Pattern Anal Applic (2015) 18:377–384
DOI 10.1007/s10044-014-0376-8

SHORT PAPER

A hybrid clustering algorithm based on missing attribute interval estimation for incomplete data

Li Zhang · Zhaohong Bing · Liyong Zhang

Received: 16 January 2013 / Accepted: 15 May 2014 / Published online: 1 June 2014

Abstract Partially missing data sets are a prevailing problem in clustering analysis. We propose a hybrid algorithm combining fuzzy clustering with particle swarm optimization (PSO) for incomplete data clustering, and missing attributes are represented as intervals. Furthermore, we develop a neighbor interval reconstruction (NIR) method based on pre-classification results that estimates the nearest-neighbor interval of missing attribute using the nearest-neighbor rule, which avoids endpoints of intervals determined by different species information, thereby improving the accuracy of missing attribute intervals and enhancing the robustness of missing attribute imputation. Then, the PSO and fuzzy c-means hybrid algorithm are used for clustering the interval-valued data set, and the global optimization ability of the PSO can improve the accuracy of clustering results compared with gradient-based optimization methods. The experimental results for several UCI data sets show the superiority of the proposed NIR hybrid algorithm.

Keywords Incomplete data set · Intervals reconstruction · Particle swarm · Fuzzy c-means · Clustering

L. Zhang · Z. Bing (✉)
School of Information, Liaoning University, Shenyang 110036, China
e-mail: bzhhappy@126.com

L. Zhang
School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China

1 Introduction

Cluster analysis is an important technique in data mining and applied to diverse areas [1]. Fuzzy c-means (FCM) is one of the most widely applied clustering methods [2–4], with which not only the final clustering results can be obtained, but also which class the extent of the data belongs to can be decided according to the membership. Therefore, FCM is a valid method for complete data. However, in practical applications, many data sets suffer from incompleteness, and FCM can not be directly applied to handle such incomplete data sets.

In the last decade, a number of new strategies based on existing clustering methods have been proposed for solving the problem of incomplete data set clustering. The expectation–maximization (EM) has been used to handle incomplete data and probabilistic clustering for a longtime. Abas [5] combined EM algorithm with finite mixture model to deal with incomplete data. Furthermore, he [6] proposed EM and particle swarm hybrid algorithm to estimate the incomplete data values. Then Lin and Su [7] combined Bayesian classifier with the EM algorithm to process incomplete data in feature extraction.

In addition, Hathaway and Bezdek [8] proposed several specific methods for incomplete data clustering based on FCM: Whole data strategy (WDS) is one of the simplest methods, the data with missing attributes are discarded, and FCM is applied to the remaining complete data. Another method is partial distance strategy (PDS), which ignores missing attributes; the distance in the FCM algorithm is calculated by the local distance proposed by Dixon [9]. The estimation of the missing attributes is regarded as the optimization problem in optimal completion strategy (OCS), and better estimated values are obtained during the process of clustering iterations. The missing attributes are

set as the corresponding attribute values from the nearest cluster prototypes during each iteration in nearest prototype strategy (NPS). Di Nuovo [10] applied a new technique in a psychological research environment using fuzzy clustering for incomplete data sets. Aydilek and Arslan [11] proposed a hybrid approach which used support vector regression and genetic algorithm to estimate missing values and optimize the parameters of FCM.

Furthermore, Simiński [12] used the marginalization and imputation to create the rough fuzzy clusters for missing data. The marginalization removes the tuples with missing values. The imputation is used to handle data with missing values. Nowicki [13] presented a new approach to fuzzy classification in the case of missing data. Rough-fuzzy sets are incorporated into logical type neuro-fuzzy structures, and a rough-neuro-fuzzy classifier is derived.

And the incomplete data can also be considered as variables to be optimized in optimization model. Dopazo and Ruiz-Tagle [14] used parameters scalarization and logarithmic target program design idea to establish the optimization model for missing data. Pei [15] established the fuzzy multi-attribute decision model to deal with incomplete data.

In order to make use of clustering distribution and data sets information to handle incomplete data, Himmelspach and Conrad [16] took the cluster dispersion into account and proposed a new membership degree for missing value estimation based on cluster dispersion. Zhang et.al [17] utilized information within incomplete instances (instances with missing values), when estimating missing values. Considering that different data types need different solutions, Subasi et.al [18] used the Boolean similarity measure to determine the missing binary data values. However, Hathaway and Bezdek [19] proposed the FCM clustering method according to the mathematical triangle inequality rules for incomplete relational data set clustering.

Recently, the nearest-neighbor [20, 21] method is increasingly used to handle incomplete data. Doquire and Verleysen [22] used the nearest-neighbor method based on mutual information to estimate the incomplete data values. Van Hulse and Khoshgoftaar [23] used complete data and incomplete data that have been estimated to search K nearest-neighbors of incomplete data, missing attributes are replaced by the mean values of these neighbors' corresponding attributes.

The missing attributes are represented by the numerical values in most of the above approaches. Because of the uncertainty of missing attributes, replacing missing attribute values by intervals can improve the robustness of the missing attributes estimation. The nearest-neighbors of missing attributes were used to determine missing attribute intervals (NNI) in the literature [24] which dealt with clustering problem for incomplete data by transforming the data set to an interval-valued one. Then the interval-valued FCM based on gradient completed the incomplete data clustering. But the restriction that missing attributes and nearest-neighbors are in the same cluster is not taken into account during determining intervals.

Based on neighbor interval reconstruction (NIR), a PSO and FCM hybrid algorithm are proposed for incomplete data in this paper. Firstly, PDS–FCM is used in the pre-classification of the incomplete data set, and then, the nearest-neighbors of incomplete data are found using the attribute distribution information. According to the results of pre-classification, the nearest-neighbors that are in different cluster with incomplete data are removed, the rest of congeneric neighbors are used to determine the missing attribute intervals, which can avoid endpoints of intervals determined by different species information. Therefore, the accuracy of missing attributes imputation will be enhanced. The incomplete data set is transformed into the interval-valued data set. Secondly, the proposed NIR hybrid algorithm is used for interval-valued data set clustering. Particles are encoded by the cluster prototypes in the hybrid algorithm; and memberships are still obtained by the gradient-based alternating iterative formula. The NIR hybrid algorithm is beneficial to improve the clustering performance.

The remaining parts of this paper are organized as follows. Section 2 presents the reconstruction of nearest-neighbor intervals for missing attributes. The PSO and FCM hybrid algorithm for incomplete data clustering are introduced in Sect. 3. Section 4 presents clustering results of several UCI data sets and a comparative study of our proposed algorithm with various other methods for handling missing values in FCM. Finally, conclusions are drawn in Sect. 5.

2 The intervals estimation

We use nearest-neighbor intervals to represent missing attributes. The partial Euclidean distance is used to calculate the distance between data in incomplete data set. According to the distance, the q nearest-neighbors [24] of an incomplete datum can be selected, and their corresponding attributes must be complete. Then, the minimum and maximum of the neighbors' corresponding attribute values are used to determine the value range of the incomplete datum's missing attribute.

$\bar{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ is an s-dimensional incomplete data set, which contains at least one incomplete datum with some (but not all) missing attribute values. The partial distance [9] calculation of $\tilde{x}_b$ and an instance $\tilde{x}_p$ (incomplete or complete) is calculated by the formula (1):

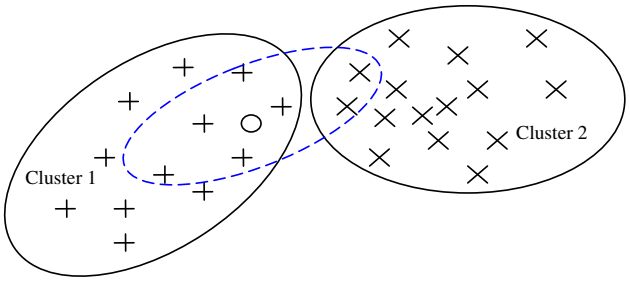

Fig. 1 The interval estimation with different species

$$D_{pb} = \frac{1}{\sum_{j=1}^{s} I_j} \sum_{j=1}^{s} \left(\tilde{x}_{jb} - \tilde{x}_{jp}\right)^2 I_j \tag{1}$$

where $\tilde{x}_{jb}$ and $\tilde{x}_{jp}$ are the jth attribute of $\tilde{x}_b$ and $\tilde{x}_p$, respectively, and the value of I_j is 0 or 1. If both $\tilde{x}_{jb}$ and $\tilde{x}_{jp}$ are non-missing, I_j is 1; otherwise, I_j is 0.

The partial distance calculation makes use of the information of complete data and known attributes in incomplete data. The missing attribute $\tilde{x}_{jb}$ can be represented by its corresponding nearest-neighbor interval $[x_{jb}^-, x_{jb}^+]$, and the non-missing attribute $\tilde{x}_{jw}$ can be rewritten into interval form $[x_{jw}^-, x_{jw}^+]$, but $x_{jw}^- = x_{jw}^+ = \tilde{x}_{jw}$.

Considering that not all neighbors are in the same cluster with the incomplete data and whetherthe intervals' endpoints are determined by corresponding attributes of those different species neighbors, it will cause an obviously unreasonable interval as shown in Fig. 1. Where, the circle sample in cluster 1 is an incomplete datum with missing attributes, and the dashed frame rings out its nearest-neighbors. It can be seen that nearest-neighbors contain samples of cluster 2.

In order to solve this problem, we adopt PDS to pre-classify incomplete data set. Whether q nearest-neighbors contain different species data or not can be judged from the pre-classification results. If q nearest-neighbors contain different species data, then different species data are removed. And the rest of congeneric neighbors are used to determine the endpoints of missing attributes intervals.

3 Particle swarm and fuzzy c-means hybrid optimization

3.1 The interval fuzzy c-means

Let $\overline{X} = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n\}$ be an interval-valued data set to be partitioned into c clusters, the attribute in the data set is $\overline{x}_k = [\overline{x}_{1k}, \overline{x}_{2k}, \ldots, \overline{x}_{sk}]^T$, $\forall j, k : \overline{x}_{jk} = \left[x_{jk}^-, x_{jk}^+\right]$. The objective function of the interval-valued data FCM is:

$$J(\boldsymbol{U}, \overline{\boldsymbol{V}}) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{v}}_i\|_2^2 \tag{2}$$

with the constraint of

$$\sum_{i=1}^{c} u_{ik} = 1, \quad k = 1, 2, \ldots, n \tag{3}$$

The interval cluster prototypes matrix is $\overline{\boldsymbol{V}} = \left[\overline{v}_{ij}\right] = [\overline{\boldsymbol{v}}_1, \overline{\boldsymbol{v}}_2, \ldots \overline{\boldsymbol{v}}_c]$, where $\overline{v}_{ij} = \left[v_{ji}^-, v_{ji}^+\right]$, $\forall i = 1, 2, \ldots c, j = 1, 2, \ldots s$. The Euclidean distance between $\overline{\boldsymbol{x}}_k$ and $\overline{\boldsymbol{v}}_i$ is defined as:

$$\begin{aligned} \|\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{v}}_i\|_2 = \Big[&\left(\boldsymbol{x}_k^- - \boldsymbol{v}_i^-\right)^T \left(\boldsymbol{x}_k^- - \boldsymbol{v}_i^-\right) \\ &+ \left(\boldsymbol{x}_k^+ - \boldsymbol{v}_i^+\right)^T \left(\boldsymbol{x}_k^+ - \boldsymbol{v}_i^+\right) \Big]^{\frac{1}{2}} \end{aligned} \tag{4}$$

where $\boldsymbol{x}_k^- = \left[x_{1k}^-, x_{2k}^-, \ldots, x_{sk}^-\right]^T$, $\boldsymbol{x}_k^+ = \left[x_{1k}^+, x_{2k}^+, \ldots x_{sk}^+\right]^T$ and $\boldsymbol{v}_i^- = \left[v_{1i}^-, v_{2i}^-, \ldots, v_{si}^-\right]^T$, $\boldsymbol{v}_i^+ = \left[v_{1i}^+, v_{2i}^+, \ldots, v_{si}^+\right]^T$.

The updating formulas of cluster prototypes are as follows:

$$\boldsymbol{v}_i^- = \frac{\sum_{k=1}^{n} u_{ik}^m \boldsymbol{x}_k^-}{\sum_{k=1}^{n} u_{ik}^m}, \quad i = 1, 2, \ldots, c \tag{5}$$

$$\boldsymbol{v}_i^+ = \frac{\sum_{k=1}^{n} u_{ik}^m \boldsymbol{x}_k^+}{\sum_{k=1}^{n} u_{ik}^m}, \quad i = 1, 2, \ldots, c \tag{6}$$

And if $\exists k, h, 1 \le k \le n, 1 \le h \le c, \forall j : \overline{x}_{jk} \subseteq \overline{v}_{jh}$, that is, $\overline{x}_k$ is within the convex hyper-polyhedron formed by $\overline{v}_h$, then $\overline{x}_k$ can be considered to belong fully to the cluster with membership 1, and belong to the other clusters with membership 0, thus [24]

$$u_{ik} = \begin{cases} 1, & i = h \\ 0, & i \ne h \end{cases}, \quad i = 1, 2, \ldots, c \tag{7}$$

In other cases, the membership is calculated by formula (8):

$$u_{ik} = \left[\sum_{t=1}^{c} \left(\frac{\|\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{v}}_i\|_2^2}{\|\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{v}}_t\|_2^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad i = 1, 2, \ldots, c \tag{8}$$

FCM can get the final clustering results; additionally, it calculates the degree of each data belonging to each cluster according to the membership, which obtains more clustering information. However, memberships and cluster prototypes must be initialized firstly in FCM, so the algorithm depends on the initial selection greatly. In addition, the memberships of FCM are calculated based on gradient descent mechanism, and the algorithm is easy to be trapped in the local optimal. To solve above problems, we introduce the swarm optimization strategy.

3.2 The hybrid optimization algorithm

PSO is a stochastic global optimization tool [25–27] which can be used to search for global optimal cluster prototypes of FCM. The PSO algorithm starts with a population of particles whose positions represent the potential solutions for the studied problem, and the velocity is randomly initialized in the search space. The performance of each particle (i.e., how close the particle is to the global optimum) is measured using a fitness function that varies depending on the optimization problem [28]. In the process of iterations, the search for optimal position is performed by updating velocities and positions of particles. The velocity of each particle is updated using two best positions, that is, the individual best position and the global best position. Each particle memorizes its own best position encountered so far which is called the individual best. On the other hand, the population memorizes the best position among all individual best positions obtained so far which is called the global best [29]. Then, a new generation of swarm is produced. The swarm conduct global search in the whole solution area during optimization process.

Particle swarm whose population size is n expressed as $X = \{x_1, x_2, \ldots, x_n\}$, x_i and v_i denote the position and velocity of the ith particle; p_b and g_b denote the individual best position and the global best position. The velocity and position updating formulas of particle are as follows:

$$v_i(t+1) = wv_i(t) + c_1\text{rand}(p_b(t) - x_i(t)) + c_2\text{rand}(g_b(t) - x_i(t)) \quad (9)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (10)$$

where w is the inertia weight, c_1, c_2 are the learning factors and rand is a randomly generated number between 0 and 1.

One shortcoming of the classic particle swarm is that it is easy to trap in the stagnation. During the stagnation time, the velocity of the particle is almost zero; particles gather near a point, which means that the algorithm is trapped in the local optimal [30]. The variation is put in the iterative process of the particle swarm to allow the particle swarm to escape from the local optima. When the particle corresponding to the global best position has not been improved for consecutive A times, the particle swarm is considered to have gathered to a local optimum location [31] and positions of the entire particle swarm are mutated as a certain variation probability. We take the mutation probability as $p(t) = \frac{1}{1+t^{0.5}}$, t is the number of iterations. The position of the particle is changed by the following formula:

$$x_i(t) = \text{rand} \times (\max - \min) + \min \quad (11)$$

max, min are the upper and lower bounds of the search space.

The NIR hybrid algorithm minimizes the objective function which is illustrated as the formula (2). Cluster prototypes are represented by particles of particle swarm; and memberships are still obtained by the gradient-based alternating iterative formula, as the formula (8). New cluster prototypes and memberships are obtained through updating velocities and positions of particles. From this, the NIR hybrid algorithm is proposed. The process is as follows:

Step (1) Determine the missing attribute intervals:

1. For each incomplete data, the q nearest-neighbors are determined according to the formula (1);
2. PDS is used in pre-classification of data set, and different species data are removed according to the results. The interval $[x_{jb}^-, x_{jb}^+]$ of x_{jb} is determined, if x_{jb} is non-missing attribute, then x_{jb}^-, x_{jb}^+.

Step (2) Initialization: Cluster prototypes are represented by particles, the number of particles in the particle swarm and the particle dimension are determined, velocities and positions of particles are initialized and the maximum number of iterations is set.

Step (3) The memberships are calculated using formula (8); the objective function values are calculated using formula (2). The objective function values are used as the fitness value of each particle.

Step (4) The global optimum options: for each particle, its fitness value is compared with the fitness value of best position which the swarm experienced, if better, it is used as the global optimum.

Step (5) Judge whether the algorithm satisfies the variation condition or not, if it satisfies the variation condition, the particles are mutated using formula (11); otherwise, go to the next step.

Step (6) The velocities and positions of particles are updated using formula (9) and formula (10).

Step (7) Judge whether the algorithm satisfies the terminated condition (the iterations get the maximum number or the error is less than a given error), if it satisfies the terminated condition, obtain the terminated classification matrix and cluster centers, otherwise go to the step 3).

4 Experimental results and analysis

Five data sets of the UCI database: Iris, Wine, Bupa, Haberman and Breast are selected to do the simulation experiments, and the experimental results of NIR are compared with five methods WDS, PDS, OCS, NPS and NNI, thereby verifying the effectiveness of NIR. The information of data sets as shown in Table 1.

The Iris data set 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.

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. The data set contains 178 data points.

The Bupa Liver Disorder data set includes 345 samples in six-dimensional space. The first five attributes are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each data point constitutes the record of a single male individual, and the data set has two clusters.

The Haberma data set contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The first three attributes are information of patients, respectively, is patient's age at time of operation, patient's year of operation and number of positive axillary nodes detected, and the last attribute is survival status (class attribute) including three classes.

Samples of the Breast data set arrive periodically as Dr. Wolberg reports his clinical cases, which contain nine attributes describing the details of each case. Besides, two additional attributes are sample code number and class attribute: benign and malignant.

Missing attributes are randomly generated by human. The relevant parameters in the experiments are set as follows: q is set to 6, A is set to 10, the size of the particle swarm is 100, the number of iterations is 500 times and the missing data rates are taken as 0, 5, 10, 15 and 20 %. The experimental results are as shown from Tables 2, 3, 4, 5 and 6, and the optimal and sub-optimal results are marked by bold and underlined types.

As can be seen from Tables 2, 3, 4, 5 and 6,

1. The test data sets in the case of different attributes missing rates, in general, the experimental results of NIR are the best, the average number of misclassification is the least. Only when the missing rate of wine was 20 %, the result is slightly worse than WDS and in sub-optimal. Moreover, results of NIR and NNI are better than PDS, OCS and NPS in general, namely, the intervals estimation of missing attributes is better than the point estimation. It is because that the interval estimation can improve the robustness of missing attributes estimation. Among them, the results of NNI on Haberman data set

Table 1 The information of data sets

The data sets	Number of samples	Number of attributes	Number of classes
Iris	150	4	3
Wine	178	13	3
Bupa	345	7	2
Haberman	306	4	2
Breast	683	11	2

Table 2 Averaged results of 30 trials using incomplete Iris data set

%Missing	The average number of misclassification						Mean number of iterations to termination					
	WDS	PDS	OCS	NPS	NNI	NIR	WDS	PDS	OCS	NPS	NNI	NIR
0	16	16	16	16	16	11	25.37	25.23	25.93	25.03	25.47	70.00
5	16.64	17.10	16.81	16.75	16.47	11.23	25.00	25.37	31.87	26.37	26.00	83.30
10	16.57	16.98	17.19	16.87	16.70	11.57	26.50	26.53	39.33	27.53	24.50	134.27
15	16.10	17.45	17.32	17.00	16.13	11.62	25.60	25.93	36.77	28.10	24.30	157.50
20	16.33	17.78	17.12	17.64	16.33	11.71	27.20	26.80	38.53	29.27	23.03	148.23

Table 3 Averaged results of 30 trials using incomplete Wine data set

%Missing	The average number of misclassification						Mean number of iterations to termination					
	WDS	PDS	OCS	NPS	NNI	NIR	WDS	PDS	OCS	NPS	NNI	NIR
0	27	27	27	27	27	**19**	24.52	25.10	25.30	**24.50**	42.13	54.67
5	29.55	38.00	37.81	38.35	35.50	**21.72**	34.10	**28.40**	41.20	34.30	45.50	63.30
10	28.47	41.30	40.11	38.49	37.41	**23.13**	38.20	**31.31**	58.43	36.43	54.43	162.00
15	29.17	43.10	42.75	45.19	42.13	**26.48**	39.40	**35.20**	59.60	37.67	85.40	155.47
20	**28.76**	48.40	45.17	46.38	42.47	29.63	45.60	44.56	63.73	**42.50**	96.50	137.63

Bold values indicate optimal results. Underlined values indicate suboptimal results

Table 4 Averaged results of 30 trials using incomplete Bupa data set

%Missing	The average number of misclassification						Mean number of iterations to termination					
	WDS	PDS	OCS	NPS	NNI	NIR	WDS	PDS	OCS	NPS	NNI	NIR
0	177	177	177	177	177	**165.34**	**33.27**	34.03	34.10	34.90	34.37	62.70
5	177.47	177.27	177.50	177.23	176.87	**166.71**	35.30	**34.17**	40.87	36.33	36.27	61.83
10	176.40	177.00	176.67	177.00	176.43	**166.89**	38.60	**33.93**	45.53	41.13	41.13	119.47
15	177.80	178.53	177.57	178.47	174.60	**167.25**	34.60	**33.07**	68.87	43.07	49.90	158.00
20	178.33	179.17	177.47	179.00	176.63	**167.87**	40.10	**33.53**	58.03	46.07	75.93	209.20

Bold values indicate optimal results. Underlined values indicate suboptimal results

Table 5 Averaged results of 30 trials using incomplete Haberman data set

%Missing	The average number of misclassification						Mean number of iterations to termination					
	WDS	PDS	OCS	NPS	NNI	NIR	WDS	PDS	OCS	NPS	NNI	NIR
0	60.58	72.38	68.62	69.43	67.77	**52.54**	33.27	**28.00**	33.73	28.63	31.10	60.00
5	65.98	74.51	71.75	73.72	71.21	**55.51**	**32.30**	32.93	37.20	33.00	34.73	77.10
10	68.16	83.09	78.88	79.13	76.80	**58.90**	39.87	**35.67**	44.00	42.67	45.70	108.53
15	71.34	84.65	82.52	79.85	78.76	**60.32**	41.43	**38.00**	49.67	47.00	56.00	130.00
20	72.96	86.26	85.10	82.17	79.20	**63.26**	**43.10**	45.00	57.93	53.47	67.97	164.87

Bold values indicate optimal results. Underlined values indicate suboptimal results

Table 6 Averaged results of 30 trials using incomplete Breast data set

%Missing	The average number of misclassification						Mean number of iterations to termination					
	WDS	PDS	OCS	NPS	NNI	NIR	WDS	PDS	OCS	NPS	NNI	NIR
0	30	30	30	30	30	**24**	25.07	24.90	**24.70**	25.62	32.00	72.00
5	30.33	30.98	31.80	31.90	30.67	**25.89**	26.77	**24.96**	25.90	27.70	32.62	83.17
10	30.25	31.00	31.76	31.87	31.00	**26.35**	27.85	**25.17**	28.57	28.45	33.28	103.73
15	32.20	33.67	33.35	32.95	31.25	**28.78**	30.70	**25.46**	30.35	29.95	32.27	122.00
20	33.25	34.18	36.20	34.71	33.07	**28.89**	34.25	**25.72**	31.75	32.43	33.94	143.87

Bold values indicate optimal results. Underlined values indicate suboptimal results

are slightly worse owing to the following restriction. Namely, it failed to take into account that the nearest-neighbors are in the same cluster with incomplete data during the determining intervals in NNI.

2. Results of NIR are better than NNI, the missing rate is higher and the accuracy of clustering results is better with NIR. This is because the estimated intervals are restricted in NIR and different species neighbors are removed. In addition, the NIR hybrid algorithm is used for interval-valued data set clustering. The hybrid algorithm utilizes the global optimization capability of particle swarm to search the optimal clustering results, and memberships are still obtained by the gradient-based alternating iterative formula. Compared with the FCM, the hybrid algorithm can avoid the algorithm to be trapped in local convergence and deal with the problem of sensitive to the initial values, thereby improving the accuracy of the results.
3. Compared with other methods, the mean number of iterations to termination of NIR is more. However, the NIR hybrid algorithm makes use of the global search ability of intelligent optimization and obtains better clustering results; thus, the algorithm efficiency must be slightly inferior to the gradient-based optimization search method. The curves between the objective function and iterations under the missing rate are 0, 5, 10, 15 and 20 % with NPF on Iris, Wine, Bupa, Harberman and Breast data sets are shown in Fig. 2.

5 Conclusions

The NIR method removes different species information based on the pre-classification results of incomplete data set, which makes the intervals estimation of missing attribute more reasonable. Then the NIR hybrid algorithm is

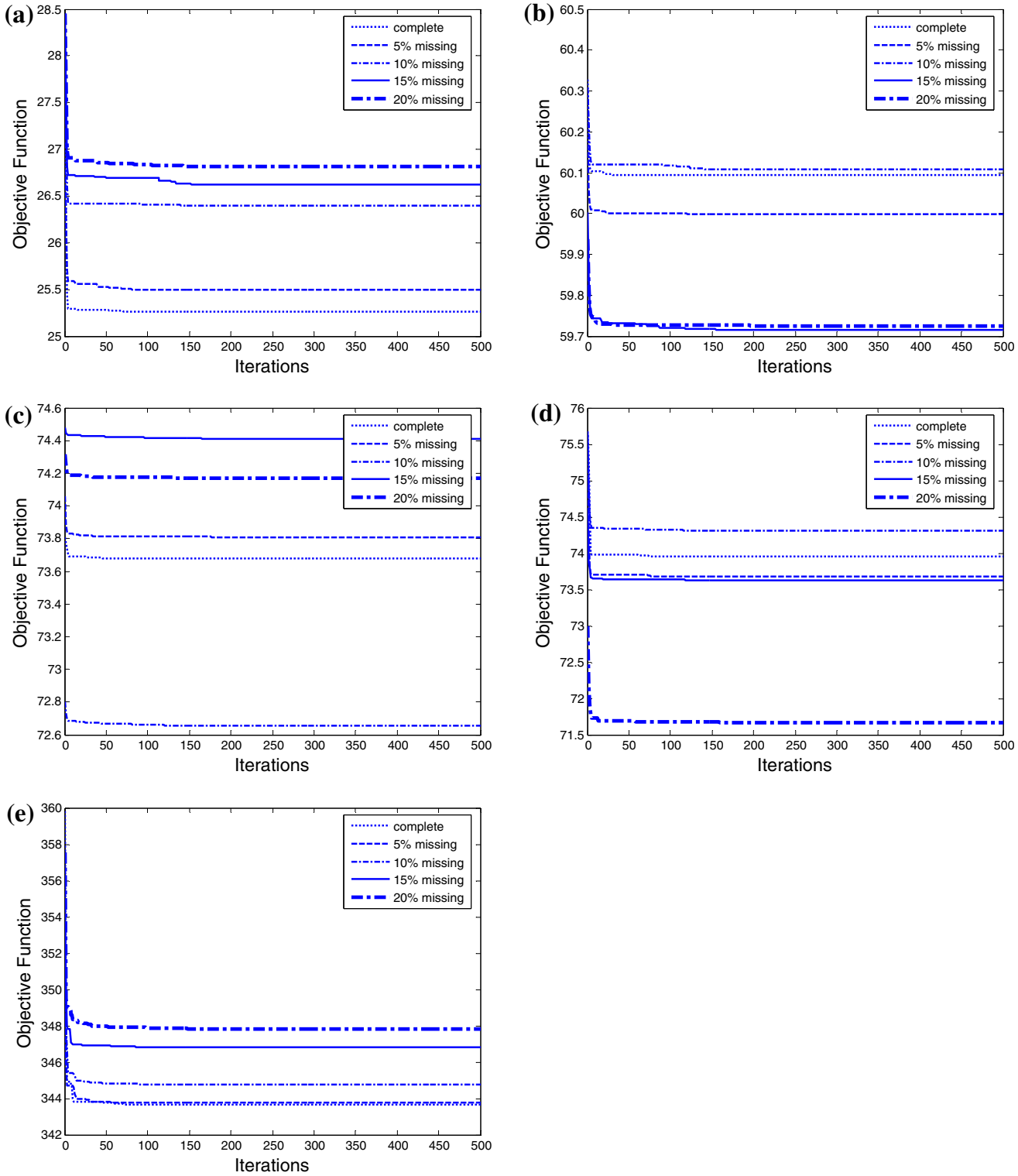

Fig. 2 The change curves between target function with iterations of data sets. **a** IRIS, **b** Wine, **c** Bupa, **d** Haberman and **e** Breast

used for the interval-valued data set clustering. The hybrid algorithm utilizes the global optimization capability of particle swarm to optimize the cluster prototypes in FCM, which can make the algorithm get more accurate clustering results. The experimental results show that NIR hybrid algorithm has more advantages over other methods in accuracy and more effective when it is applied to the incomplete data clustering.

Acknowledgments This work is supported by the National Nature Science Foundation of China (No. 61174115, No. 51104044).

References

1. Chen M, Miao DQ (2011) Interval set clustering. Expert Syst Appl 38(4):2923–2932
2. Wang J, Chung FL, Wang ST, Deng ZH (2013) Double indices-induced FCM clustering and its integration with fuzzy subspace clustering. Pattern Anal Appl 6:1433–7541
3. Chang CT, Lai JZ, Jeng MD (2011) A fuzzy K-means clustering algorithm using cluster center displacement. J Inf Sci Eng 27(3):995–1009
4. Taherdangkoo M, Bagheri MH (2013) A powerful hybrid clustering method based on modified stem cells and Fuzzy C-means algorithms. Eng Appl Artif Intell 26(5–6):1493–1502
5. Abas AR (2010) Using general regression with local tuning for learning mixture models from incomplete data sets. Egypt Inform J 11(2):49–57
6. Abas AR (2012) Unsupervised learning of mixture models based on swarm intelligence and neural networks with optimal completion using incomplete data. Egypt Inform J 13(2):103–109
7. Lin HC, Su CT (2013) A selective Bayes classifier with meta-heuristics for incomplete data. Neurocomputing 15(106):95–102
8. Hathaway RJ, Bezdek JC (2001) Fuzzy c-means clustering of incomplete data. IEEE Trans Syst Man Cybern Part B Cybern 31(5):735–744
9. Dixon JK (1979) Pattern recognition with partly missing data. IEEE Trans Syst Man Cybern 9(10):617–621
10. Di Nuovo AG (2011) Missing data analysis with fuzzy C-means: a study of its application in a psychological scenario. Expert Syst Appl 38(6):6793–6797
11. Aydilek IB, Arslan A (2013) A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci 233:25–35
12. Simiński K (2013) Clustering with missing values. Fundam Inform 123(3):331–350
13. Nowicki RK (2010) On classification with missing data using rough-neuro-fuzzy systems. Int J Appl Math Comput Sci 20(1):55–67
14. Dopazo E, Ruiz-Tagle M (2011) A parametric GP model dealing with incomplete information for group decision-making. Appl Math Comput 218(2):514–519
15. Pei Z (2012) Rational decision making models with incomplete weight information for production line assessment. Inf Sci 222(10):696–716
16. Himmelspach L, Conrad S (2010) Fuzzy clustering of incomplete data based on cluster dispersion. Comput Intell Knowl Based Syst Des 6178:59–68
17. Zhang SC, Jin Z, Zhu XF (2011) Missing data imputation by utilizing information within incomplete instances. J Syst Softw 84(3):452–459
18. Subasi MM, Subasi E, Anthony M, Hammer PL (2011) A new imputation method for incomplete binary data. Discrete Appl Math 159(10):1040–1047
19. Hathaway RJ, Bezdek JC (2002) Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm. Pattern Recogn Lett 23(1):151–160
20. Sánchez JS, Mollineda RA, Sotoca JM (2007) An analysis of how training data complexity affects the nearest neighbor classifiers. Pattern Anal Appl 10(3):189–201
21. Franco A, Maltoni D, Nanni L (2010) Data pre-processing through reward–punishment editing. Pattern Anal Appl 13(4): 367–381
22. Doquire G, Verleysen M (2012) Feature selection with missing data using mutual information estimators. Neurocomputing 90:3–11
23. Van Hulse J, Khoshgoftaar TM (2011) Incomplete-case nearest neighbor imputation in software measurement data. In: Proceedings of Information Sciences, pp 1–15
24. Li D, Gu H, Zhang L (2010) A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. Expert Syst Appl 37(10):6942–6947
25. Izakian H, Abraham A (2011) Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Syst Appl 38(3):1835–1838
26. Benaichouche AN, Oulhadj H, Siarry P (2013) Improved spatial fuzzy c-means clustering for image segmentation using PSO initialization, Mahalanobis distance and post-segmentation correction. Digit Signal Process 23(5):1390–1400
27. Yu SW, Wei YM, Fan JL, Zhang X, Wang K (2012) Exploring the regional characteristics of inter-provincial CO_2 emissions in China: an improved fuzzy clustering analysis based on particle swarm optimization. Appl Energy 92:552–562
28. Omran MG, Salman A, Engelbrecht AP (2006) Dynamic clustering using particle swarm optimization with application in image segmentation. Pattern Anal Appl 8(4):332–344
29. Mohandes MA (2012) Modeling global solar radiation using particle swarm optimization (PSO). Sol Energy 86(11):3137–3145
30. Farahmand H, Rashidinejad M, Mousavi A, Gharaveisi AA, Irving MR, Taylor GA (2012) Hybrid mutation particle swarm optimization method for available transfer capability enhancement. Int J Electr Power Energy Syst 42(1):240–249
31. Zhang L, Zhao JQ, Zhang XN, Zhang SL (2013) Study of a new improved PSO-BP neural network algorithm. J Harbin Inst Technol 20(5):99–105