Fuzzy Clustering Algorithm for Interval Data Based on Feedback RBF Neural Network

Hao Luo[1], Qing Hou[1], Yang Liu[1], Li Zhang[1], Yuanzhi Li[2]

[1] School of Information, Liaoning University, China
[2] Fu Foundation School of Engineering and Applied Science, Columbia University
luohao8711@163.com, yl3833@columbia.edu, 2550803788@qq.com, zhang_li@lnu.edu.cn, 2361676926@qq.com

Abstract

Data set with missing attribute is often encountered in practical applications. To solve the problem that fuzzy c-means clustering algorithm can't be directly used for fuzzy clustering of incomplete data, a feedback Radial Basis Function neural network (FRBF) is proposed to estimate the missing attribute values for incomplete data. The error between the actual output value of RBF neural network and the expected value is fed back to the input layer, then a feedback RBF neural network is constructed. Further, due to the numerical data can't accurately describe the incomplete data, we provide an interval approach, which can convert the numerical data set into an interval valued data set. Thus, an interval fuzzy c-means clustering algorithm based on improved RBF neural network (FRBF-IFCM) is proposed to perform clustering analysis. Experimental results show that this algorithm has better accuracy in data clustering performance than similar algorithms.

Keywords: Incomplete data, Interval value, RBF neural network, Fuzzy C-means

1 Introduction

Clustering analysis is used to analyze the huge and complex data in various fields and is an important tool in data mining. The purpose of clustering analysis is to discover the hidden information of data structure from an unknown data set and divide the data into several disjoint subsets [1]. Data in the same cluster often have similar characteristics, and the similarity between different clusters is relatively small [2]. Fuzzy c-means (FCM) algorithm is an effective clustering method, which uses membership degree to determine the degree to which each sample point belongs to a certain cluster, achieving a fuzzy clustering of data sets [3]. When clustering samples, it can consider the relationship with other clusters comprehensively, so that not only the final clustering results can be obtained, but also which class the extent of each sample belongs to can be decided according to the membership [4]. Therefore, FCM algorithm is widely used in clustering, but FCM needs complete data sets, which can't directly use incomplete data sets [5-6]. However, in practical applications, many data sets suffer from incompleteness, most of the data collection and transmission are carried out in the case of unattended, which is easy to cause the phenomenon of data loss once it is disturbed by the outside world [7]. If the missing data is not processed in time, the validity of large data sets will be seriously affected. How to deal with incomplete data sets is an urgent problem in fuzzy clustering.

In recent years, aiming at the problem of incomplete data clustering, many scholars have proposed some new strategies based on the existing clustering methods [8-10].

The expectation–maximization (EM) has been used to handle incomplete data and probabilistic clustering for a longtime. Abas combined the EM algorithm with finite mixture model to deal with incomplete data. Furthermore, to overcome an issue of the local optimality of the algorithm, he integrated the EM algorithm with Particle Swarm Optimization to enhance the clustering effect [11]. Lin and Su combined EM with Robust Bayes Classifier for feature selection and classification of incomplete data [12]. Ding and Song developed the EM algorithm in the framework of Gaussian copula models for the imputation of missing values [13]. However, when there is a large number of data missing or a large class of attributes missing, the method often fails to get the ideal filling effect and may fall into local optimization.

As an active branch of machine learning, the neural network has been widely applied in imputations of missing values and handle incomplete data clustering. The autoencoder (AE) architecture has obtained great achievements in the field of missing value imputation. Lai et al. [14] propose an architecture named tracking removed autoencoder (TRAE) to model the incomplete data for imputations of missing values by redesigning the input structure of hidden neurons in a dynamic way on the basis of the traditional autoencoder (AE), which strengthens the dependence of missing values on

*Corresponding Author: Li Zhang; E-mail: zhang_li@lnu.edu.cn
DOI: 10.3966/160792642020052103017

known attribute values for each incomplete record. Besides, Zhang et al. [15] propose a fuzzy C-means algorithm based on Missing-data Back Propagation (MBP) neural network. They use MBP neural network to fill the missing data, and FCM algorithm was use to cluster datasets. Zhang and Xu et al. [16] proposed a missing data filling method based on convolutional network by combining spatio-temporal correlation of data with deep learning model. However, considering the efficiency of parameter learning in those neural networks, the construction of the model generally does not exceed three layers in practical application.

As a typical model of machine learning, RBF neural network has been widely applied in function approximation, pattern recognition and many other domains due to its superior predictive accuracy and robustness performance [17]. RBF neural network is a feedforward neural network with good performance. RBF neural network can approach any nonlinear function with arbitrary accuracy, and it has global approximation ability and fast convergence speed [18], which fundamentally solves the local optimal problem of BP network. Therefore, we consider using RBF neural network to model the impution for incomplete data. Further, in order to make RBF network get more information, and improve the accuracy of the algorithm, feedback mechanisms is analyzed in this paper, which feeds back the difference between the predicted value of RBF neural network and the theoretical expected value of data back to the input layer, so as to get the FRBF model. Through feedback, the neural network can retain the data information at the previous moment and add it into the calculation of the data at the next moment, which makes the network not only dynamic but also more complete in the system information retained.

Nevertheless, the missing data filled by the above methods are all numerical data, which can't fully describe the uncertainties present in the missing data.Therefore the interval valued fuzzy clustering theory is introduced and has been used to handle incomplete data recently [19]. Li et al. [20] proposed an interval kernel Fuzzy C-Means clustering of incomplete data. They estimate missing values in the form of intervals using the nearest neighbor method, and use a kernel method to replace Euclidean distance with kernel-induced distance. By clustering interval data based on gradient based on interval kernel distance, interval kernel fuzzy c-means clustering of incomplete data sets is realized. Pham et al. [21] proposed an interval-valued approach to fuzzy co-clustering algorithms for multidimensional data classification. Long Thanh Ngo et al. [22] proposed to a novel interval-valued fuzzy set-based approach to realize collaborative clustering. Jinhua Li et al. [23] proposed a Robust Fuzzy C-means (RFCM) clustering algorithm, which represents missing values by intervals and uses K-nearest neighbor method to construct missing eigenvalues conveniently. Using the interval valued data to express the fuzziness of the boundary between data items is more effective [24-25]. The above literatures show that intervals are applicable for expressing missing values, which are helpful for improving the performance of incomplete data clustering.

Based on the above theories which is used to handle the uncertainty of missing data, a fuzzy clustering algorithm for interval data is proposed based on feedback RBF neural network (FRBF-IFCM). FRBF neural network is used to estimate the missing attributes, and then the numeric data is converted into interval data, forming interval FCM clustering algorithm. 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. A brief review of theoretical review is given in Section 2. In Section 3, we present the interval fuzzy c-means clustering algorithm based on feedback RBF neural network (FRBF-IFCM) and its application analysis. In Section 4, we first describe data sets and the model parameters, and then the detailed comparative experiment is done to illustrate the performance of FRBF-IFCM by comparing with other methods. Finally, this paper concludes with a summary in Section 5.

2 Theoretical Review

This section mainly summarizes the basic network model structure of FRBF neural network and its working principle. Meanwhile, it introduces the basic idea and operation process of IFCM algorithm.

2.1 FRBF Neural Network

In the training stage, most schemes only use the set of complete data to train the network. The decrease in the number of complete data will cause a serious decline in training accuracy as the missing rate increases. Moreover, the information of the incomplete data is as important as complete data when training the model for imputation of missing data. For the basic RBF neural network, the error between the prediction output and the expected output of the output layer of the missing attribute can't be calculated. Therefore, the weights of the neural network can't be adjusted. But the training dataset of FRBF neural network contains incomplete data, which can replace the error of missing attributes with the mean error of complete attributes in the training process. Further, RBF network introduces feedback, which fed back the error between the actual value of incomplete data and expected value to the input layer, so that the network can remember more information in the past. Therefore, the whole FRBF

model not only refers to the incomplete data, but also introduces a feedback mechanism to make the whole network obtain more information. This can make the filling value of the missing attribute more reasonable, thus improving the effectiveness of clustering analysis. After several iterations of learning, the FRBF 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. Because the data samples in the nearest neighborhood have similar structures, using the nearest neighbor sample to guide the estimation of incomplete data attributes can make the estimated value closer to its actual value. Therefore, we use the nearest neighbor rule [26-27] to select the corresponding training sample set for missing data attributes in incomplete data samples. Thus the missing attributes can be more effectively estimated and filled.

For an incomplete data set of s dimension $X=\{x_1, x_2, \ldots, x_n\}$, the similarity measurement formula of incomplete data sample x_a and data sample x_b (with or without missing attributes) is shown in Eq. (1) :

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

Where, x_{ia} and x_{ib} is the ith attribute of x_a and x_b respectively, and I_i satisfies:

$$I_i=\begin{cases}1, & x_{ia} \text{ and are complete attribute}\\ 0.5, & x_{ia} \text{ or } x_{ib} \text{ is the missing attribute}\\ 0, & \text{other}\end{cases} \qquad (2)$$

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.

2.1.2 Sample Training with FRBF Neural Network

When obtaining data samples, data attributes are often lost for various reasons, resulting in incomplete data. FRBF is used to estimate the missing attributes in the incomplete data set. The process is as follows:

Step 1: Normalize input data set. All the data are converted to a number of interval [0,1].

Step 2: Determinate and optimize training sample.

Step 3: Initialize FRBF network. Determine the input nodes $n+m$, hidden nodes l and output nodes m; initialize the weight w_{ik}, center vector C_i, and width σ^2. Determine the maximum training times M, error accuracy ε_1, and each learning rate of the network η_1, η_2, η_3.

Step 4: Calculate the output value of hidden layer of the network according to Eq. (3). Where, C_i is the center vector and σ^2 is the width and l is the node of hidden layer. Assuming that input parameters of FRBF neural network is $X=(x_1, x_2, \ldots, x_{n+m})$, the output parameters is $Y=(y_1, y_2, \ldots, y_m)$ and the error between expected output value and actual output value is e_k.The output value of hidden layer calculated by Eq. (1):

$$\phi_i(X)=\exp\left(-\frac{\left\|X-C_i\right\|^2}{\sigma^2}\right), i=1,2,\ldots,l \qquad (3)$$

Step 5: Calculate the output value of the network output layer according to Eq. (4). Where, w_{ik} is the connection weight between the hidden layer and the output layer.

$$O_{k=}\sum_{i=1}^{l} w_{ik}\phi_i(X), k=1,2,\ldots,m \qquad (4)$$

Step 6: Calculate the error e_k according to Eq. (5). Then the error e_k is fed back to the input layer. Where $\overline{e}$ is the average error of complete attribute.

$$e_k=\begin{cases}\overline{e}, & \text{if } Y_k \text{ is a missing attribute}\\ Y_k-O_k, & \text{otherwise}\end{cases}, k=1,2,\ldots,m \qquad (5)$$

Step 7: Adjust and update the weight w_{ik}, center vector C_i, and width σ^2 of the network using Eq. (6), Eq. (7) and Eq. (8). The resulting error is fed back to the input layer.Where $w^n{}_{ik}$ represent the weight when the iteration times reach n, C_i^n and σ^n is also like this.

$$w_{ik}^{n+1}=w_{ik}^{n}-\eta_1\frac{\partial E^n}{\partial w_{ik}^n} \qquad (6)$$

$$C_i^{n+1}=C_i^n-\eta_2\frac{\partial E^n}{\partial C_i^n} \qquad (7)$$

$$\sigma_i^{n+1}=\sigma_i^n-\eta_3\frac{\partial E^n}{\partial \sigma_i^n} \qquad (8)$$

Eq. (9), Eq. (10) and Eq. (11) are the adjustment direction of above parameters, where $\phi_i^{'}(X)$ is derivative of $\phi_i(X)$.

$$\frac{\partial E}{\partial w_{ik}}=\sum_{k=1}^{m}\sum_{i=1}^{l} w_{ik}\phi_i(X) \qquad (9)$$

$$\frac{\partial E}{\partial C_i}=\frac{1}{\sigma^2}\sum_{k=1}^{m}\sum_{i=1}^{l}w_{ik}\phi_i^{'}(X)\left\|X-C_i\right\| \tag{10}$$

$$\frac{\partial E}{\partial \sigma_i}=\frac{1}{\sigma^3}\sum_{k=1}^{m}\sum_{i=1}^{l}w_{ik}\phi_i^{'}(X)(\left\|X-C_i\right\|)^2 \tag{11}$$

The objective function of the FRBF is:

$$E=\frac{1}{2}\sum_{k=1}^{m}e_k{}^2 \tag{12}$$

Step 8: Terminate the iterations if e < ε_1, or iteration number is greater than the maximum number of training; otherwise, repeat steps 4 to 8.

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

In order to solve the problem of uncertainty in missing data, interval data is introduced and FCM algorithm is used to cluster interval valued data, then obtained IFCM algorithm. FCM and IFCM algorithms are described below.

2.2.1 FCM Algorithm

The FCM algorithm uses the idea of iteration, and then optimizes the objective function. Fuzzy membership degree is introduced in FCM algorithm to express the degree of each data sample belonging to each cluster. In the initial stage of FCM clustering algorithm, a number of clustering centers are randomly selected, and the initial clustering results are obtained by taking these clustering centers as the starting point. Then, the ultimate membership value of each node is obtained by optimizing the objective function, and the cluster nodes are classified into the corresponding region by comparing their membership value [28].

The elements in the membership matrix satisfy the following conditions:

$$\sum_{i=1}^{c}u_{ij}=1,\ j=1,2,\ldots,n \tag{13}$$

The objective function of the FCM is:

$$J_m\left(U,V\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij}\right)^m d^2\left(x_j,v_i\right) \tag{14}$$

The updating formulas of the membership degree u_{ij} and the cluster center are as follows:

$$v_i=\frac{\sum_{j=1}^{n}\left(u_{ij}\right)^m x_j}{\sum_{j=1}^{n}\left(u_{ij}\right)^m},\ i=1,2,\ldots,c \tag{15}$$

$$u_{ij}=\left[\sum_{t=1}^{c}\left(\frac{\left\|x_j-v_i\right\|_2^2}{\left\|x_j-v_t\right\|_2^2}\right)^{\frac{1}{m-1}}\right]^{-1},\quad i=1,2,\ldots,c;\ j=1,2,\ldots,c \tag{16}$$

2.2.2 IFCM Algorithm

In order to handle the uncertainty of missing data, we convert the numerical data into the interval valued data. The IFCM algorithm is used to clustering analysis of interval valued data sets in this paper.

Let $X=\{x_1,x_2,\ldots,x_n\}$ be the interval valued data sets, n be the number of data sets. Each attribute value in data sample x_k is represented by interval, which is $x_{kj}=[x_{kj}^{-},x_{kj}^{+}],(1\le k\le s)$. The data set X is divided into c categories, where the clustering center is represented as $V=[v_{ik}]=[v_1,v_2,\ldots,v_c]$ and $v_{ik}=[v_{ik}^{-},v_{ik}^{+}]$, $(i=1,2,...,s)$. We use the membership matrix $U_{(c\times n)}$ to represent the clustering result of data sets.

The objective function of IFCM algorithm is as follows:

$$J_m\left(U,V\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij}\right)^m d^2\left(x_j,v_i\right) \tag{17}$$

Where, u_{ij} satisfies the constraints of formula $\sum_{i=1}^{q}u_{ij}=1,j=1,2,...,n$ and $0<\sum_{j=1}^{n}u_{ij}<1,i=1,2,...,c.$ The Euclidean distance between data x_j and cluster center v_i is: $d^2(x_j,v_i)$. Where, m is the fuzzy index and be set to 2 satisfying and $m\in(1,+\infty)$.

The specific calculation equation of $d^2(x_j,v_i).$ is as follows:

$$\begin{aligned}&\left\|x_j-v_i\right\|_2^2\\&=\sqrt{\left(x_j^{-}-v_i^{-}\right)^T\left(x_j^{-}-v_i^{-}\right)+\left(x_j^{+}-v_i^{+}\right)^T\left(x_j^{+}-v_i^{+}\right)}\end{aligned} \tag{18}$$

$x_j^{-}=[x_{1j}^{-},x_{2j}^{-},...v_{sj}^{-}]^T$ and $x_j^{+}=[x_{1j}^{+},x_{2j}^{+},...v_{sj}^{+}]^T$ are the left and right boundary of interval data attribute x_j. The vectors of left and right boundary of interval clustering center v_i are expressed as $v_i^{-}=[v_{1i}^{-},v_{2i}^{-},...v_{si}^{-}]^T$ and $v_i^{+}=[v_{1i}^{+},v_{2i}^{+},...v_{si}^{+}]^T$, respectively.

The updating formulas of interval for cluster center are as follows:

$$v_i^+ = \frac{\sum_{j=1}^{b} u_{ij}^m x_j^+}{\sum_{j=1}^{n} u_{ij}^m}, i = 1, 2, \ldots, c \quad (19)$$

$$v_i^- = \frac{\sum_{j=1}^{n} u_{ij}^m x_j^-}{\sum_{j=1}^{n} u_{ij}^m}, i = 1, 2, \ldots, c \quad (20)$$

If the interval data sample x_j belongs to the interval value range of cluster center $v_h (1 \le h \le c)$, its membership degree is 1. If the interval data sample x_j does not belong to the interval value range of cluster center v_h, its membership degree is 0, that is:

$$u_{ij} = \begin{cases} 0, i \ne h \\ 1, i = h \end{cases} \quad (21)$$

Otherwise, the membership update formula of each data sample is follows Eq. (16).

2.2.3 Algorithm Process for IFCM

The main process of the IFCM algorithm is as follows:

Step 1: Parameter initialization. Clustering number is set to c and the maximum iteration number is G; determine fuzzy index m and iterative termination threshold ε, and the membership matrix $U^{(0)}$ is initialized.

Step 2: When the iteration reached to $l (l = 1, 2, ...)$, update equation of clustering center according to $U^{(l-1)}$, and calculate the left interval value $V^{(l)^-}$ and the right interval value $V^{(l)^+}$ of clustering center $V^{(l)}$ using (19) and (20).

Step 3: Update the membership matrix using $V^{(l)}$ according to (16) or (21).

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

3 FRBF-IFCM Algorithm

By training the FRBF network proposed in this paper, the estimated values of incomplete data are numerical. However, numerical data can't accurately describe the uncertainty of incomplete data, and there will be some errors. Therefore, this section mainly introduces the conversion between the numerical value and interval value. Then, how to perform clustering analysis with FRBF-IFCM Algorithm is introduced in this section.

3.1 Conversion between Missing Data Sets and Interval Data Sets

In this paper, numerical data are converted into interval data to express the uncertainty of missing data. The transformation rules and the transformation process are described below.

3.1.1 The Interval Estimation of Missing Attributes

The value of interval data is not a fixed value, but a continuous interval range value, which is also the difference between interval data and numerical data. Although the value estimated by the FRBF is close to its original value, it can't accurately describe the incomplete data. Moreover, the use of interval data to describe the uncertainty of data samples has been widely recognized [29]. The rules of conversion for interval valued data sets are as follows:

(1) After estimating the missing attribute x_{ik} in incomplete data sets by FRBF network, the estimation value of missing attributes can be obtained by the output value of FRBF output neurons, so does the estimation of complete data. Therefore, for the complete data, the error between expected output value and actual output value is obtained. Compute the mean error and it is used to decided the interval of missing attribute. If the mean value is $\overline{e}$, the estimation of missing attribute is x, then the interval of missing attribute is set to $[x - \overline{e}, x + \overline{e}]$.

(2) The complete attribute value x_{ik} in the incomplete data set is expressed in the form of interval $[x_{ik}^-, x_{ik}^+]$, and satisfies the requirement of $x_{ij}^- = x_{ij} = x_{ij}^+$. In other words, the values of the left endpoint and right endpoint of the complete attribute interval are equal and equal to the original value of the attribute.

3.1.2 Conversion of Interval Valued Data Sets

The interval conversion process of the entire data set is as follows:

Step 1: The missing attribute is estimated through FRBF network and the estimation error of complete attributes in incomplete data can also be obtained.

Step 2: The average value $\overline{e}$ of the error is obtained by calculating the estimated error of complete attribute in the incomplete data sample.

Step 3: The interval values of missing data attributes in incomplete data sets was expressed as $[x - \overline{e}, x + \overline{e}]$, where the median value x is the estimation of the missing attribute, $\overline{e}$ is the mean value of the errors.

Step 4: The interval range of the missing attribute was checked and it should be limited to the range of the interval [0, 1]. If $x^- < 0$, the value of the left endpoint of the estimated interval of the missing attribute is set

to 0, namely $x^- = 0$. If $x^+ > 1$, the right endpoint of the estimated interval of the missing attribute is set to 1, namely $x^+ = 1$.

Step 5: The complete attributes in the incomplete data set was converted to interval.

3.2 Proposed FRBF-IFCM Algorithm

The processing of the incomplete data sets using FRBF-IFCM algorithm as follows: First, the estimation of the missing attribute was obtained by FRBF-IFCM algorithm. Then, the numerical value is converted into the form of interval. Besides, the complete attribute is also transformed into an interval. Finally, FCM algorithm is used to cluster interval data sets. The flow chart of FRBF-IFCM is shown in Figure 1 and the clustering analysis of steps are as follows:

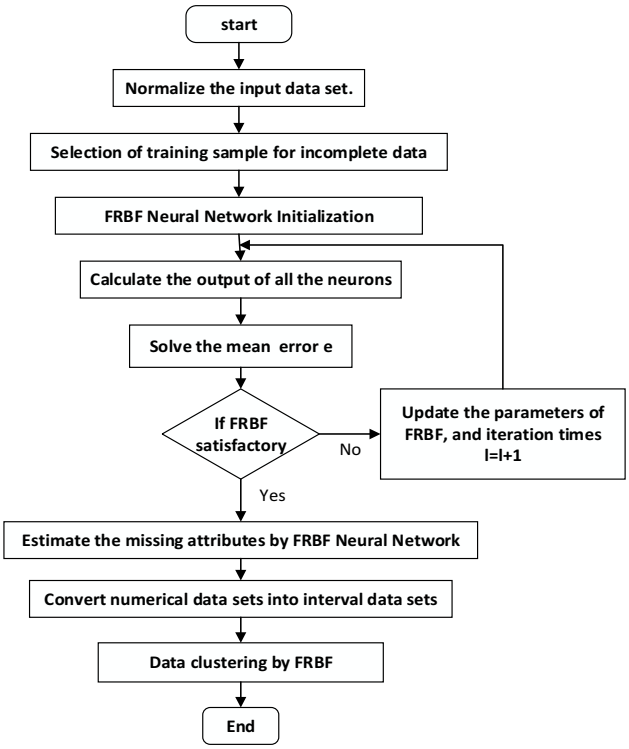

Figure 1. The flow chart of FRBF-IFCM algorithm

Step 1: Normalize the input data set. All data are converted into numbers between intervals [0, 1].

Step 2: Select training sample set. The training samples are selected for incomplete data samples according to the nearest neighbor rule.

Step 3: Initialize FRBF network. Define the number of nodes in each layer of the neural network as $n+m$, l and m respectively. Initialize the weight w, center vector C_i and width σ^2. Determine the maximum number of training M. Determine the error accuracy ε_1 and each learning rate η_1, η_2, η_3.

Step 4: Train FRBF network. The FRBF network is trained using the selected training sample, then otaine the well trained FRBF network for each missing attribute.

Step 5: Fill the missing attributes with estimations. The trained FRBF network is used to estimate the missing data attributes in incomplete data sets, and the estimation error of the complete attributes in data sets is obtained.

Step 6: The evaluation interval of missing attributes was determined. According to the mean error of the estimated error of complete attribute in the data set by the FRBF network, the left endpoint and right endpoint of the missing attribute intervals is determined. Then the interval of the missing attribute x is $[x-\overline{e}, x+\overline{e}]$, that is $[x^-, x^+]$. And $[x^-, x^+] \subset [0,1]$ also needs to make sure. Thus, the estimated values of all missing attributes in the data set are converted into intervals.

Step 7: The interval transformation of the complete attribute value is carried out. All the complete attributes in the data set are converted into the form of an interval, that is, the left and right endpoint values of interval are equal and equal to their actual values. So that the entire numerical data set is interval.

Step 8: Initialize the parameters of IFCM algorithm. The number of clusters is c. The maximum number of iterations is G. Fuzzy index is m and iteration termination threshold is ε. The membership matrix $U^{(0)}$ is initialized.

Step 9: Update the cluster center matrix. When the iteration reaches the $l(l=1,2,\ldots)$, according to $U^{(l-1)}$, the left endpoint value $V^{(l)^-}$ and right endpoint value $V^{(l)^+}$ of the clustering center matrix $V^{(l)}$ are updated by using Eq. (19) and Eq. (20).

Step 10: Update the membership matrix. According to $V^{(l)}$, the membership matrix $U^{(l)}$ is updated by Eq. (16) and Eq. (21).

Step 11: Judge the terminated condition. Terminate the iterations when the training times reach the maximum G, or $\max|U^{(l+1)} - U^{(l)}| \le \varepsilon$; otherwise, increase the iteration($l = l+1$) and repeat steps 9 to 11.

4 Experimental Results and Discussion

4.1 Experimental Results

Three data sets in UCI database, Iris, Bupa and Breast, are selected, along with two artificial data sets, to perform the simulation experiments. Table 1 below describes the information related to the above data sets.

Table 1. The information of data sets

The data sets	Number of simples	Number of attributes	Number of classes
Iris	150	4	3
Bupa	345	7	2
Breast	683	11	2
Artificial data sets I	200	N/A	2
Artificial data sets II	400	N/A	3

UCI data sets. Iris data set is also often used in cluster analysis experiments. It is first a data set for multidimensional attribute analysis of Iris flowers. There are 150 sample data in this data set and they are divided into three categories. Each of the three categories contains 50 samples and each sample contains 4 attributes.

The Bupa dataset is a sample data set for liver disease studies. The data set contains 345 sample data. It is divided into two categories, in which the sample number of each category is 145 and 200 respectively. This data sample contains 7 attributes, but the 7th attribute is the category identification, and does not participate in the experiment.

The Breast dataset is the dataset for clinical case descriptions of Breast cancer. There are 699 sample data in this data set, but there are 16 sample data with attribute missing, so 683 data are used in the actual data analysis. The data set is divided into two categories: benign breast tumor sample data and malignant breast tumor sample data. It contains 444 and 239 data samples, respectively. There are 11 attribute columns in the sample data.

Artificial data sets. Two artificial data sets are generated using the data generation method given by B. A. Pimentel [30]. The number of data samples in manual data set I is 200, including 2 categories, and each subclass contains 100 data samples. In artificial data set II, the number of data samples is 400 and the number of categories is 3. Each subclass contains 80, 100 and 220 data samples respectively. The data samples in the above two artificial data all obey independent two-dimensional normal distribution. The expectation and variance matrices are defined as follows:

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad \sum = \begin{bmatrix} \sigma_1^{\,2} & 0 \\ 0 & \sigma_2^{\,2} \end{bmatrix}$$

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

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

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

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

(1) The first category: $u_1 = 20,\ \ u_2 = 20\ \ \sigma_1^{\,2} = 2,\ \sigma_2^{\,2} = 4$.

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

(3) The third category: $u_1 = 36,\ u_2 = 36\ \ \sigma_1^{\,2} = 16,\ \sigma_2^{\,2} = 16$.

Evaluation Index of Algorithms. The proposed FRBF-IFCM algorithm was evaluated from two aspects: average number of misclassification and several external evaluation indexes. Rand Index [31]. Adjusted Rand Index; Jaccard Coefficient; Minkowski Measure and Γ Statistics was selected in this paper, which can evaluate the similarity between the real division of experimental data and the corresponding fuzzy division results. Among them,only the smaller the value of the evaluation index Minkowski Measure is, the better the performance of the corresponding clustering algorithm is, others are opposite. The formulas of the above external evaluation indexes are shown in Table 2, where $a = |V \cap Y|$, $v = |V \cap Z|$, $c = |X \cap Y|$, $d = |X \cap Z|$. Among them, the matrix R and Q respectively represent the real division of experimental data and the fuzzy division result of clustering algorithm. V represents the fuzzy set of data sample pairs belonging to the same category in R; X represents the fuzzy set of data sample pairs belonging to different classes in R; Y represents the fuzzy set of data sample pairs belonging to the same category in Q; Z represents the fuzzy set of data sample pairs belonging to different classes in Q.

4.2 Discussion

The experimental results of the method in this paper (FRBF-IFCM) are compared with the four classical algorithms proposed by Hathaway and Bezdek [9]: WDS-FCM, PDS-FCM, OCS-FCM, NPS-FCM, and the recent algorithm MBP-FCM [15] presented by Zhang, to verify the effectiveness of the method in the paper. Missing attributes are randomly generated by humans. The relevant parameters in the experiments are set as follows: the maximum training times of FRBF network are set to M=500, the error precision is set to $\varepsilon_1 = 0.01$, each learning rate is set to $\eta_1 = 0.1$, $\eta_2 = 0.1$, $\eta_3 = 0.1$, the maximum iteration times of IFCM algorithm is set to G = 100, the fuzzy index is set to m=2, and the iteration termination threshold is set to $\varepsilon = 0.001$. The missing rate is taken as 5%, 10%, 15% and 20%. The experimental results are shown from Table 3 to Table 8. The optimal results are marked by bold types, and the suboptimal results are marked by underline.

We can see from Table 3 to Table 6 that, in general, FRBF-IFCM algorithm is better than FRBF-FCM algorithm and MBP-FCM algorithm in terms of the average number of misclassification results. FRBF-FCM algorithm adopts numerical value estimation, which does not make full use of data set information and causes loss of information and degrade the clustering performance. Compared with the method, FRBF-IFCM utilizes the information of interval value estimation, which improved the accuracy of clustering results. The network structure of FRBF-IFCM is

Table 2. The formulas of the external evaluation indexes

External effectiveness evaluation index	The formula
Rand Index (RI)	$w_{RI}=\frac{a+d}{a+b+c+d}$
Adjusted Rand Index (ARI)	$w_{ARI}=\frac{a-\frac{(a+b)(a+c)}{a+b+c+d}}{\frac{(a+b)+(a+c)}{2}-\frac{(a+b)(a+c)}{a+b+c+d}}$
Jaccard coefficient (JC)	$w_{JC}=\frac{a}{a+b+c}$
Minkowski measure (MM)	$w_{MM}=\sqrt{\frac{b+c}{b+a}}$
Γ statistics (Γ S)	$w_{\Gamma S}=\frac{Ma-(a+b)(a+c)}{\sqrt{(a+b)(a+c)(M-(a+b))(M-(a+c))}}$, $M=a+b+c+d$

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

% miss	The average number of misclassification						
	WDS-FCM	PDS-FCM	OCS-FCM	NPS-FCM	MBP-FCM	FRBF-FCM	FRBF-IFCM
0	16	16	16	16	16	16	16
5	16.4	17.1	16.8	16.7	16.3	16.0	**15.6**
10	16.5	16.9	17.1	16.8	16.2	16.2	**15.9**
15	16.1	17.4	17.3	17.0	16.7	**15.5**	15.7
20	16.3	17.7	17.1	17.6	16.9	15.7	**15.3**

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

% miss	The average number of misclassification						
	WDS-FCM	PDS-FCM	OCS-FCM	NPS-FCM	MBP-FCM	FRBF-FCM	FRBF-IFCM
0	177	177	177	177	177	177	177
5	177.4	177.2	177.5	177.2	176.5	175.5	**174.6**
10	176.4	177.0	176.6	177.0	176.9	**173.7**	174.0
15	177.8	178.5	177.5	178.4	178.0	177.7	**176.5**
20	178.3	179.1	177.4	179.0	177.3	175.3	**175.1**

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

% miss	The average number of misclassification						
	WDS-FCM	PDS-FCM	OCS-FCM	NPS-FCM	MBP-FCM	FRBF-FCM	FRBF-IFCM
0	30	30	30	30	30	30	30
5	30.3	30.9	31.8	31.9	29.5	29.5	**29.3**
10	30.2	31.0	31.7	31.8	29.2	29.4	**29.1**
15	32.2	33.6	33.3	32.9	30.5	31.6	**30.4**
20	33.2	34.1	36.2	34.7	33.3	33.1	**31.0**

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

% miss	The average number of misclassification			
	artificial data set I		artificial data set II	
	FRBF-FCM	FRBF-IFCM	FRBF-FCM	FRBF-IFCM
5	8.8	**8.5**	56.2	**56.0**
10	**10.8**	11.2	58.4	**58.0**
15	13.1	**12.8**	63.8	**61.7**
20	14.3	**13.5**	65.2	**63.5**

Table 7. Average effectiveness evaluation results of 10 trails using incomplete Iris data set

Algorithm	5% missing					10% missing				
	RI	ARI	JC	MM	Γ S	RI	ARI	JC	MM	Γ S
WDS-FCM	0.7764	0.4802	0.6425	0.6850	0.4848	0.7799	0.4856	0.6529	0.6727	0.4903
PDS-FCM	0.7801	0.5546	0.6303	0.6991	0.5608	0.7713	0.5058	0.6305	0.7005	0.5119
OCS-FCM	0.7625	0.4556	0.6358	0.6934	0.4618	0.7927	0.5412	0.6334	0.6956	0.5458
NPS-FCM	0.7634	0.4576	0.6370	0.6920	0.4628	0.7928	0.5419	0.6327	0.6962	0.5466
MBP-FCM	0.7991	0.5989	0.7005	0.5755	0.5607	0.7928	0.5418	0.6526	0.6730	0.5547
FRBF-IFCM	**0.8532**	**0.6424**	**0.7305**	**0.4856**	**0.6208**	**0.8501**	**0.6501**	**0.7289**	**0.4907**	**0.6098**
Algorithm	15% missing					20% missing				
	RI	ARI	JC	MM	Γ S	RI	ARI	JC	MM	Γ S
WDS-FCM	0.8237	0.5981	0.6683	0.6548	0.6025	0.8158	0.5849	0.6674	0.6558	0.5901
PDS-FCM	0.8093	0.6054	0.6247	0.7073	**0.6106**	0.7347	0.3528	0.6219	0.7100	0.3560
OCS-FCM	0.7642	0.4613	0.6331	0.6976	0.5478	0.7495	0.4022	0.6316	0.6991	0.4064
NPS-FCM	0.7086	0.4677	0.6396	0.6891	0.4727	0.7522	0.4105	0.6369	0.6925	0.4147
MBP-FCM	0.8235	0.6013	0.7010	0.6550	0.5901	0.8179	0.5813	0.6637	0.6557	0.5901
FRBF-IFCM	**0.8475**	**0.6412**	**0.7276**	**0.4983**	0.6030	**0.8320**	**0.6356**	**0.7158**	**0.5010**	**0.5910**

Table 8. Average effectiveness evaluation results of 10 trails using incomplete Bupa data set

Algorithm	5% missing					10% missing				
	RI	ARI	JC	MM	Γ S	RI	ARI	JC	MM	Γ S
WDS-FCM	0.4997	-0.0021	0.3537	0.9901	-0.0021	0.4997	-0.0019	0.3510	**0.9903**	-0.0019
PDS-FCM	0.4997	-0.0023	**0.3551**	0.9904	-0.0023	0.4996	-0.0023	**0.3555**	0.9904	-0.0023
OCS-FCM	0.4997	-0.0023	**0.3551**	0.9904	-0.0022	0.4996	-0.0023	0.3552	0.9905	-0.0023
NPS-FCM	0.4996	-0.0022	0.3545	0.9904	-0.0022	0.4997	-0.0021	0.3544	**0.9903**	-0.0021
MBP-FCM	**0.4998**	-0.0021	0.3548	0.9902	-0.0021	0.4998	-0.0021	0.3545	0.9904	-0.0021
FRBF-IFCM	**0.4998**	**-0.0016**	0.3476	**0.9900**	**-0.0016**	**0.4999**	**-0.0016**	0.3550	**0.9903**	**-0.0016**
Algorithm	15% missing					20% missing				
	RI	ARI	JC	MM	Γ S	RI	ARI	JC	MM	Γ S
WDS-FCM	0.4996	-0.0021	0.3523	0.9902	-0.0021	0.4995	-0.0024	0.3521	0.9898	-0.0024
PDS-FCM	0.4997	-0.0023	0.3551	0.9904	-0.0023	0.4997	-0.0023	0.3558	0.9904	-0.0023
OCS-FCM	0.4997	-0.0023	0.3552	0.9904	-0.0023	0.4997	-0.0022	0.3552	0.9904	-0.0022
NPS-FCM	0.4997	-0.0021	0.3543	0.9904	-0.0021	0.4997	**-0.0019**	0.3533	0.9903	**-0.0019**
MBP-FCM	0.4996	-0.0019	0.3515	0.9899	-0.0019	0.4997	-0.0021	0.3548	0.9900	-0.0021
FRBF-IFCM	**0.4999**	**-0.0018**	**0.3644**	**0.9893**	**-0.0018**	**0.5003**	-0.0021	**0.3649**	**0.9888**	-0.0021

Table 9. Average effectiveness evaluation results of 10 trails using incomplete Breast data set

Algorithm	5% missing					10% missing				
	RI	ARI	JC	MM	Γ S	RI	ARI	JC	MM	Γ S
WDS-FCM	0.7694	0.5329	0.6592	0.6494	0.5346	0.7286	0.4549	0.6115	0.7013	0.4560
PDS-FCM	0.8136	0.6216	0.7173	**0.5852**	0.6245	0.7819	0.5588	0.6789	0.6242	0.5609
OCS-FCM	0.8136	0.6217	**0.7174**	0.5865	0.6251	0.8143	0.5669	**0.7182**	0.5861	0.6265
NPS-FCM	0.8093	0.6139	0.7120	0.5931	0.6170	0.8066	0.6076	0.7076	0.5975	0.6103
MBP-FCM	0.8125	0.6218	0.7163	0.5891	0.6217	0.8114	0.6151	0.7147	0.5907	0.6207
FRBF-IFCM	**0.8156**	**0.6220**	0.7173	0.5854	**0.6332**	**0.8145**	**0.6189**	0.7156	**0.5832**	**0.6270**
Algorithm	15% missing					20% missing				
	RI	ARI	JC	MM	Γ S	RI	ARI	JC	MM	Γ S
WDS-FCM	0.7717	0.5383	0.6649	0.6401	0.5401	0.7637	0.5226	0.6566	0.6478	0.5242
PDS-FCM	0.8134	0.6213	0.7171	0.5875	0.6246	0.8131	0.6207	0.7167	0.5879	0.6241
OCS-FCM	0.8142	**0.6229**	**0.7181**	0.5862	**0.6263**	**0.8148**	0.6240	0.7187	0.5853	**0.6271**
NPS-FCM	0.8018	0.5974	0.7010	0.6046	0.6002	0.7985	0.5906	0.6965	0.6094	0.5933
MBP-FCM	0.8118	0.6179	0.7171	0.5902	0.6215	0.8111	0.6164	0.7187	0.5914	0.6200
FRBF-IFCM	**0.8143**	0.6228	**0.7181**	**0.5769**	0.6250	0.8132	**0.6242**	**0.7188**	**0.5693**	0.6247

simple, compared with the complex network structure, the three-layer network can well estimate the missing attributes.

The results presented in Table 3 to Table 5 show misclassification results of WDS-FCM, PDS-FCM, OCS-FCM, NPS-FCM, MBP-FCM, FRBF-FCM and FRBF-IFCM. Different methods for handling missing attributes in FCM lead to different clustering results. In general, the average number of FRBF-IFCM misclassification is the least in the case of different

attributes missing rates. Only when the missing rate of Iris data set is 15% and that of Bupa data set is 10%, the experimental results of FRBF-FCM algorithm are better than the results of FRBF-IFCM algorithm.

From the results in Table 7 to Table 9, it can be seen that, for the other external effectiveness evaluation indexes, the FRBF-IFCM algorithm proposed in this paper is better than the others, can obtain relatively better experimental results compared with the other algorithms.

The convergence analysis is shown from Figure 2 to Figure 6, which describe the change curve between objective function and iterations of three data sets with different missing rates by FRBF-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.

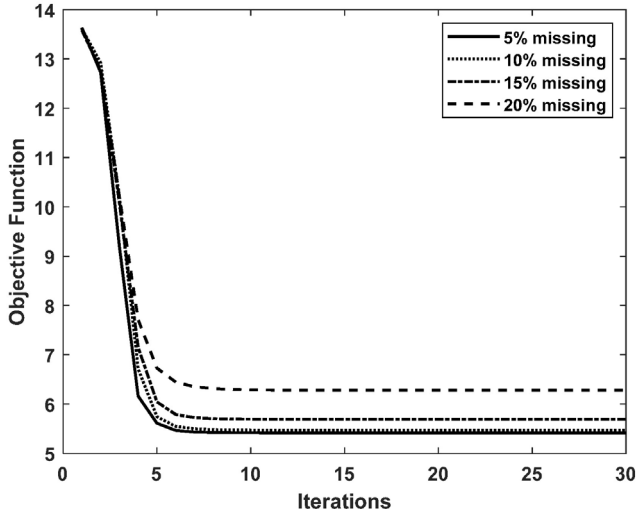

Figure 2. The change curve between objective function and iterations of Iris data set

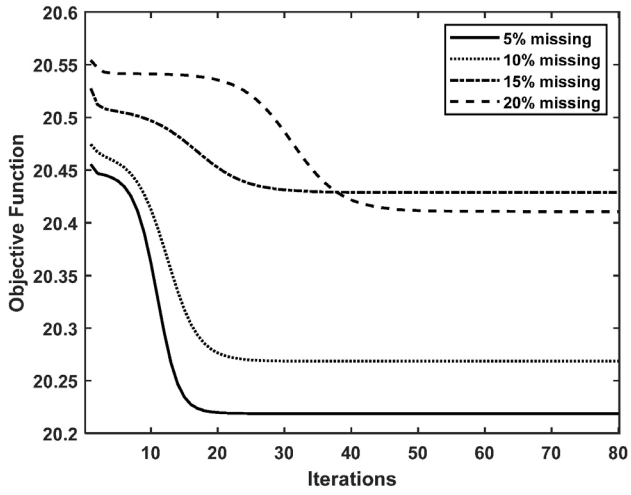

Figure 3. The change curve between objective function and iterations of Bupa data set

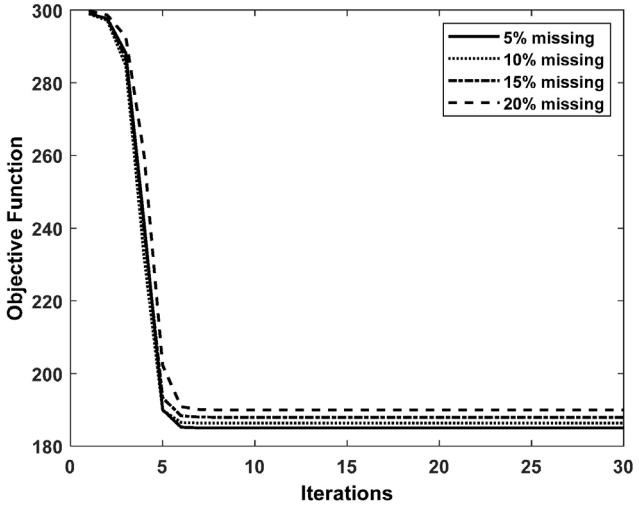

Figure 4. The change curve between objective function and iterations of Breast data set

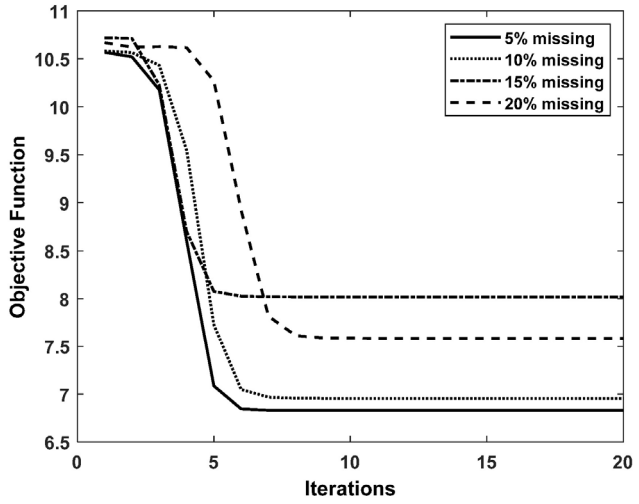

Figure 5. The change curve between objective function and iterations of artificial data set I

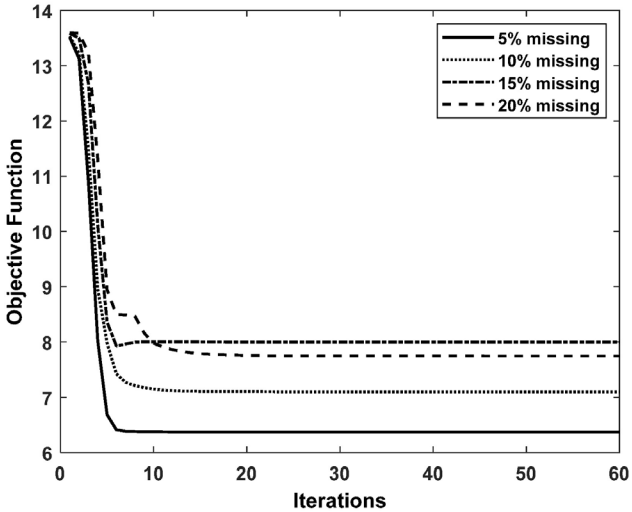

Figure 6. The change curve between objective function and iterations of artificial data set II

5 Conclusion

This paper presents a fuzzy c-means algorithm based on FRBF for clustering analysis. The contribution of this paper is to propose a method to estimate missing attributes for incomplete data and convert numerical data to interval valued data for clustering analysis. FRBF algorithm is used to estimate the missing attribute of incomplete data sets, where the error between the predicted value of FRBF neural network and the actual value of data is fed back to the input layer makes the estimation of missing attribute more reasonable. Furthermore, the missing attribute is replaced by intervals, which more accurately describes the information of missing attribute. Therefore, the proposed algorithm can obtain the more reasonable estimation and better clustering results. The experimental results show that the proposed algorithm has more advantages over other methods in accuracy and more effective when it is applied to the incomplete data classification. However, the intervals may have influence on the structure of data sets. In the future, the problem of interval determination remains a challenge.

Acknowledgements

This work is supported by the National Nature Science Foundation of China under Grant No. 51704138. It is supported by the Educational Department of Liaoning Province Science and Technology Research Projects under Grant LQN201910.

References

[1] Y. Yang, Z. Ma, Y. Yang, F. Nie, H. T. Shen, Multitask Spectral Clustering by Exploring Intertask Correlation, *IEEE Transactions on Cybernetics*, Vol. 45, No. 5, pp. 1083-1094, May, 2015.

[2] Q. D. Liu, R. S. Zhang, R. J. Hu, G. J. Wang, Z. H. Wang, Z. L. Zhao, An Improved Path-based Clustering Algorithm, *Knowledge-Based Systems*, Vol. 163, pp. 69-81, January, 2019.

[3] S. B. Zhou, W. X. Xu, L. K. Xu, Improved FCM Algorithm Based on Density Peaks and Spatial Neighborhood Information, *Chinese Journal of Scientific Instrument*, Vol 40, No. 4, pp. 137-144, April, 2019.

[4] J. X. Lin, L. P. Wu, J. W. Wu, Z. J. Zhang, Adaptive FCM Clustering Algorithm Based on Sample and Feature Weights, *Journal of Natural Science of Heilongjiang University*, Vol. 35, No. 2, pp. 244-252, April, 2018.

[5] X. Li, Z. J. Fan, Y. Yao, Q. Xu, W. Y. Zhu, Improved Automated Graph and FCM Based DDoS Attack Detection Mechanism in Software Defined Networks, *Journal of Internet Technology*, Vol. 20, No. 7, pp. 2117-2127, December, 2019.

[6] 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.

[7] W. Wang, Q. Su, W. Zhou, Y. Liu, B. Zhang, Missing Data Filling Algorithm in Different Types of Incomplete Large Data, *Science Technology and Engineering*, Vol. 18, No. 8, pp. 91-96, March, 2018.

[8] J. Nayak, B. Naik, D. P. Kanungo, H. S. Behera, A Hybrid Elicit Teaching Learning Based Optimization with Fuzzy C-Means (ETLBO-FCM) Algorithm for Data Clustering, *Ain Shams Engineering Journal*, Vol. 9, No. 3, pp. 379-393, September, 2018.

[9] G. Q. Zhao, L. Zhang, C. Tang, W. F. Hao, Y. Luo, Clustering of AE Signals Collected during Torsional Tests of 3D Braiding Composite Shafts Using PCA and FCM, *Composites Part B: Engineering*, Vol. 161, pp. 547-554, March, 2019.

[10] A. Majdi, M. Beiki, Applying Evolutionary Optimization Algorithms for Improving Fuzzy C-Mean Clustering Performance to Predict the Deformation Modulus of Rock Mass, *International Journal of Rock Mechanics and Mining Sciences*, Vol. 113, pp. 172-182, January, 2019.

[11] A. R. Abas, Unsupervised Learning of Mixture Models Based on Swarm Intelligence and Neural Networks with Optimal Completion Using Incomplete Data, *Egyptian Informatics Journal*, Vol. 13, No. 2, pp. 103-109, July, 2012.

[12] H. C. Lin, C. T. Su, A Selective Bayes Classifier with Meta-Heuristics for Incomplete Data, *Neurocomputing*, Vol. 106, pp. 95-102, April, 2013.

[13] W. Ding, P. X.-K. Song, EM Algorithm in Gaussian Copula with Missing Data, *Computational Statistics & Data Analysis*, Vol. 101, pp. 1-11, September, 2016.

[14] X. C. Lai, X. Wu, L. Y. Zhang, W. Lu, C. Q. Zhong, Imputations of Missing Values Using a Tracking-removed Autoencoder Trained with Incomplete Data, *Neurocomputing*, Vol. 366, pp. 54-65, November, 2019.

[15] L. Zhang, H. Pan, B. L. Wang, L. Y Zhang, Z. J. Fu, Interval Fuzzy C-Means Approach for Incomplete Data Clustering Based on Neural Networks, *Journal of Internet Technology*, Vol. 19, No. 4, pp. 1089-1098, July, 2018.

[16] W. J. Zhang, G. Y. Xu, M. J. Li, S. Zhu, Missing Data Imputation Approach Based on Convolutional Neural Network, *Microelectronics and computer*, Vol. 36, No. 3, pp. 48-52+57, March, 2019.

[17] J. Dong, Y. X. Zhao, L. Chang, Z.-F. Han, C.-S. Leung, Orthogonal Least Squares Based Center Selection for Fault-tolerant RBF Networks, *Neurocomputing*, Vol. 339, pp. 217-231, April, 2019.

[18] L. D. Su, A Radial Basis Function (RBF)-Finite Difference (FD) Method for the Backward Heat Conduction Problem, *Applied Mathematics and Computation*, Vol. 354, pp. 232-247, August, 2019.

[19] Z. S. Xu, X. J. Gou, An Overview of Interval-valued Intuitionistic Fuzzy Information Aggregations and Applications, *Granular Computing*, Vol. 2, No. 1, pp. 13-39, March, 2017.

[20] T. H. Li, L. Y. Zhang, W. Lu, H. Hou, X. D. Liu, W. Pedrycz, C. Q. Zhong, Interval Kernel Fuzzy C-Means Clustering of Incomplete Data, *Neurocomputing*, Vol. 237, pp. 316-331, May, 2017.

[21] V. N. Pham, L. T. Ngo, W. Pedrycz, Interval-valued Fuzzy Set Approach to Fuzzy Co-clustering for Data Classification, *Knowledge-Based Systems*, Vol. 107, pp. 1-13, September, 2016.

[22] L. T. Ngo, T. H. Dang, W. Pedrycz, Towards Interval-valued Fuzzy Set-based Collaborative Fuzzy Clustering Algorithms, *Pattern Recognition*, Vol. 81, pp. 404-416, September, 2018.

[23] J. H. Li, S. J. Song, Y. L. Zhang, K. Li, A Robust Fuzzy C-Means Clustering Algorithm for Incomplete Data, *International Conference on Intelligent Computing for Sustainable Energy and Environment (ICSEE 2017), International Conference on Life System Modeling and Simulation (LSMS 2017)*, Nanjing, China, 2017, pp. 3-12.

[24] R. Sarrazin, Y. D. Smet, J. Rosenfeld, An Extension of PROMETHEE to Interval Clustering, *Omega*, Vol. 80, pp. 12-21, October, 2018.

[25] A. K. Shukla, P. K. Muhuri, Big-data Clustering with Interval Type-2 Fuzzy Uncertainty Modeling in Gene Expression Datasets, *Engineering Applications of Artificial Intelligence*, Vol. 77, pp. 268-282, January, 2019.

[26] L. W. Feng, C. Zhang, Y. Li, Y. H. Xie, Fault Detection for Multistage Process Based on Improved Local Neighborhood Standardization and KNN, *Journal of Computer applications*, Vol. 38, No. 7, pp. 2130-2135, July, 2018.

[27] P. H. Kassani, A. B. J. Teoh, E. Kim, Evolutionary-modified Fuzzy Nearest-neighbor Rule for Pattern Classification, *Expert Systems with Applications*, Vol. 88, pp. 258-269, December, 2017.

[28] L. X. Zhao, C. X. Dong, L. Zhao, WSN Weighted Probability Cluster Head Selection Algorithm Based on FCM Clustering, *Control Engineering of China*, Vol. 26, No. 6, pp. 1211-1215, June, 2019.

[29] S. V. Muravyov, L. I. Khudonogova, E. Y. Emelyanova, Interval Data Fusion with Preference Aggregation, *Measurement*, Vol. 116, pp. 621-630, February, 2018.

[30] B. A. Pimentel, R. M. C. R. de Souza, A Multivariate Fuzzy C-Means Method, *Applied Soft Computing*, Vol. 13, No. 4, pp. 1592-1607, April, 2013.

[31] C. C. Yeh, M. S. Yang, Evaluation Measures for Cluster Ensembles Based on a Fuzzy Generalized Rand Index, *Applied Soft Computing*, Vol. 57, pp. 225-234, August, 2017.

Biographies

Hao Luo was born in Liaoning Province, China in 1987. He received the Ph.D. degree in engineering mechanics from Liaoning Techanical University, Fuxin, in 2016. He works as a lecturer at Liaoning University, China. His research interests include monitoring and early warning equipment research of coal mine, and data mining.

Qing Hou was born in Liaoning Province, China. She is a M.S. candidate at Liaoning University, Shenyang, China. Her research interests is fuzzy clustering for incomplete data.

Yang Liu was born in Liaoning Province, China. She received the Master degrees from Liaoning University, Shenyang, China. Her research interests is fuzzy clustering for incomplete data.

Li Zhang received the Ph.D. degree in material processing engineering from Northeastern University, China, in 2000. From 2002 to 2009, he was with Dalian University of Technology, Institute of Automation, China. Since 2009, he is with Liaoning University, Department of Computer Science, China. His research interests include data cluster and fault diagnosis.

Yuanzhi Li received her bachelor degree in computer science from Columbia University, in 2019. From 2014 to 2017, she was in Bard College, studying mathematics. Since 2019, she is pursuing her master degree in computer science engineering from Santa Clara University. Her research interests include machine learning and neural network.