{ji} - v_{ji}^*)^2, \tag{17}$$

where V^* is the actual cluster prototypes of the IRIS data as shown in (15).

4.3 Discussion

From Tables 1, 2, and 3 and Fig. 3, it is easy to see that for 0 % missing data, all approaches reduce to regular FCM. And for other cases, different methods for handling missing attributes in FCM lead to different clustering results. In terms of misclassification error, a commonly used clustering criterion, the proposed hybrid IGA–FCM approach can always perform better than the compared methods on the three data sets. And as for the imputation error, IGA–FCM can always get the smallest values expect for the 15 % case of incomplete IRIS data sets and 20 % case of incomplete Thyroid data sets, where IGA–FCM gives suboptimal solutions. Besides, for incomplete IRIS data sets, the cluster prototypes obtained by IGA–FCM are closer to the actual ones, based on the curves in Fig. 3. The above experimental results imply that the nearest-neighbor interval representation captures the essence of pattern similarities in the original data sets, and hybrid IGA–FCM has the ability to estimate more accurate imputations of missing attributes in the nearest-neighbor intervals, which is naturally helpful to get more satisfying clustering results. As for the efficiency of the algorithms, the proposed IGA–FCM algorithm is slower than the compared algorithms, and because cluster analysis is an off-line data analysis approach, the convergence rate is not as important as the evaluation indexes mentioned above.

Furthermore, as one can observe, the WDS, PDS, OCS, and NPS versions of FCM can perform well in some cases on the three data sets, whereas in the other cases the methods cannot generate satisfying results. As an example, consider the PDS–FCM approach, in the cases that the IRIS data misses 5 and 20 %, the Wine data misses 15 % and the New-Thyroid data misses 15 and 20 % of their attributes, PDS–FCM can obtain the second smallest misclassification errors, while in the other cases PDS–FCM fails to generate satisfying results. And it is similar for the other three approaches, that is, WDS–FCM, OCS–FCM and NPS–FCM. It illustrates that these compared methods may be applicable to some certain missing cases of the data sets, and fail to exhibit robustness as the proposed IGA–FCM algorithm.

In the compared approaches, WDS–FCM simply deletes all the incomplete data, whereas PDS–FCM ignores missing attributes belonging to the discarding/ignoring methods. These two approaches (eliminating and ignoring) do not make full use of data set information and cause loss of information, which could degrade the clustering performance. The other two compared methods, NPS–FCM and OCS–FCM, are imputation methods as IGA–FCM. The former one, NPS–FCM, replaces each missing attribute by the corresponding attribute of the nearest prototype in each iteration. In addition, OCS–FCM optimizes the missing attributes by gradient optimization in the entire attribute space, in terms of the ranges of missing attributes; this approach is equivalent given that no nearest-neighbor information is taken into account but all the interval representations of missing attributes in IGA–FCM are [0,1]. Both of the two approaches do not take the attribute distribution information of the data sets into account, which affects the missing attribute estimation and the subsequent

Table 1 Averaged results of ten trials using incomplete IRIS data set

% Missing	Mean number of iterations to termination					Mean number of misclassification					Mean imputation error		
	WDS	PDS	OCS	NPS	IGA	WDS	PDS	OCS	NPS	IGA	OCS	NPS	IGA
0	24.0	26.4	23.7	24.6	50.0	16.0	16.0	16.0	16.0	16.0	0	0	0
5	24.9	24.6	30.3	25.6	50.0	**16.3**	16.9	17.3	16.9	**16.3**	0.3762	0.3720	**0.3148**
10	24.5	23.9	34.1	29.9	50.0	17.1	16.8	16.3	16.7	**16.2**	0.9111	0.9163	**0.8328**
15	24.8	22.7	37.9	30.1	50.0	16.4	16.8	16.8	16.4	**15.8**	2.0695	**1.3369**	1.4555
20	25.4	25.8	39.8	30.8	50.0	16.4	16.2	16.6	16.2	**16.1**	2.2760	3.8582	**2.0533**

Table 2 Averaged results of ten trials using incomplete Wine data set

% Missing	Mean number of iterations to termination					Mean number of misclassification					Mean imputation error		
	WDS	PDS	OCS	NPS	IGA	WDS	PDS	OCS	NPS	IGA	OCS	NPS	IGA
0	28.0	20.0	22.0	22.0	50.0	9.0	9.0	9.0	9.0	9.0	0	0	0
5	21.6	23.1	27.6	26.5	50.0	10.3	10.0	10.0	9.9	**9.6**	2.8380	2.8538	**2.4241**
10	25.5	23.2	39.3	28.1	50.0	12.7	10.2	10.7	10.1	**9.2**	5.8875	5.8695	**5.0264**
15	37.1	22.4	37.7	29.7	50.0	21.8	12.4	13.2	12.5	**10.9**	8.7404	8.6543	**7.9555**
20	42.9	22.1	46.7	32.3	50.0	45.1	12.0	12.7	11.9	**10.8**	11.0539	11.0553	**10.8624**

Table 3 Averaged results of ten trials using incomplete New-Thyroid data set

% Missing	Mean number of iterations to termination					Mean number of misclassification					Mean imputation error		
	WDS	PDS	OCS	NPS	IGA	WDS	PDS	OCS	NPS	IGA	OCS	NPS	IGA
0	97.0	79.0	81.0	175.0	50.0	20.0	20.0	20.0	20.0	20.0	0	0	0
5	88.1	105.2	198.9	159.3	50.0	35.8	40.7	38.9	32.2	**29.0**	0.6790	0.6121	**0.5758**
10	97.8	132.9	111.4	108.5	50.0	32.3	41.3	43.2	44.9	**27.3**	1.7632	1.6584	**1.4727**
15	83.8	106.0	153.9	102.3	50.0	57.1	29.5	45.9	54.4	**25.2**	2.9275	2.7455	**2.6579**
20	117.7	104.7	127.4	94.0	50.0	64.9	37.1	49.2	49.0	**34.3**	4.5956	**3.6984**	4.0785

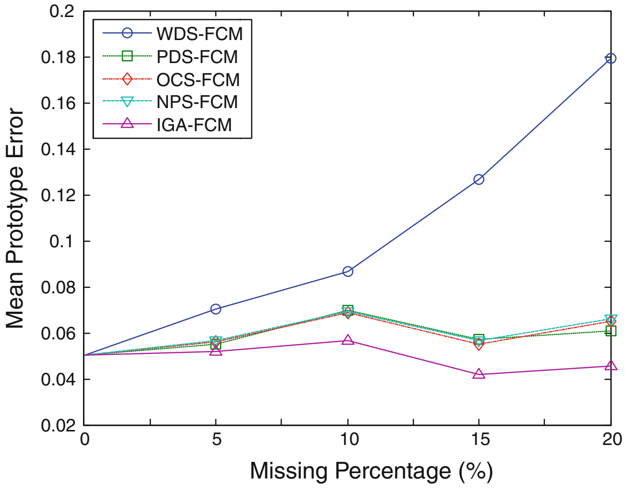

Fig. 3 Comparison of averaged prototype error of 10 trials using incomplete IRIS data by five algorithms

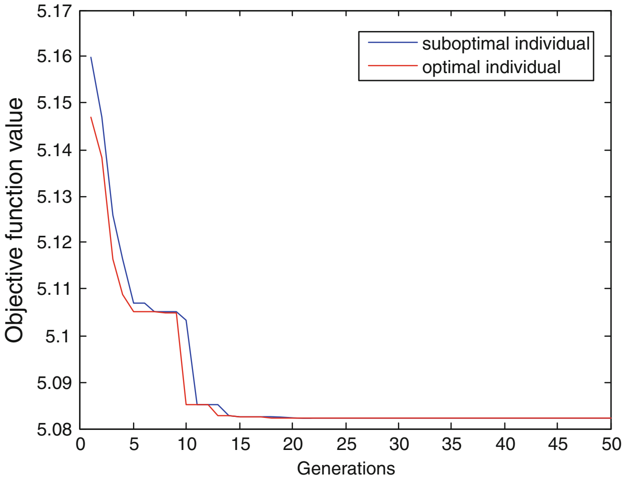

Fig. 4 The genetic iteration trend lines for the optimal and suboptimal individuals

clustering analysis. In comparison, IGA–FCM adopts nearest-neighbor intervals to represent missing attributes, which can limit missing attribute estimation to appropriate ranges so as to avoid that the improper information misleads the missing attribute estimation. Moreover, compared with the gradient optimization used in OCS–FCM, GA employed in IGA–FCM has excellent optimization ability. Thus, with appropriate imputations of missing attributes that consider nearest-neighbor information, improved clustering results can be obtained by the proposed IGA–FCM algorithm.

Figure 4 shows the variations of objective function values (9) for the optimal and suboptimal individuals in 50 generations when clustering the IRIS data set with 5 % of its data missing (tested on a dual-core 2.53 GHz PC with 4 GB of RAM). We can see that, in the genetic process, the optimal and suboptimal solutions tend to be consistent gradually and can finally achieve convergence.

5 Conclusion

In this paper, we have presented a hybrid genetic algorithm–fuzzy *c*-means approach for the problem of incomplete data clustering. The proposed algorithm has two main characteristics. Firstly, based on the partial distance between incomplete data and other samples in data set, missing attributes are represented by nearest-neighbor intervals that can capture the essence of pattern similarities in data sets. Accordingly, the missing attribute estimation can be limited to the subsets that contain nearest neighbors of incomplete data rather than the entire attribute space, which can avoid the effect of improper information on missing attribute estimation effectively. Secondly, based on the interval representation of missing attributes, the proposed algorithm hybridizes GA and FCM, and optimizes missing attribute imputations in corresponding nearest-neighbor intervals, and clustering results of incomplete data set can be obtained simultaneously. Experiments on several famous UCI data sets have demonstrated the performance of the proposed hybrid algorithm; the proposed algorithm is clearly superior to the compared methods in terms of clustering performance, which shows that the hybrid IGA–FCM algorithm effectively solves the incomplete data clustering problem.

References

Acuna E, Rodriguez C (2004) The treatment of missing values and its effect in the classifier accuracy. In: Banks D, House L, McMorris F, Arabie P, Gaul W (eds) Classification, clustering and data mining applications. Springer, Berlin, pp 639–648

Bai H, Zhang P, Ajjarapu V (2009) A novel parameter identification approach via hybrid learning for aggregate load modeling. IEEE Trans Power Syst 24:1145–1154

Bandyopadhyay S (2005) Simulated annealing using a reversible jump Markov chain Monte Carlo algorithm for fuzzy clustering. IEEE Trans Knowl Data Eng 17:479–490

Bandyopadhyay S, Saha S (2008) A point symmetry-based clustering technique for automatic evolution of clusters. IEEE Trans Knowl Data Eng 20:1441–1457

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York

Blake CL, Merz CJ (1998) UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA

Blickle T, Thiele L (1996) A comparison of selection schemes used in evolutionary algorithms. Evol Comput 4:361–394

Chang PC, Liao TW (2006) Combing SOM and fuzzy rule base for flow time prediction in semiconductor manufacturing factory. Appl Soft Comput 6:198–206

Chang PC, Liu CH, Fan CY (2009) Data clustering and fuzzy neural network for sales forecasting: a case study in printed circuit board industry. Knowl Based Syst 22:344–355

Chang PC, Fan CY, Dzan WY (2010) A CBR-based fuzzy decision tree approach for database classification. Expert Syst Appl 37:214–225

Davis L (1991) Handbook of genetic algorithms. Van Nostrand Reinhold, New York

Deb K (2001) Multiobjective optimization using evolutionary algorithms. Wiley, Chichester

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Dixon JK (1979) Pattern recognition with partly missing data. IEEE Trans Syst Man Cybern 9:617–621

Farhangfar A, Kurgan LA, Pedrycz W (2007) A novel framework for imputation of missing values in databases. IEEE Trans Syst Man Cybern A 37:692–709

Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Menlo Park

Hathaway RJ, Bezdek JC (1995) Optimization of clustering criteria by reformulation. IEEE Trans Fuzzy Syst 3:241–245

Hathaway RJ, Bezdek JC (2001) Fuzzy *c*-means clustering of incomplete data. IEEE Trans Syst Man Cybern Part B 31:735–744

Hathaway RJ, Bezdek JC (2002) Clustering incomplete relational data using the non-Euclidean relational fuzzy *c*-means algorithm. Pattern Recognit Lett 23:151–160

Honda K, Ichihashi H (2004) Linear fuzzy clustering techniques with missing values and their application to local principle component analysis. IEEE Trans Fuzzy Syst 12:183–193

Hoppner F, Klawonn F, Kruse R, Runkler T (1999) Fuzzy cluster analysis: methods for classification data analysis and image recognition. Wiley, New York

Huang X, Zhu Q (2002) A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets. Pattern Recognit Lett 23:1613–1622

Leung FHF, Lam HK, Ling SH, Tam PKS (2003) Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE Trans Neural Netw 14:79–88

Li D, Gu H, Zhang LY (2010a) A fuzzy *c*-means clustering algorithm based on nearest-neighbor intervals for incomplete data. Expert Syst Appl 37:6942–6947

Li D, Zhong CQ, Zhang LY (2010b) Fuzzy *c*-means Clustering of partially missing data sets based on statistical representation. In: Proceedings of the 7th international conference on fuzzy systems and knowledge discovery, pp 460–464

Lim CP, Leong JH, Kuan MM (2005) A hybrid neural network system for pattern classification tasks with missing features. IEEE Trans Pattern Anal Mach Intell 27:648–653

Liu YG, Chen KF, Liao XF, Zhang W (2004) A genetic clustering method for intrusion detection. Pattern Recognit 37:927–942

Mclachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York

Michalewicz Z (1994) Genetic algorithms + data structure = evolution programs. Springer, New York

Miyamoto S, Takata O, Umayahara K (1998) Handling missing values in fuzzy *c*-means. In: Proceedings of the third Asian fuzzy systems symposium, pp 139–142

Mukhopadhyay A, Maulik U, Bandyopadhyay S (2009) Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes. IEEE Trans Evol Comput 13:991–1005

Ren ZW, San Y (2007) Improvement of real-valued genetic algorithm and performance study. Acta Electronica Sinica 35:269–274 (in Chinese)

Silva EL, Gil HA, Areiza JM (2000) Transmission network expansion planning under an improved genetic algorithm. IEEE Trans Power Syst 15:1168–1175

Stade I (1996) Hot deck imputation procedures. In: Incomplete data in sample survey symposium on incomplete data proceedings, pp 225–248

Su JP, Lee TE, Yu KW (2009) A combined hard and soft variable-structure control scheme for a class of nonlinear systems. IEEE Trans Ind Electron 56:3305–3313

Timm H, Doring C, Kruse R (2004) Different approaches to fuzzy clustering of incomplete data sets. Int J Approx Reason 35:239–249

Wei CH, Fahn CS (2002) The multisynapse neural network and its application to fuzzy clustering. IEEE Trans Neural Netw 13: 600–618

Zhu JJ, Liu SX, Wang MG (2004) Estimation of weight vector of interval numbers judgment matrix in AHP using genetic algorithm. J Syst Eng 19:343–349 (in Chinese)