# Interval Fuzzy C-means Approach for Incomplete Data Clustering Based on Neural Networks

Li Zhang[1], Hui Pan[1], Beilei Wang[1], Liyong Zhang[2], Zhangjie Fu[3]

[1] School of Information, Liaoning University, China
[2] School of Control Science and Engineering, Dalian University of Technology, China
[3] School of Computer and Software, Nanjing University of Information Science & Technology, China
zhang_li@lnu.edu.cn, iwajob2014@163.com, zhly@dlut.edu.cn,{1048895252, 155128115}@qq.com

## Abstract

In the field of data mining and machine learning, the problem of recovering missing values from a dataset has become an important research issue. Recently, the numerical values may not be suitable for describing the uncertainty of missing attributes, and there is a certain degree of error. Hence, we propose an efficient interval approach which utilizes a missing-data back propagation to estimate the error value of the complete property of the missing samples and convert the value of the missing attribute to the form of an interval. Furthermore, fuzzy C-means performs clustering analysis on the recovered data set. Therefore, the numerical data set is converted into an interval valued fuzzy C means clustering analysis, and the final clustering results are obtained. The experimental results demonstrate that our algorithm has good accuracy in data clustering performance.

**Keywords**: Incomplete data, Interval valued estimation, Fuzzy C-means clustering

## 1 Introduction

Clustering is the process of grouping data objects into a set of disjoint classes, called clusters, so that objects within a class have high similarity to each other, while objects in separate classes are more dissimilar [1]. Clustering analysis is an unsupervised classification method, which has a significant impact on all fields. The Fuzzy C-means (FCM) algorithm is a useful tool for clustering real s-dimensional data, but it is not directly applicable to the case of incomplete data [2]. However, in practice, missing values occur with collection of any data set due to various reasons including equipment malfunctioning, human errors, and faulty data transmission. If an organization does not take extreme care during data collection, then approximately 5% or more missing/corrupt data may be introduced in the data sets [3-4]. Any missing data in the database could prevent the discovery of important factors, and lead to invalid conclusions. The problem of recovering missing values from a dataset has become an important research issue in the field of data mining and machine learning [5]. Hence, the effective treatment of missing data is an important problem in the real world. Dealing with missing data effectively can help us to make full use of the information of datasets, so as to improve the accuracy and robustness of fuzzy clustering analysis results.

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. [6-9]. As far as we know, the interval valued fuzzy clustering theory has been used to handle incomplete data recently.

Deb [10] proposed a novel imputation method that exploits the within-record and between-record correlations to impute missing data of numerical or categorical values. Hong et al. [11] dealt with the problem of learning from incomplete quantitative data sets based on rough sets, and proposed an algorithm that can simultaneously derive certain and possible fuzzy rules from incomplete quantitative data sets and estimate the missing values in the learning process. Then he introduced an iterative missing-value completion method based on the robust association rules to extract useful association rules for inferring missing values in an iterative way [12]. Mutual information is one of the widely used criteria in feature selection, which determines the relevance between features and target classes. Qian et al. first validated the feasibility of the mutual information. And then an effective mutual information-based feature selection algorithm with a forward greedy strategy was developed in incomplete data [13].

Gao et al. [14] put forward two interval valued data fuzzy C-means clustering algorithms and then developed these two kinds of algorithms according to considering influence factors of interval size on clustering results, and proposed using weighted factor to control the interval size of clustering effect. Yue et

*Corresponding Author: Hui Pan; E-mail: iwajob2014@163.com

DOI: 10.3966/160792642018081904012

al. [15] proposed a new type of interval valued data fuzzy clustering algorithm, which adopted interval partition strategy to calculate the distance between interval data, and improved the flaws of the interval distance calculation. In order to reduce the complexity of computation, the interval fuzzy sets were presented by [16], which unified the weights of the secondary of the interval valued fuzzy sets and reduced the complexity of the computation. Interval fuzzy sets have been widely used in pattern recognition and clustering analysis [17-18].

Bing et al. used particle swarm global optimization ability to find the best estimates of the missing attributes in recent neighboring interval, and optimized the fuzzy c-means clustering center to get the better clustering results. After that, she put forward the nearest neighbor interval reconstruction rule. The nearest neighbor interval is reconstructed according to the result of the pre classification, which excludes the nearest neighbor samples from the same class as the missing samples to get the more accurate missing attribute interval [19].

With the further research on the fuzzy clustering algorithm, it finds that data clustering performance is not satisfactory due to incomplete data. Using the interval valued data to express the fuzziness of the boundary between data items is more effective [20-22]. We propose a fuzzy C-means algorithm based on Missing-data Back Propagation (MBP) neural network. The experimental verification of the University of California, Irvine (UCI) data set and the artificial data set are used to show that our algorithm has good accuracy in data clustering.

The remainder of this paper is organized as follows. In order to make this contribution self-contained, a brief review of theoretical review is given in Section 2. Section 3 presents our proposed interval fuzzy C-means (IFCM) algorithm and its application analysis. In Section 4, we present our experimental results. Finally, the conclusion is drawn in Section 5.

## 2 Theoretical Review

### 2.1 MBP Neural Network

For the basic Back Propagation (BP) neural network, the training sample contains the missing attribute, besides, the error between the prediction output and the expected output of the output layer of the missing attribute cannot be calculated. Therefore, the weights and thresholds of the neural network cannot be adjusted. But in the testing process of MBP neural network, the output of the output layer uses the network to fill in the missing attribute data of the estimation of missing attribute. The average value of all integrity property valuation error is replaced by the estimation error of missing attribute, in order to satisfy the network weights and thresholds for correction. After several iterations of learning, the BP neural network trained with the corresponding missing data is obtained.

#### 2.1.1 Sample Selection and Optimization

The selection of training samples can greatly affect the performance of the network. The conditions for selecting the training samples are as follows: the number of training samples is large enough, and the training samples must be complete so that it can include all the characteristics of the data in the sample.

**Selection of training samples.** Let s-dimensional incomplete data set $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ contain at least one incomplete datum with some (but not all) missing attribute values. For missing data sets, the partial distance [23] of $\tilde{x}_a$ and an instance $\tilde{x}_b$ (incomplete or complete) is calculated using formula (1):

$$D_{ba} = \frac{s}{\sum_{i=1}^{s} I_i} \sum_{i=1}^{s} \left(x_{ia} - x_{ib}\right)^2 I_i \quad i = 1, 2, ..., s \tag{1}$$

Where $x_{ia}$ and $x_{ib}$ are the $i_{th}$ attribute of $\tilde{x}_a$ and $\tilde{x}_b$ respectively, and the value of $I_i$ is 0 or 1. If both $x_{ia}$ and $x_{ib}$ are non-missing, $I_i$ is 1; otherwise, $I_i$ is 0.

The partial distance formula is used to calculate the distance, which is between each missing datum and all other data, and then the distance values are deposited into the corresponding matrix in ascending order. The smaller the distance value is, the lager the similarity of each attribute values is. According to the nearest neighbor rule, the nearest neighbor sample set is selected as the preparatory training sample set. The selected nearest neighbor sample set contains not only the complete data, but also the missing data. This selection method can let the missing data as training sample to estimate the missing attribute. Furthermore, it can make full use of both the complete and missing attribute information of the data.

**Optimization of training samples.** For each missing data, the missing position of the missing data is processed by preparatory training sample set. In this paper, the sample set of optimization is used as the first half of the training sample set, and the original training sample set is the latter part of the training sample set. The network will be fine tuned using the latter part of the training sample, so that the performance of the network to get the overall improvement.

As shown in Figure 1, for a randomly generated 2-dimensional data set $\tilde{X}$, which contains 10 data samples and the first dimension attribute of the sample data (?, 0.1) is missing. It is represented by the horizontal line and others are represented by dots. Due to the existence of missing attributes, the partial distance formula is used as the similarity measure of the missing samples and all the other samples. In this case, calculate the partial distance values of the missing samples (?, 0.1) and the other nine data samples. The smaller the partial distance value is, the more the similarity is. According to the similarity of each sample and missing data values in descending order, and the large similarity of the data samples is selected as the pre-training sample set. Four larger black points, which are represented by the similarity measure, are selected as the nearest neighbor sample set for missing data samples in Figure 1.

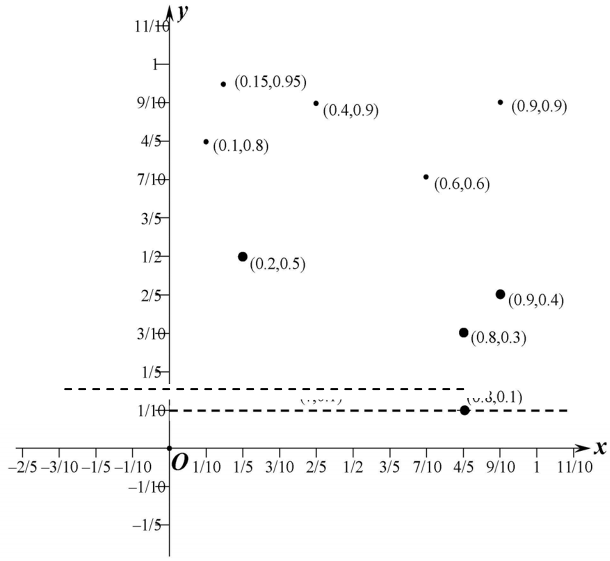

**Figure 1.** Selection of training samples

### 2.1.2 Sample Selection and Optimization

In the acquisition of data samples, because some interference factors will produce missing data sample, overmuch missing data samples for neural network training leads to serious influence that affects the performance of the neural network. Therefore, the preprocessing of the missing data is an important part of the training phase of the neural network. MBP neural network algorithm process is as follows:

**Step1:** Normalize input data set, that is, all the data are converted to a number of interval [0, 1].

**Step2:** Determinate and optimize training sample.

**Step3:** Initialize network. Determine the input nodes *n*, hidden nodes *l* and output nodes *m*; initialize the network weights $w_{ij}$, $w_{jk}$ and the thresholds *a* and *b*; determine the maximum number of training *M*, error accuracy $\varepsilon_1$ and learning rate $\eta$.

**Step 4:** Calculate hidden layer output $sH$ using (2).

$$h_j = f(\frac{n}{I}\sum_{i=1}^{n} w_{ij}x_iI_i - a_j),\ j=1,2,\ldots,l \quad \textbf{(2)}$$

Where
$$I=\sum_{i=1}^{n} I_i \quad \textbf{(3)}$$

$$I_i = \begin{cases} 0, & if\ x_i\ is\ missing; \\ 1, & otherwise\quad ; \end{cases} \quad \textbf{(4)}$$

where $x_i$ is the $i_{th}$ attribute of $x$ respectively, $n/I$ is Input layer node number recovery factor, and implicit layer excitation function $f$ is as follows:

$$f(x)=\frac{1}{1+e^{-x}} \quad \textbf{(5)}$$

**Step 5:** Calculate the outputs of the neural network $O$ using (6).

$$O_k = \sum_{j=1}^{l} H_j w_{jk} - b_k),\quad k=1,2,\ldots,m \quad \textbf{(6)}$$

**Step 6:** Calculate error $e_k$ using (7).

$$e_k = \begin{cases} \bar{e}, & if\ Y_k\ is\ missing; \\ Y_k - O_k, & otherwise\quad ; \end{cases} \quad k=1,2,\ldots,m \quad \textbf{(7)}$$

where $Y_k$ is $k_{th}$ attribute of sample expected output, $O_k$ is $k_{th}$ attribute of sample expected output, and $\bar{e}$ is the mean of the actual output and the expected output error of all the complete attributes of the data sample $\tilde{Y}$.

**Step 7:** Update network weights $w_{ij}$ and $w_{jk}$ using (8) and (10).

$$w_{ij} = w_{ij} + \eta H_j(1-H_j)x(i)I_i\sum_{k=1}^{m} w_{jk}e_k$$
$$i=1,2,\ldots,n;\ j=1,2,\ldots,l \quad \textbf{(8)}$$

$$I_i = \begin{cases} 0, & if\ x_i\ is\ missing; \\ 1, & otherwise\quad ; \end{cases} \quad \textbf{(9)}$$

$$w_{jk} = w_{jk} + \eta H_j e_k,\ j=1,2,\ldots,l,\ k=1,2,\ldots,m \quad \textbf{(10)}$$

**Step 8:** Update threshold *a* and *b* using (11) and (12).

$$a_j = a_j + \eta H_j(1-H_j)\sum_{k=1}^{m} w_{jk}e_k,\ j=1,2,\ldots,l \quad \textbf{(11)}$$

$$b_k = b_k + e_k,\ k=1,2,\ldots,m \quad \textbf{(12)}$$

**Step 9:** Terminate the iterations if $e<\varepsilon_1$, or iteration number is greater than the maximum number of training; otherwise, increase the iteration ($l=1+1$) repeat steps 3 through 9.

## 2.2 Interval Valued Fuzzy C-means (IFCM) Clustering Algorithm

**FCM algorithm.** The FCM algorithm proposed by Bezdek purposes is to partition a numerical object data set $X=\{x_1,x_2,\ldots,x_n\}\subset R^s$ into c clusters. The objective function of the interval FCM is:

$$J(U,V)=\sum_{i=1}^{C}\sum_{J=1}^{n}u_{ik}^{m}\|x_k-v_i\|_2^2 \tag{13}$$

with the constraint of

$$\sum_{i=1}^{c}u_{ik}=1, k=1,2,\ldots,n \tag{14}$$

The updating formulas of the membership degree $u_{ik}$ and the cluster prototype $v_i$ are as follows:

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}x_k}{\sum_{k=1}^{n}u_{ik}^{m}}, i=1,2,\ldots,c, \tag{15}$$

$$u_{ik}=\left[\sum_{t=1}^{c}\left(\frac{\|x_k-v_i\|_2^2}{\|x_k-v_t\|_2^2}\right)^{\frac{1}{m-1}}\right]^{-1}, \quad \begin{array}{l} i=1,2,\ldots,c; \\ k=1,2,\ldots,n; \end{array} \tag{16}$$

**IFCM algorithm.** The IFCM algorithm is used to cluster analysis of interval valued data sets in this paper. Let $\overline{X}=\{\overline{x_1},\overline{x_2},\ldots,\overline{x_n}\}$ be interval valued data sets, $n$ be the number of data sets, and data sample $\overline{x_j}(1\le j\le n)$ expressed as $\overline{x_j}=[\overline{x_{1j}},\overline{x_{2j}},\ldots,\overline{x_{sj}}]^T$, and each attribute value in the data sample $\overline{x}_k$ is expressed as the interval, which is $\overline{x_{kj}}=[\overline{x_{kj}},x_{kj}^{+}](1\le k\le s)$. Data set $\overline{X}$ is divided into $c$ classes, and its clustering center is expressed as $\overline{V}=[\overline{V_{1k}}]=[\overline{v_1},\overline{v_2},\ldots,\overline{v_c}]$, where $\overline{v_{ik}}=[\overline{v_{1k}},v_{ik}^{+}]$ $(i=1,2,...,c;k=1,2,...,s)$.

In many applications it is useful to use the median instead of the mean to measure the center of a group. Essentially, one uses the median because it is not sensitive to outliers, but this robustness comes at a price: computing the median takes much longer than computing the mean. Sometimes one finds the median as a step in a larger iterative process (like in many optimization algorithms), and this step is the bottleneck. Therefore, we use the mean to measure the center of a group to avoid time consuming and to meet the need of real time processing. The basic flow chart of the IFCM is shown in Figure 2.

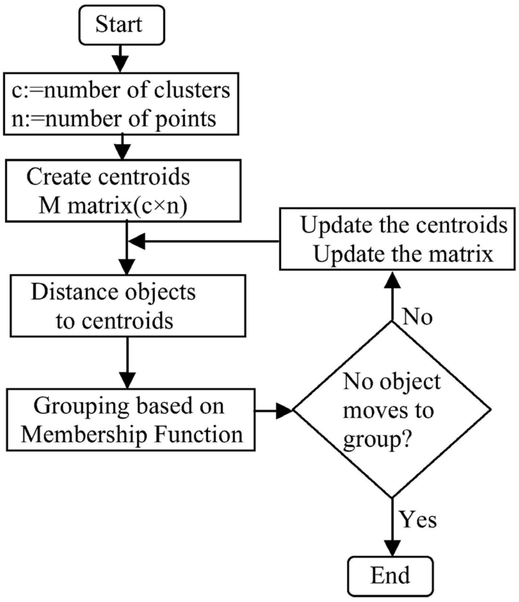

**Figure 2.** The basic flow chart of the IFCM

The main process of the IFCM algorithm is as follows:

**Step 1:** Initialize network, set clustering number $c(2\le c\le n)$, where $n$ is data sample size, and $G$ is the maximum iteration number; determine fuzzy weighting coefficient $m$ and iterative termination threshold $\varepsilon$, and initialize the partition matrix $U^{(0)}$.

**Step 2:** Update formula of clustering center, according to $U^{(l-1)}$, when the iteration to $l(l=1,2,...)$ times, and calculate the left interval value $\overline{V}^{(1)-}$ and the right interval value $\overline{V}^{(1)+}$ of clustering center $\overline{V}^{(1)}$ using (17) and (18).

$$v_i^{-}=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j^{-}}{\sum_{j=1}^{n}u_{ij}^{m}}, i=1,2,\ldots,c \tag{17}$$

$$v_i^{+}=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j^{+}}{\sum_{j=1}^{n}u_{ij}^{m}}, i=1,2,\ldots,c \tag{18}$$

**Step 3:** Update the membership matrix, according to $\overline{V}^{(1)}$, update the membership matrix $U^{(1)}$ using (16) and (19).

$$u_{ij}=\begin{cases}0, i\ne h\\ 1, i=h\end{cases} \tag{19}$$

**Step 4:** Terminate the iterations if $\max|U^{(l+1)}-U^{(1)}|\le\varepsilon$, or iteration number $1>G$; otherwise, increase the iteration ($l=1+1$) repeat steps 2 through 4.

## 3 MBP-IFCM Algorithm for Incomplete Data Clustering

### 3.1 Conversion between Missing Data Sets and Interval Data Sets

**The interval estimation of missing attributes.** The estimation error of complete data which is calculated by MBP neural network is used to determine the estimation interval of missing values. In the prediction stage of MBP neural network, the estimation value of missing attributes can be obtained by the output value of MBP output neurons. So does the estimation of complete data. Therefore, for the complete data, the discrepancy between expected output value and actual output value is obtained by the squared error measure. It can be used to get the maximum absolute error that the estimated value is less than the expected value, and the maximum absolute error that the estimated value is greater than the expected value. And the two largest absolute errors are defined as the boundary value of the left and right boundaries of the missing attribute. It can limit the value of the missing attribute to a reasonable range.

Let 4-dimension incomplete data set $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$, with $s$ missing attributes. The third attribute value $x_{33}$ of the data sample $\tilde{x}_3 =$ (0.3889, 0.7500, ?, 0.0833) is missing. After the valuation of the MBP network, the output value is $\tilde{x}_{o3} =$ (0.3995, 0.6931, 0.1132, 0.0879). The maximum value of the error absolute value, which is determined by the estimated value of the full property that is less than the expected output value, is $e_{rmax} = 0.056$. And the maximum value of the error absolute value, which is determined by the estimated value of the full property that is greater than the expected output value, is $e_{rmax} = 0.0106$. Thus, the missing attribute value interval is $[0.1132 - e_{lmax}, 0.1132 + e_{rmax}]$, that is [0.0563, 0.1238].

As shown in Figure 3, $e$ represents the estimated value of the MBP network for missing attributes, $e_i$ represents the left boundary value for the missing attribute value range, and $e_r$ represents the right boundary value for the missing attribute value range. After normalization, all the attributes of the data sample are between 0 and 1.

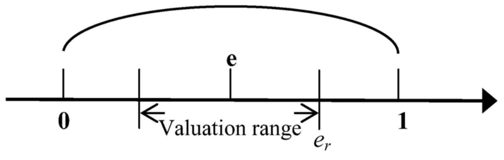

**Figure 3.** Determination of missing attribute value range

**Conversion of interval valued data sets.** The missing attribute is estimated through the MBP network, and then the numerical value of the missing attribute is converted into interval valued by the estimation error of full attribute with MBP network. Therefore, all of the complete attribute values of missing data need to be transformed into interval valued. A multistep transformation process is as follows:

**Step 1:** Estimate the missing data using MBP network, and obtain the estimated value error between the missing attribute and complete attribute.

**Step 2:** Compare each complete attribute value error of missing data samples to get the maximum error absolute value $e_{lmax}$ that the estimated value of the full property is less than the expected output value, and the maximum error absolute value $e_{rmax}$ that the estimated value of the full property is greater than the expected output value.

**Step 3:** Convert the numerical value of the missing attribute to interval valued estimation, and missing attribute value interval is $[e - e_{lmax}, e + e_{rmax}]$, that is $[e_l, e_r]$.

**Step 4:** Judge whether the valuation range of the missing attributes in the [0, 1]. If $e_1 < 0$, set the value of the missing attribute to the left boundary value to 0; if $e_r < 1$, set the value of the missing attribute to the left boundary value to 1.

**Step 5:** Express the missing data in the full range of all the properties expressed as interval, that is, the left and right boundaries of the interval are equal and they are the original value of the full property.

### 3.2 MBP-IFCM Algorithm

The MBP-IFCM algorithm is first used to process the missing data sets and the estimated value of the missing attribute. Furthermore, the numerical value is converted into the form of interval. Besides, the full property is also transformed into an interval. Finally, the analysis of interval valued data sets is carried out. The flow diagram is shown in Figure 4 and the concrete steps are as follows:

**Step 1:** Normalize the input data set. Turn all the data into a number between 0 and 1, so as to eliminate the difference between the number of dimensions of data.

**Step 2:** Determine and optimize the training samples. The similarity of each sample and the other is calculated by using the partial distance formula. The nearest neighbor training sample set based on nearest neighbor rule is determined for each missing attribute. The missing data processing is done for the corresponding properties of the nearest neighbor training sample, which is determined by the location of the missing attribute of each sample. Then, take the collection set of optimized training and original training as the training sample set.

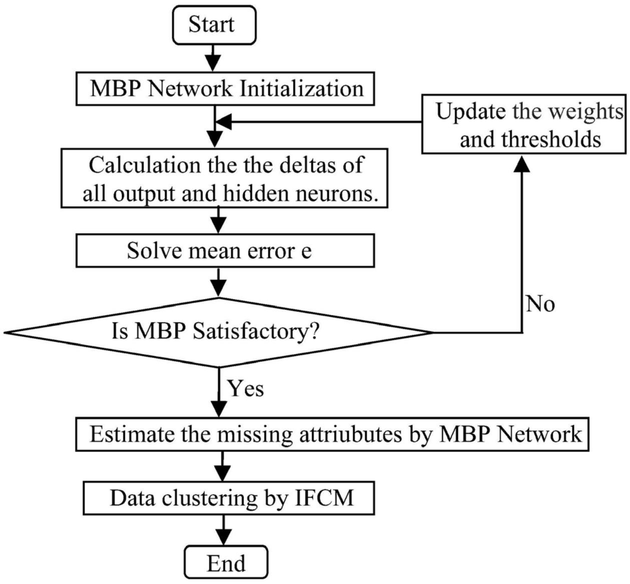
Start
MBP Network Initialization
Update the weights and thresholds
Calculation the the deltas of all output and hidden neurons.
Solve mean error e
No
Is MBP Satisfactory?
Yes
Estimate the missing attriubutes by MBP Network
Data clustering by IFCM
End

**Figure 4.** The flow diagram of MBP-IFCM algorithm

**Step3:** Initialize the network. Determine the numbers of nodes $n, l, m$, and initialize weights $w_{ij}, w_{jk},$ and $a, b$. Determine the maximum number of training $M$, error accuracy $\varepsilon_1$, and learning rate $\eta$.

**Step 4:** Train the MBP network. Using the training sample set of the missing attributes to train the MBP network, and the neural network is trained for each missing attribute.

**Step5:** Estimate the missing attribute. Using the trained MBP network to estimate the value of each attribute, and then get the estimated value of the missing attribute and the estimated value of the complete attribute in the missing data via the MBP network.

**Step 6:** Determine the value of the missing attribute interval. Calculate the error between the estimated value and the expected output value of the full attribute, and get the maximum error absolute value $e_l$, where the estimates value is less than the expected value, and the maximum error absolute value $e_r$, where the estimated value is greater than the expected value. The two maximum errors absolute values of $e_l$ and $e_r$ are used as the left and right boundary values of the missing attribute values. Furthermore, it guarantees to be obtained in the range of 0 to 1. Thus the value of the data can be gotten to focus on all the missing attributes of the valuation range $[e_l, e_r] \subset [0,1]$.

**Step7:** Transform valuation range data set. All the integrity of the data set $x_{ij}^-$ is converted into the form of interval $[x_{ij}^-, x_{ij}^+]$, and $x_{ij}^- = x_{ij} = x_{ij}^+$. Convert the entire numerical data set into the valuation interval data set.

**Step 8:** Initialize parameters of IFCM, determine the clustering number $c$, the maximum number of iterations $G$, the fuzzy weighting coefficient $m$ and the iterative termination threshold $\varepsilon$, and initial membership matrix $U^{(0)}$.

**Step 9:** Update formula of clustering center: according to $U^{(l-1)}$, when the iteration to $l(l = 1, 2, \ldots)$ times, calculate the left interval value $V^{(1)-}$ and the right interval value $V^{(1)+}$ of clustering center $V^{(1)}$ using (16) and (17).

**Step 10:** Update the membership's matrix: according to $V^{(1)}$, namely, update the memberships matrix $U^{(1)}$ using (15) and (18).

**Step11:** Judge the algorithm whether satisfies the terminated condition. Terminate the iterations if $\max|U^{(l+1)} - U^1| \le \varepsilon$, or iteration number $1 > G$; otherwise, increase the iteration $(l = l + 1)$ repeat steps 9 through 11.

## 4 Experimental Results and Discussion

### 4.1 Experimental Results

Three data sets of the UCI database: Wine, Bupa, Breast and two artificial data sets are selected to perform the simulation experiments. The information of data sets is shown in Table 1.

**Table 1.** The information of data sets

| The data sets | Number of samples | Number of attributes | Number of classes |
|---|---|---|---|
| Wine | 178 | 13 | 3 |
| Bupa | 345 | 7 | 2 |
| Breast | 683 | 9 | 9 |
| Artificial data sets 1 | 200 | N/A | 2 |
| Artificial data sets 1 | 350 | N/A | 3 |

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

The Bupa Liver Disorder data set contains 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. Namely, the two output clusters are liver disorder patient and normal patient that are represented by 0 and 1 respectively.

The Breast data set contains nine attributes describing the details of each case. Besides, two additional attributes are sample code number and class attribute: malignant and benign.

**Artificial data sets.** Two artificial data sets are generated using the data generation method given by B

A Pimentel [24]. The artificial data set I contains 200 samples that are divided into two clusters, each of which contains 100 data samples in two-dimensional space. The artificial data set II contains 350 samples that are divided into three clusters, each of which contains 50, 80, 220 data samples in two-dimensional space.

The data sample points are subject to an independent two-dimensional normal distribution, and the expectation and variance matrix of the two artificial data sets are defined as follows:

$$u\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \text{ and } \Sigma\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

The data sample points of each class in the artificial data set I are generated according to the following parameters:

(1) The first category: $u_1 = 3, u_2 = 5, \sigma_1^2 = 1, \sigma_2^2 = 2$.

(2) The second category: $u_1 = 6, u_2 = 8, \sigma_1^2 = 1, \sigma_2^2 = 2$.

The data sample points of each class in the artificial data set II are generated according to the following parameters:

(1) The first category: $u_1 = 48, u_2 = 25, \sigma_1^2 = 4, \sigma_2^2 = 2$.

(2) The second category: $u_1 = 60, u_2 = 30, \sigma_1^2 = 9, \sigma_2^2 = 25$.

(3) The third category: $u_1 = 45, u_2 = 38, \sigma_1^2 = 16, \sigma_2^2 = 16$.

## 4.2 Discussion

The experimental results of the method in this paper (MBP-IFCM) are compared with the method WDS-FCM, PDS-FCM, OCS-FCM, NPS-FCM, and MBP-FCM proposed by Wang [21], thereby to verify the effectiveness of the method in the paper. The missing rate is taken as 5%, 10%, 15% and 20%. Experimental results are shown from Table 2 to Table 8, and the optimal results are marked by bold types.

**Table 2.** Averaged number of misclassification results of 10 trails using incomplete Wine data set

| %missing | The average number of misclassification | | | | |
|---|---|---|---|---|---|
| | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | MBP-IFCM |
| 5 | 10.3 | 10.0 | 10.0 | 9.9 | **9.3** |
| 10 | 12.7 | 10.2 | 10.7 | 10.1 | **9.6** |
| 15 | 21.8 | 12.4 | 13.2 | 12.5 | **10.0** |
| 20 | 45.2 | 12.0 | 12.7 | 11.9 | **9.7** |

**Table 3.** Averaged number of misclassification results of 10 trails using incomplete Bupa data set

| %missing | The average number of misclassification | | | | |
|---|---|---|---|---|---|
| | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | MBP-IFCM |
| 5 | 177.4 | 177.2 | 177.5 | 177.2 | **176.2** |
| 10 | **176.4** | 177.0 | 176.6 | 177.0 | **176.4** |
| 15 | 177.8 | 178.5 | 177.5 | 178.4 | **176.7** |
| 20 | 178.3 | 179.1 | 177.4 | 179.0 | **176.0** |

**Table 4.** Averaged number of misclassification results of 10 trails using incomplete Breast data set

| %missing | The average number of misclassification | | | | |
|---|---|---|---|---|---|
| | WDS-FCM | PDS-FCM | OCS-FCM | NPS-FCM | MBP-IFCM |
| 5 | 51.3 | 30.9 | 31.8 | 31.9 | **29.1** |
| 10 | 73.1 | 31.0 | 31.7 | 31.8 | **29.2** |
| 15 | 65.1 | 33.6 | 33.3 | 32.9 | **30.2** |
| 20 | 64.0 | 34.1 | 36.2 | 34.7 | **29.3** |

**Table 5.** Averaged number of misclassification results of 10 trails using incomplete UCI data set

| %missing | The average number of misclassification | | | | | |
|---|---|---|---|---|---|---|
| | Wine | | Bupa | | Breast | |
| | MBP-FCM | MBP-IFCM | MBP-FCM | MBP-IFCM | MBP-FCM | MBP-IFCM |
| 5 | **9.3** | **9.3** | 176.5 | **176.2** | 29.5 | **29.1** |
| 10 | 9.7 | **9.6** | 176.9 | **176.4** | **29.2** | **29.2** |
| 15 | 10.4 | **10.0** | 178.0 | **176.7** | 30.5 | **30.2** |
| 20 | 9.9 | **9.7** | 177.3 | **176.0** | 30.3 | **29.3** |

**Table 6.** Averaged number of misclassification results of 10 trails using incomplete artificial data set

| %missing | The average number of misclassification | | | |
|---|---|---|---|---|
| | artificial data sets 1 | | artificial data sets 2 | |
| | MBP-FCM | MBP-IFCM | MBP-FCM | MBP-IFCM |
| 5 | 6.8 | **5.9** | 28.5 | **28.1** |
| 10 | **7.1** | 8.0 | 38.0 | **33.5** |
| 15 | 10.4 | **10.1** | 41.6 | **40.6** |
| 20 | 13.9 | **12.3** | 60.9 | **47** |

**Table 7.** Averaged iteration number results of 10 trails using incomplete UCI data set

| %missing | Iteration Number | | | | | |
|---|---|---|---|---|---|---|
| | Wine | | Bupa | | Breast | |
| | MBP-FCM | MBP-IFCM | MBP-FCM | MBP-IFCM | MBP-FCM | MBP-IFCM |
| 5 | **25.5** | 25.6 | 41.2 | **40.8** | **14.7** | 15.2 |
| 10 | **25.1** | 28.6 | 42.0 | **39.6** | **14.7** | 14.8 |
| 15 | **27.1** | 27.9 | **42.0** | 42.7 | **15.4** | 15.8 |
| 20 | 27.7 | **27.2** | 44.6 | **43.0** | **15.0** | 16.3 |

**Table 8.** Averaged iteration number results of 10 trails using incomplete artificial data set

| %missing | Iteration Number | | | |
|---|---|---|---|---|
| | artificial data sets 1 | | artificial data sets 2 | |
| | MBP-FCM | MBP-IFCM | MMBP-FCM | MBP-IFCM |
| 5 | **12.4** | 13.0 | **40.9** | 42.6 |
| 10 | 13.8 | **13.1** | **44.8** | 47.4 |
| 15 | **14.2** | 15.0 | **49.5** | 52.4 |
| 20 | 15.2 | **15.0** | **50.6** | 52.7 |

MBP-FCM algorithm adopts numerical value estimation, which does not make full use of data set information and causes loss of information, which could degrade the clustering performance. Compared with the method, MBP-IFCM utilizes the information of interval value estimation, which can use the attribute distribution information of data sets sufficiently to training the improved FCM for each missing attribute, thereby improves the accuracy of clustering results.

The results presented in Tables 2 and 5 show misclassification results of WDS-FCM, PDS-FCM, OCS-FCM, NPS-FCM, and MBP-IFCM. And the results presented in Tables 6 and 8 show misclassification results of MBP-FCM and MBP-IFCM. Different methods for handling missing attributes in FCM lead to different clustering results. In general, the averaged number of MBP-IFCM misclassification is the least in the case of different attributes missing rates. The results are slightly worse than MBP-FCM only when the missing rates of Wine was 5% and Breast was 10%.

The convergence analysis is shown from Figure 5 to Figure 9, which describe the change curve between objective function and iterations of three data sets with different missing rates by MBP-IFCM. As it can be seen from these figures, the objective function value is decreasing with the increasing iteration number. The algorithm can obtain convergence by optimization iteration methods on above various data sets with different missing rates.

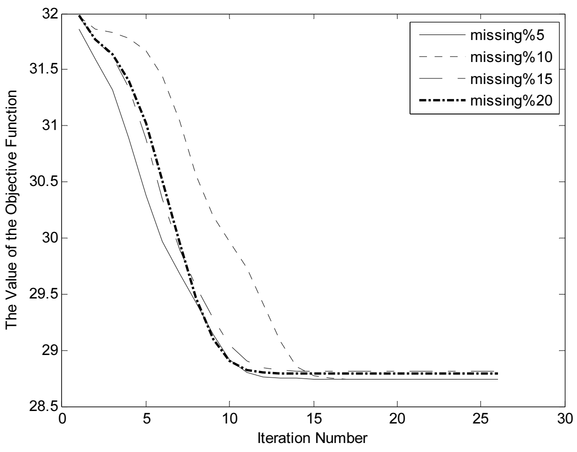

**Figure 5.** The change curve between objective function and iterations of Wine data set

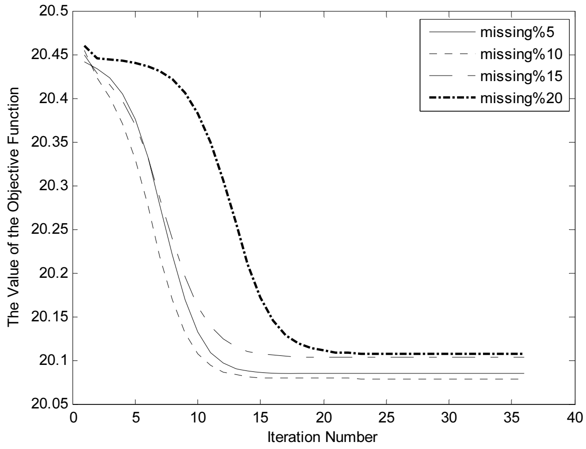

**Figure 6.** The change curve between objective function and iterations of Bupa data set

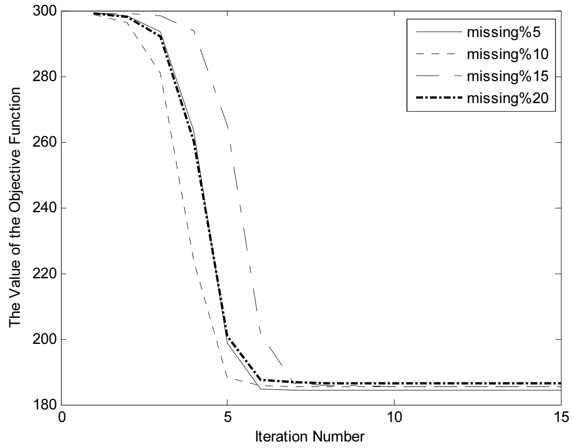

**Figure 7.** The change curve between objective function and iterations of Breast data set

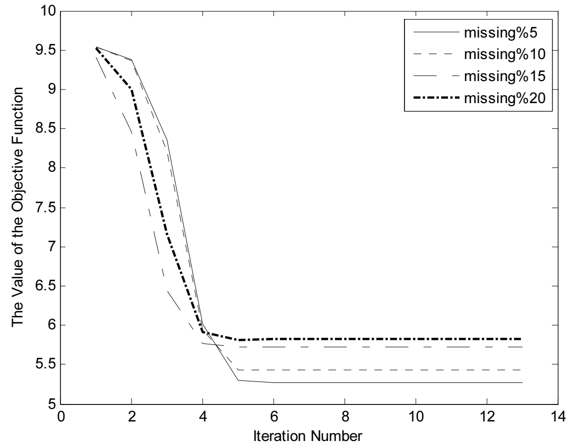

**Figure 8.** The change curve between objective function and iterations of artificial data set 1

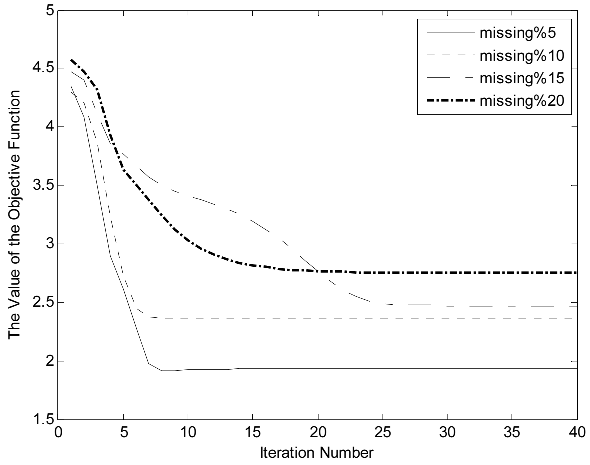

**Figure 9.** The change curve between objective function and iterations of artificial data set 2

## 5 Conclusions

This paper presents a fuzzy c-means algorithm based on MBP for clustering analysis. The proposed algorithm has two main characteristics. Firstly, the missing attributes are replaced by intervals based on the nearest-neighbor information, which makes the estimation of missing attribute more reasonable. Secondly, an MBP can be trained by incomplete data set, which can use the attribute distribution information of data sets sufficiently. Thus, the proposed algorithm can obtain more reasonable imputations of missing attributes and more satisfying clustering results. The experimental results show that the proposed algorithm has better performance than comparative methods in accuracy and more effective when it is applied to the incomplete data classification.

## Acknowledgements

This work is supported by the National Nature Science Foundation of China (No. 61174115). It was also supported by PAPD and CICAEET.

## References

[1] D. X. Jiang, C. Tang, A. D. Zhang, Cluster Analysis for Gene Expression Data: A Survey, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 11, pp. 1370-1386, November, 2004.

[2] R. J. Hathaway, J. C. Bezdek, Fuzzy C-means Clustering of Incomplete Data, *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, Vol. 31, No. 5, pp. 735-744, October, 2001.

[3] K. Takai, Y. Kano, Asymptotic Inference with Incomplete Data, *Communications in Statistics— Theory and Methods*, Vol. 42, No. 17, pp. 3174-3190, April, 2013.

[4] C. S. Lin, PFCF: An Effective VBR Stream Scheduling Algorithm for Clustered Media Systems, *Journal of Internet Technology*, Vol. 7, No. 3, pp. 293-304, July, 2006.

[5] T. P. Hong, L. H. Tseng, B. C. Chien, Mining from Incomplete Quantitative Data by Fuzzy Rough Sets, *Expert Systems with Applications*, Vol. 37, No. 3, pp. 2644-2653, March, 2010.

[6] L. Y. Zhang, W. Lu, X. Liu, W. Pedrycz, C. Zhong, L. Wang, A Global Clustering Approach using Hybrid Optimization for Incomplete Data based on Interval Reconstruction of Missing Value, *International Journal of Intelligent Systems*, Vol. 31, No. 4, pp. 297-313, April, 2016.

[7] Y. H. Zheng, B. Jeon, D. Xu, Q. M. J. Wu, H. Zhang, Image Segmentation by Generalized Hierarchical Fuzzy C-means Algorithm, *Journal of Intelligent & Fuzzy Systems*, Vol. 28, No. 2, pp. 961-973, March, 2015.

[8] B. Gu, X. M. Sun, V. S. Sheng, Structural Minimax Probability Machine, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 7, pp. 1646-1656, July, 2017.

[9] P. Guo, J. Wang, X. H. Geng, C. S. Kim, J. U. Kim, A Variable Threshold-Value Authentication Architecture for Wireless Mesh Networks, *Journal of Internet Technology*, Vol. 15, No. 6, pp. 929-935, November, 2014.

[10] R. Deb, A. W. Liew, Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data, *Information Sciences*, Vol. 339, pp. 274-289, April 2016.

[11] T. P. Hong, L. H. Tseng, S. L. Wang, Learning Rules from Incomplete Training Examples by Rough Sets, *Expert Systems with Applications*, Vol. 22, No. 4, pp. 285-293, May, 2002.

[12] T. P. Hong, C. W. Wu, Mining Rules from An Incomplete Dataset with a High Missing Rate, *Expert Systems with Applications*, Vol. 38, No. 4, pp. 3931-3936, April, 2011.

[13] W. B. Qian, W. H. Shu, Mutual Information Criterion for Feature Selection from Incomplete Data, *Neurocomputing*, Vol. 168, pp. 210-220, November, 2015.

[14] X. B. Gao, J. L. Fan, W. X. Xie, A Novel Algorithm of FCM Clustering for Interval Valued Data, *Journal of Xidian University (Natural Science)*, Vol. 26, No. 5, pp. 604-609, 1999.

[15] M. D. Yue, Novel Fuzzy C-means Clustering Algorithm for Interval Data, *Computer Engineering and Applications*, Vol. 47, No. 13, pp. 157-160, 2011.

[16] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*, https://books.google.com.tw/books/about/Uncertain_Rule_based_Fuzzy_Logic_Systems.html?id=j3NQAAAAMAAJ&redir_esc=y

[17] B. I. Choi, F. C. Rhee, Interval Type-2 Fuzzy Membership Function Generation Methods for Pattern Recognition, *Information Sciences*, Vol. 179, No. 13, pp. 2102-2122, June, 2009.

[18] Z. X. Ji, Y. Xia, Q. S. Sun, G. Cao, Interval-valued Possibilistic Fuzzy C-means Clustering Algorithm, *Fuzzy Sets and Systems*, Vol. 253, pp. 138-156, October, 2014.

[19] L. Zhang, Z. H. Bing, L. Y. Zhang, A Hybrid Clustering Algorithm based on Missing Attribute Interval Estimation for

Incomplete Data, *Pattern Analysis and Applications*, Vol. 18, No. 2, pp. 377-384, May, 2015.

[20] D. D. Nguyen, L. T. Ngo, L. T. Pham, W. Pedrycz, Towards Hybrid Clustering Approach to Data Classification: Multiple Kernels based Interval-valued Fuzzy C-Means Algorithms, *Fuzzy Sets and Systems*, Vol. 279, pp. 17-39, November, 2015.

[21] L. Y. Zhang, W. Pedrycz, W. Lu, X. D. Liu, L. Zhang, An Interval Weighted Fuzzy C-means Clustering by Genetically Guided Alternating Optimization, *Expert Systems with Applications*, Vol. 41, No. 13, pp. 5960-5971, October, 2014.

[22] B. L. Wang, L. Y. Zhang, L. Zhang, Z. Bing, X. Xu, Missing Data Imputation by Nearest-neighbor Trained BP for Fuzzy Clustering, *Journal of Information & Computational Science*, Vol. 11, No. 15, pp. 5367-5375, October, 2014.

[23] J. K. Dixon, Pattern Recognition with Partly Missing Data, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 10, pp. 617-621, October, 1979.

[24] B. A. Pimentel, R. M. C. R. de Souza, A Multivariate Fuzzy C-means Method, *Applied Soft Computing*, Vol. 13, No. 4, pp. 1529-1607, April, 2013.

## Biographies

**Li Zhang** received his Ph.D. in Materials Processing Engineering from The State Key Laboratory of Rolling and Automation, Northeastern University, China, in 2000. Currently, he works as a professor in School of Information, Liaoning University, China. His research interests include pattern recognition, data cluster analysis and fault diagnosis.

**Hui Pan** is a graduate student in Computer System Structure at Liaoning University, Shenyang, China. Her researched interests is fuzzy clustering for incomplete data.

**Beilei Wang** received the Master degrees in Computer Software and Theory from Liaoning University, Shenyang, China, in 2015. She is currently a Junior Software Developer, Shanghai, China. She is the author of 2 research papers. Her researched interests is fuzzy clustering analysis.

**Liyong Zhang** received the B.S. degree in Automation and the M.S. degree in Control Theory and Control Engineering from Dalian University of Technology, Dalian, China, in 1999 and 2002 respectively. He is currently a Ph.D. student and working as a Lecturer in the School of Control Science and Engineering, Dalian University of Technology. He has published more than 50 research papers. His current research interests include fuzzy clustering and granular computing.

**Zhangjie Fu** obtained his Ph.D. in computer science from the College of Computer, Hunan University, China, in 2012. Currently, he works as an assistant professor in School of Computer and Software, Nanjing University of Information Science and Technology, China. His research interests include cloud computing, data analysis, network and information security.