Hybrid Missing Value Imputation Algorithms Using Fuzzy C-Means and Vaguely Quantified Rough Set

Daiwei Li, Haiqing Zhang, Tianrui Li, *Senior Member, IEEE*, Abdelaziz Bouras, *Member, IEEE*, Xi Yu, and Tao Wang

***Abstract*—In real cases, missing values tend to contain meaningful information that should be acquired or should be analyzed before the incomplete dataset is used for machine learning tasks. In this work, two algorithms named jointly fuzzy c-means and VQNN (Vaguely Quantified Nearest Neighbor) imputation (JFCM-VQNNI) and jointly fuzzy c-means and fitted VQNN imputation (JFCM-FVQNNI) have been proposed by considering clustering conception and sufficient extraction of uncertain information. In the proposed JFCM-VQNNI and JFCM-FVQNNI algorithm, the missing value is regarded as a decision feature and then the prediction is generated for the objects that containing at least one missing value. Specially, as for JFCM-VQNNI algorithm, indistinguishable matrixes, tolerance relations, and fuzzy membership relations are adopted to identify the potential closest filled values based on corresponding similar objects and related clusters. On the basis of JFCM-VQNNI algorithm, JFCM-FVQNNI algorithm synthetic analyzes the fuzzy membership of the dependent features for instances with each cluster. In order to fill the missing values more accurately, JFCM-FVQNNI algorithm performs fuzzy decision membership adjustment in each object with respect to the related clusters by considering highly relevant decision attributes. The experiments have been carried out on five datasets. Based on the analysis of RMSE, MAE, imputation values with actual values comparison, and classification accuracy results analysis, we can draw the conclusion that the proposed JFCM-FVQNNI and JFCM-VQNNI algorithms yields sufficient and reasonable imputation performance results by comparing with fuzzy c-means parameter-based imputation algorithm and fuzzy c-means rough parameter-based imputation algorithm.**

***Index Terms*—Missing value imputation (MVI), fuzzy membership relations, fuzzy c-means clustering imputation, rough set, nearest neighbor imputation.**

I. Introduction

MISSING dataset tends to happen due to several reasons, for instance, malfunctioning measurement equipment, incorrection data collection operation, unexpected changes in experimental sessions during data collection, and management of similar but not identical datasets. Some rows of data in missing dataset may contain one or more attributes are not present. These rows with missing dataset will cause the problems of introducing a substantial amount of bias, increasing the difficulties of analyzing dataset, and reducing the efficiency of research results. Although some research works related to missing value imputation have emerged, still most direct default way of handling missing dataset is to discard the missing cases or impute by using simple statistic methods, which may introduce bias or affect the effectiveness of the analysis results [1]. Thus, according to the comprehensive internal and external relationship analysis among data objects, it is essential to propose a high-performance algorithm that can replace missing data with reasonable estimated value and preserve the

This research is supported by the National Natural Science Foundation of China (NSFC) (No. 61602064), the Sichuan Province Science and Technology Program (Nos. 2018JY0273, 2018GZDZX0028, and 2019YFG0398), and Building Skills 4.0 through University and Enterprise Collaboration (Erasmus+ Shyfte 4.0) project (Erasmus+Programme：598649-EPP-1-2018-1FR-EPPKA2-CBHE-JP, http://shyfte.eu).

Daiwei Li is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China and also with the College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China (e-mail: ldwcuit@cuit.edu.cn).

Haiqing Zhang (Corresponding author) is with the College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China (e-mail: haiqing_zhang_zhq@163.com).

Tianrui Li is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: trli@swjtu.edu.cn).

Abdelaziz Bouras is with the Computer Science Department, Qatar University, ictQATAR, Box.2731, Doha, Qatar (e-mail: abdelaziz.bouras@qu.edu.qa).

Xi Yu is with the College of Information Science and Technology, Chengdu University, Chengdu 610106, China (e-mail: yuxi@cdu.edu.cn).

Tao Wang is with the DISP Laboratory, INSA Lyon, UJM-Saint Etienne 69621, France (e-mail: tao.wang@univ-st-etienne.fr).

original information to a maximum extend.

According to the missingness mechanism that how the missing dataset generalized and also the relations with the dataset itself, the missing types can be concluded into three categories in terms of missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [1]. As for the type of MCAR, the missing probability is same for all items and deleting the cases with missing data will not bias the inferences. In practice, most missing cases are not MCAR. A more general missing case is called MAR, which the missing probability of a variable depends only on other available recorded variables. As for the type of NMAR, it can be categorized into two situations of missingness that depends on unobserved predictors and missingness that depends on the missing value itself. For instance, some missing values may depend on some information that has not been written down, however, the unobserved information may also need to be used to forecast the missing values. In order to reduce the bias in the inference of results, this type missingness should be explicitly modeled. Another particularly difficult situation in NMAR is that the missing items have strong correlation with other missing items. In practice, it is very difficult to infer the missing values based on the other unobserved values, thus current research works mainly focus on the missing type of MAR.

Based on literature review, the approaches of missing value imputation (MVI) are mainly categorized into two groups. One group is statistical technique that includes mean, expectation maximization (EM) [2], linear regression [3], and least squares imputation [4], etc. And the other group is machine learning based techniques that contain fuzzy c-means based clustering approaches [5], [6], [7], decision tree (DT) methods [8], nearest neighbor-based algorithms [9], random forest (RF) [10], and rough/fuzzy set-based algorithms [9], [11], etc.

First of all, the details of statistical imputation techniques are discussed. Mean imputation tends to fill the average value of all of the existing data to the missing item in the same missing attribute. The EM approach contains two main steps named as expectation step (E-Step) and the maximization step (M-Step) [2]. Firstly, the missing values are assigned based on the predicted mean and covariance values. And then the imputation values are judged by the optimal possible strategy based on maximum likelihood approach. Next, the assigned mean and covariance values are updated based on the M-Step by considering the imputed values. The whole process is iterative until the efficiency of imputation task stops increasing. As for linear regression-based imputation approaches, the relations among features are studied, and then the corresponding regression coefficients are adopted to predict the missing values. In terms of least squares imputation, it has many variants and it has been used less frequently in recent years [1]. In conclusion, since the statistical imputation techniques are easy and simple to operate, meanwhile, they also have the basic filling effect, thus, statistical imputation techniques are usually combined with other advanced techniques to conduct the imputation task and also often used as the benchmark comparison methods [2], [3], [4], [12], [13].

Next, the machine learning based techniques are further studied. Fuzzy c-means based clustering approaches are recognized as unsupervised learning means that category the objects with similar relationships into one group [5], [6], [7]. Typically, clustering the objects is based on the cluster center and the membership degree of the object. Two indicators are very important to obtain more accurate filling values in this type method. One is the distance that is calculated between the incomplete data and the corresponding cluster centroids, and the other one is the relations that are defined between the incomplete objects and the membership function with the center objects. In decision tree models, the tree nodes represent the test of attributes and the generation of the tree usually based on largest entropy (or weight) [8]. Based on literature review, C4.5 and CART are the most used tree imputation strategies for categorical and numerical values. In RF, adding the bootstrapping strategy by comparing with DT, multiple decision trees are used and the final missing imputation values are assigned based on comprehensive analyze of the majority votes of multiple trees [10]. Usually, nearest neighbor-based algorithms are based on supervised learning strategy, which also the most commonly used missing value filling approaches [1], [9]. The nearest neighbors between complete and incomplete objects tend to be calculated by distance functions or similarity relations. The missing values are used as the testing cases, and the complete and missing features indicate as input data and predict data label. Rough/fuzzy set-based algorithms introduce rough/rough set into MVI, which are proved to be effective approaches to handle vagueness and uncertainty information in MVI [9], [11]. However, only a few researches work currently adopts this type method.

Based on above analysis, we can draw the conclusion that three issues are vital important in MVI problem solving. The first one is to reduce bias and noise and reduce the complexity of introducing parameters in missing value handling approaches. The second one is to reduce the time complexity under the circumstances of finding the global optimum point. The last one is to maximum keep the original information to increase the imputation efficiency. In order to reduce the limitations of current research works and aim to solve above three issues, this paper has developed a novel hybrid approach by combining fuzzy c-mean technique, nearest neighbor

conception, and rough set.

The reminder of this paper is organized as follows. Section 2 reviews the necessary theoretical background. The running mechanisms of fuzzy c-means, rough c-means, and rough-fuzzy c-means algorithms have been detailed analyzed. The MVI approaches related to cluster algorithms in terms of fuzzy c-means parameter-based imputation algorithm and fuzzy c-means rough parameter-based imputation algorithm have been analyzed carefully. Section 3 details the proposed jointly fuzzy c-means and VQNN and jointly fuzzy c-means and fitted VQNN Imputation algorithms. The fuzzy decision membership of the target nearest neighbors has been adjusted, and the fuzzy similarity relations between training and testing instances has been balanced. Comparative experiments have been conducted to judge the performance of the proposed algorithm and parameters have been analyzed in Section 4. Finally, we conclude our work and present future challenges in Section 5.

II. THEORETICAL BACKGROUND

Missing value imputation is vital important for improving data quality, which is also a quite challenging task. Based on the above explanation in literature review, there is no holistically effective method for MVI. In this paper, the conception of center cluster is used to explore the closest proper imputation value for missing values.

A. Fuzzy C-Means algorithm

The fuzzy C-means (FCM) clustering algorithm has been adopted to handle the MVI task and proved to be capable of predicting missing values efficiently [14], [5] [6]. In FCM, each object is determined by a fuzzy membership function that shows how much degree of the similarity between objects and clusters. The greater of the fuzzy membership function, and then the greater of the similarity degree between the object and the cluster. In order to clearly utilize FCM algorithm, the notation definitions are given in the following explanation.

Suppose $\mathrm{X}_{n\times f}$ is the raw matrix, *n* represents the row and *f* represents the corresponding features for each row. Let *c* denotes the expected number of clusters. In order to better increase the predication accuracy, we consider the involved f features for each cluster $cluster_i (1 \le i \le c)$, which can be shown as $cluster_i = \{cluster_{1i}, cluster_{2i}, \cdots, cluster_{ji}, cluster_{fi}\}$. The membership degree $u(x_{kf}, cluster_i)$ is used to judge the close degree between the object x_{kf} and the cluster $cluster_i$. $m_{cluster_i}$ is defined as the cluster center of $cluster_i$.

Definition 1 [15]. The objective of the FCM algorithm is to minimize the following equation (Eq. (1)):

$$\min(J_v) = \sum_{k=1}^{N}\sum_{i=1}^{C} u^v(x_{kf}, cluster_i) x_{kf} - m^2{}_{cluster_i} \quad (1)$$

Where, *v* indicates the fuzziness degree of clusters. Meanwhile, $u(x_{kf}, cluster_i)$ and $m_{cluster_i}$ are defined in Eq. (2) and Eq. (3), respectively.

$$u(x_{kf}, cluster_i) = \frac{1}{\sum_{j=1}^{C} \left(\frac{\left\| x_{kf} - m_{cluster_i} \right\|}{\left\| x_{kf} - m_{cluster_j} \right\|} \right)^{\frac{2}{v-1}}} \quad (2)$$

$$m_{cluster_i} = \frac{\sum_{i=1}^{N} u^v(x_{kf}, cluster_i) \cdot x_{kf}}{\sum_{i=1}^{N} u^v(x_{kf}, cluster_i)} \quad (3)$$

We should emphasize that there is no theoretical proof of how to select the optimal value of cluster number c and the adjustment fuzziness parameter of v in current research works. Therefore, in this paper, based on practical experiments, the values of v and c can be capable of obtaining by the analysis of dataset characteristics and the relation of features.

The FCM algorithm can be summarized as the following steps:

Step 1. Initialize the corresponding membership function of each object and the centroid based on Eq. (2) and Eq. (3), respectively. Note: $U^0 = \left[u(x_{kf}, cluster_i) \right]$

Step 2. Suppose in *k*-step, calculate and update the centroid $m_{cluster_i}$ with U^k.

Step 3. Calculate and update the membership function of U^k and U^{k+1}.

Step 4. The FCM algorithm is stop when the gap between two adjacent iterations is less than the termination condition, which is noted as $\left\| U^{k+1} - U^k \right\| < \theta$.

B. Rough C-Means and Rough-Fuzzy C-Means Algorithm

1) Rough set Theory

Rough set theory [16] is widely used to handle dataset in various research tasks due to its capability of dealing with vagueness information. The crisp (or imprecision) value is expressed by a boundary region of lower and upper approximation. Rough set theory is developed from the indiscernibility relation that each object in the universe of discourse is intrinsically linked and the objects characterized by the same relation (or information) are recognized as indiscernible.

Definition 2 [16]. Let the tuple (*U*, *A*) be an information system, where U is a nonempty finite set of objects and *A* is the corresponding attributes. The nonempty finite set of attributes A is the union of sets

C and D (and $C \cap D = \varnothing$), where C and D represent condition attributes and decision attributes, respectively. For each attribute $a_i \in A$, the domain of a_i named V_{a_i} is determined based on the associated value set of attributes. An information function f is defined as $f: U \times A \to V$, which is used to assign values for each object on its corresponding attributes. For an object x in U, the mathematical expression of information function is shown as: $\forall a_i \in A, x \in U, f(x, a_i) \in V_{a_i}$.

Definition 3 [16]. The indiscernibility relation is an equivalence relation, let the subset of attributes $F \subseteq A$, and then the indiscernibility relation of attributes F is defined as follows:

$$IND(F) = \{(x, y) \in U \times U : f(x, a) = f(y, a), \forall a \in F\} \quad (4)$$

Based on the indiscernibility relation definition, let R be an equivalence relation on U, $R \subseteq U \times U$. The tuple (U, R) is an approximation space. The equivalence relation R caused a partition on U, which is U / R. The equivalence class involving object x is marked as relation $[x]_R = \{y \mid (x, y) \in R\}$.

Definition 4 [16]. Based on the equivalence class, for each object $X \subseteq U$, the lower and upper approximations of object X can be constructed in Eq. (5) and Eq. (6), respectively.

$$\underline{R}(X) = \bigcup \{Y \in U / R \mid Y \subseteq X\} \quad (5)$$

$$\overline{R}(X) = \bigcup \{Y \in U / R \mid Y \cap X \neq \phi\} \quad (6)$$

2) Rough C-Means Algorithm

Suppose the lower and upper approximations of cluster α_i is $\underline{A}(\alpha_i)$ and $\overline{A}(\alpha_i)$, and the boundary region of cluster $B(\alpha_i)$ is denoted as $B(\alpha_i) = \{\underline{A}(\alpha_i) - \overline{A}(\alpha_i)\}$. In **R**ough **C**-**M**eans (RCM) algorithm, the cluster α_i is analyzed by the lower and upper approximations in rough set and also follow the fundamental rough set properties.

RCM algorithm [17] is proposed by adding the concept of lower and upper approximation into clustering task. The space of the objects is classified into boundary and approximation region. The mean value of cluster centroid is obtained based on the weighted lower approximation and boundary region.

Definition 5 [17]. The cluster center m_{α_i} in RCM is defined as follows:

$$m_{\alpha_i} = \begin{cases} \omega \times \mathrm{A} + \tilde{\omega} \times \mathrm{B} & \text{if } \underline{\mathrm{A}}(\alpha_i) \neq \phi, \mathrm{B}(\alpha_i) \neq \phi \\ \mathrm{A} & \text{if } \underline{\mathrm{A}}(\alpha_i) \neq \phi, \mathrm{B}(\alpha_i) = \phi \\ \mathrm{B} & \text{if } \underline{\mathrm{A}}(\alpha_i) = \phi, \mathrm{B}(\alpha_i) \neq \phi \end{cases} \quad (7)$$

$$\mathrm{A} = \frac{1}{\left|\underline{\mathrm{A}}(\alpha_i)\right|} \sum_{x_j \in \underline{\mathrm{A}}(\alpha_i)} x_j \text{ ; and } \mathrm{B} = \frac{1}{\left|\mathrm{B}(\alpha_i)\right|} \sum_{x_j \in \mathrm{B}(\alpha_i)} x_j \quad (8)$$

Where, the parameter ω means the weight in lower approximation for all objects, parameter $\tilde{\omega}$ means the weight in boundary region for the corresponding objects, and it has the relation: $\tilde{\omega} = 1 - \omega$. The centroid is affected by the lower and boundary region based on Eq. (7).

Based on above analysis, the main steps of RCM algorithm can be concluded as follows:

Step 1. Initialize the parameters of $m_{\alpha_i} (i = 1, 2, \cdots, c)$.

Step 2. Calculate the distance d_{ij} between the objects and the centroid value m_{α_i} of cluster α_i.

Step 3. Suppose d_{ij} reach to the minimum for all clusters and $(d_{ij} - d_{kj}) \leq \eta$, then $x_j \in \overline{\mathrm{A}}(\alpha_i)$ and $x_j \in \overline{\mathrm{A}}(\alpha_k)$, meanwhile, x_j is not belong to the lower bound.

Step 4. In the other circumstances, $x_j \in \underline{\mathrm{A}}(\alpha_i)$ and d_{ij} is minimum for all clusters. Based on the properties of rough set, the object x_j should also belong to the upper approximation of cluster α_i, which is $x_j \in \overline{\mathrm{A}}(\alpha_i)$.

Step 5. Update the cluster centroid based on Eq. (7).

Step 6. Repeat the steps from 2 to 5 until convergence, which means all of the clusters have been assigned for all objects and no more updating.

3) Rough-Fuzzy C-Means Algorithm

In order to combine the advantages of fuzzy set and rough set in terms of efficiently dealing with overlapping regions and effectively handling uncertainty and vagueness information, rough-fuzzy c-means (RFCM) algorithm [7] is then proposed by incorporating rough and fuzzy sets. RFCM algorithm allows the incorporation of fuzzy membership value with cluster centroid, and also enhances the cluster by fuzzy lower approximation and fuzzy boundary. RFCM algorithm is capable of handling the overlapping partitions efficiently by comparing with FCM and RCM algorithms, which has been applied in many practical cases[19], [20], [21].

Definition 6 [18]. The objective of RFCM algorithm is to minimize the following equation (Eq. (9)):

$$\min(J_{RFC}) = \begin{cases} \omega \times \mathrm{P} + \tilde{\omega} \times \mathrm{Z} & \text{if } \underline{\mathrm{A}}(\alpha_i) \neq \phi, \mathrm{B}(\alpha_i) \neq \phi \\ \mathrm{P} & \text{if } \underline{\mathrm{A}}(\alpha_i) \neq \phi, \mathrm{B}(\alpha_i) = \phi \\ \mathrm{Z} & \text{if } \underline{\mathrm{A}}(\alpha_i) = \phi, \mathrm{B}(\alpha_i) \neq \phi \end{cases} \quad (9)$$

$$\text{Where, } \mathrm{P} = \sum_{i=1}^{c} \sum_{x_{kf} \in \underline{\mathrm{A}}(\alpha_i)} u^{v}(x_{kf}, \alpha_i) \left\| x_{kf} - m_{\alpha_i} \right\|^2 \quad (10)$$

and $Z = \sum_{i=1}^{c} \sum_{x_{kf} \in Z(\alpha_i)} u^{v}(x_{kf}, \alpha_i) \left\| x_{kf} - m_{\alpha_i} \right\|^2$ (11)

Definition 7 [18]**.** The centroid of RFCM is obtained based on the lower approximation and fuzzy boundary. The new centroid of RFCM effected by fuzzy memberships and lower and upper bounds. In order to solve Eq. (9), the modified centroid is calculated as follows:

$$m_{\alpha_i} = \begin{cases} \omega \times q + \tilde{\omega} \times l & \text{if } \underline{\mathrm{A}}(\alpha_i) \neq \phi, \mathrm{B}(\alpha_i) \neq \phi \\ q & \text{if } \underline{\mathrm{A}}(\alpha_i) \neq \phi, \mathrm{B}(\alpha_i) = \phi \\ l & \text{if } \underline{\mathrm{A}}(\alpha_i) = \phi, \mathrm{B}(\alpha_i) \neq \phi \end{cases} \quad (12)$$

Where, $q = \frac{1}{|\underline{\mathrm{A}}(\alpha_i)|} \sum_{x_{kf} \in \underline{\mathrm{A}}(\alpha_i)} x_{kf}$ (13)

$$l = \frac{1}{\sum_{x_{kf} \in \mathrm{B}(\alpha_i)} u^{v}(x_{kf}, \alpha_i)} u^{v}(x_{kf}, \alpha_i) \times x_{kf} \quad (14)$$

The main steps of RFCM algorithm can be summarized as follows:

Step 1. Initialize the centroids $m_{\alpha_i} (i = 1, 2, \cdots, c)$ for all clusters.

Step 2. Calculate the fuzzy membership function $u^{v}(x_{kf}, \alpha_i)$ based on Eq. (2) for all clusters with the corresponding objects.

Step 3. Let $u(x_{kf}, \alpha_i)$ and $u(x_{kf}, \alpha_j)$ be the highest membership of object x_{kf}.

if $(u(x_{kf}, \alpha_i) - u(x_{kf}, \alpha_j)) \leq \delta$

then the object x_{kf} belong to the upper approximations $\overline{\mathrm{A}}(\alpha_i)$ and $\overline{\mathrm{A}}(\alpha_j)$ corresponding with clusters α_i and α_j.

else $x_{kf} \in \underline{\mathrm{A}}(\alpha_i)$. In addition, based on the properties of rough set, it also has the relation $x_{kf} \in \overline{\mathrm{A}}(\alpha_i)$.

Step 4. Updating the fuzzy membership $u(x_{kf}, \alpha_i)$ by considering lower and boundary regions for clusters and objects.

Step 5. Updating new centroid value based on Eq. (12).

Step 6. **Repeat** Steps 2 to 5 **until** convergence, and record the number of iterations.

In order to deal with the problems of uncertainty and incompleteness in RFCM algorithm, the target objects are handled by the combination of fuzzy set theory and the lower and upper estimation of rough sets. Extensive performance has been conducted to compare RFCM, FCM, and RCM. Here, we only show part of the results. Since RFCM is an advanced method by comparing with RCM [22], thus we focus on advantages and disadvantages analysis of RFCM and FCM. We randomly generate 10000 objects in two dimensions, and cluster them into three, six, nine, and twelve groups by using the algorithms of RFCM and FCM, respectively. The results are shown from Fig.1 to Fig.4 and the cluster center is marked as "*". The cluster generated by RFCM algorithm consists of cluster centroid, lower approximation, and fuzzy boundary from Fig.1 to Fig.4.

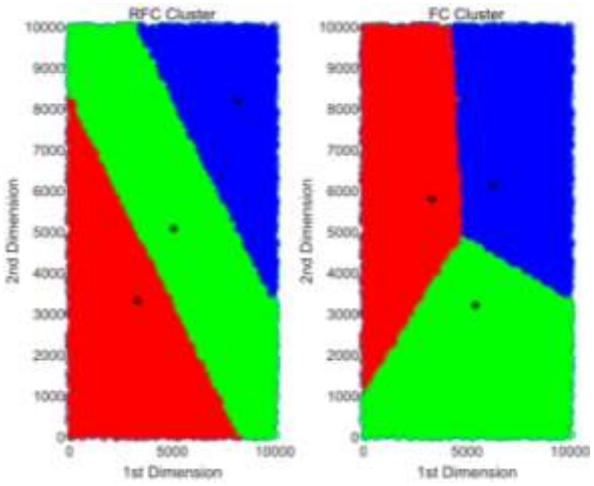

Fig.1. RFCM and FCM comparation: three clusters

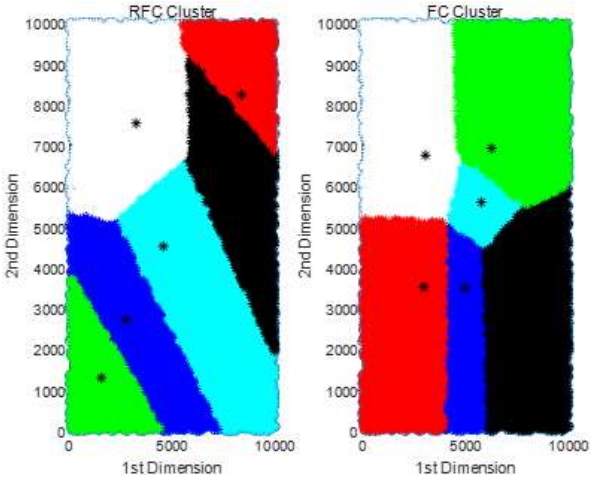

Fig.2. RFCM and FCM comparation: six clusters

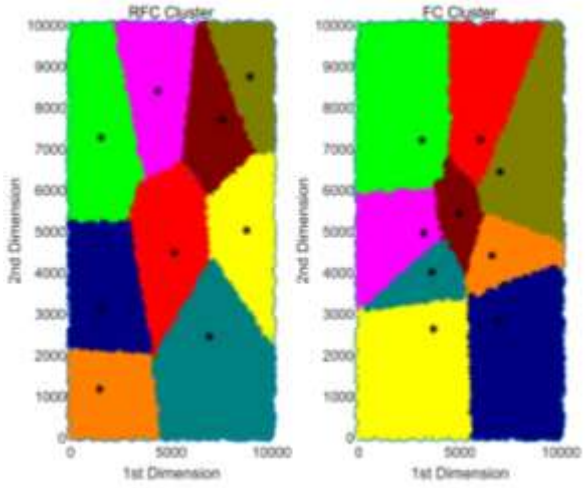

Fig.3. RFCM and FCM comparation: nine clusters

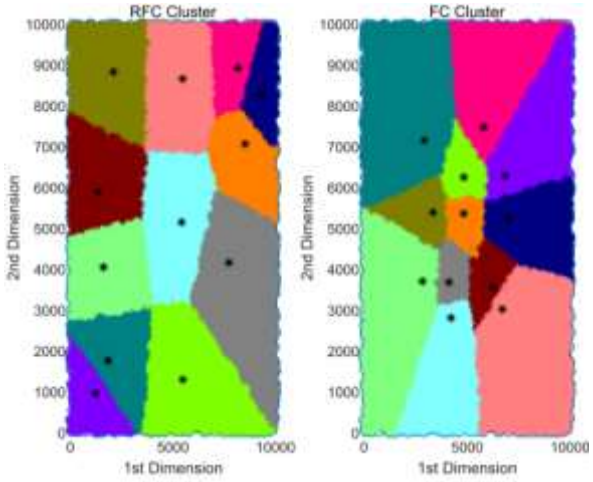

Fig.4. RFCM and FCM comparation: twelve clusters

The experimental results have shown that RFCM algorithm can reduce the fuzziness of FCM and can handle fuzziness and incompleteness more efficiently by comparing with FCM and RCM. In addition, RFCM can have higher similarity values within the clusters and higher distance between cluster separation. Furthermore, the RFCM handles uncertainty and incompleteness in class level based on the conception of fuzzy boundary and lower and upper approximation. The defined membership function of RFCM can efficiently handle overlapping partitions by comparing with FCM cluster algorithm based on the results in Fig.1-Fig.4.

C. Missing Value Imputation Based on Cluster Algorithms

The imputation algorithms based on cluster algorithms can be categorized into two groups of centroid-based imputation methods and parameter-based imputation methods [23]. The running mechanism of the two types of algorithms will be explained in the following.

In the first category of centroid-based method, the raw dataset can be divided into complete objects and missing objects. Firstly, the cluster algorithm is employed on complete objects and then the complete objects are assigned into several clusters based on calculated centroids. The centroid is calculated based on the mean value of each attribute in each cluster. Secondly, the missing values of each object is imputed one by one. The missing column value of an object is changed by the corresponding column values of the related centroid. The missing column value of an object is assigned the value of the minimum distance between centroids and also the object will be placed in the minimum distance cluster. Finally, perform the same process till all missing values are assigned. The centroid-based imputation methods are *k*-means centroid-based imputation algorithm, fuzzy C-means centroid-based imputation algorithm [7], and rough *k*-means centroid-based imputation algorithm [24]. In detail, as for the fuzzy C-means centroid-based imputation algorithm, the objects are clustered based on the fuzzy membership function and the outstanding segmentation of objects into a cluster is achieved by the higher fuzzy membership. Meanwhile, the fuzzy membership function of all objects is calculated with respect to all clusters. In addition, the cluster center is calculated based on Eq. (3) and the missing values are imputed based on the cluster center. Finally, the missing value is iteratively updated based on the distance analysis with each center. Similarly, the rough *k*-means centroid-based imputation algorithm treats the missing values by using centroid values. In conclusion, the fuzzy C-means centroid-based and rough *k*-means centroid-based algorithm, the accuracy of filling imputation values depends on the gap between the objects with the corresponding clusters, which is the inherent disadvantage of centroid-based algorithm.

In parameter-based imputation algorithms, the missing values of an object are filled in based on the relations among clusters and also the properties of cluster [24]. Similarly, the original dataset is separated into complete dataset and missing dataset. Next, the cluster imputation algorithm is used to partition the complete objects into several clusters based on the value of centroid. In addition, the information of the closest object within the cluster is applied to achieve the optimal accuracy to the missing object. For instance, in fuzzy C-means parameter-based imputation algorithm, the missing value is assigned based on the combination of cluster centroid and the fuzzy membership function. In rough *k*-means parameter-based imputation algorithm, the missing value is obtained based on the closest approximation objects. Some representative parameter-based algorithms are described in the following.

1) Fuzzy C-Means Parameter-based Imputation Algorithm

The **F**uzzy **C**-**M**eans **P**arameter-based (Short for FCMP) imputation algorithm [5] is proposed to overcome the limitations of FCM centroid-based imputation algorithm. This algorithm imputes the missing values based on the combination of membership values and cluster centroid values. The steps of the FCMP imputation algorithm can be concluded as follows:

Step 1. Splitting the raw matrix $X_{n\times f}$ into two parts *CX* (objects with complete objects) and *MX* (objects with missing value objects). Which has $CX \bigcup MX = X$.

Step 2. Applying FCM algorithm to *CX* and grouping *CX* into *C* clusters.

Step 3. Obtaining the centroid value $m_{cluster_i}(i = 1,2,\cdots,C)$ for clusters (complete objects) and figuring out the membership functions of *MX* for all clusters based on Eq. (2).

Step 4. The missing values of *MX* is calculated based on the Eq. (15).

$$mx_{pq} = \sum_{i=1}^{C} u(mx_{pq}, cluster_i) \times m_{p,cluster_i}, p \le n, q \le f \quad (15)$$

Fig.5 illustrates that how to assign the missing values based on FCM parameter-based imputation. All objects are divided into three clusters (Represented in red, magenta, and blue in Fig.5) based on FCM clustering algorithm. Suppose the mark '?' (blackened rectangle) is one of the missing values that should be imputed. And then, based on FCM parameter-based imputation, building the relations of the missing object, the membership functions, and the cluster centroid values. Suppose the membership values of the missing object are 0.4,0.4, and 0.1, and the cluster centroids are inferred as 0.2, 0.45, and 0.7. Then, the missing object is computed as $'?' = 0.4*0.2 + 0.4*0.45 + 0.1*0.7 = 0.33$.

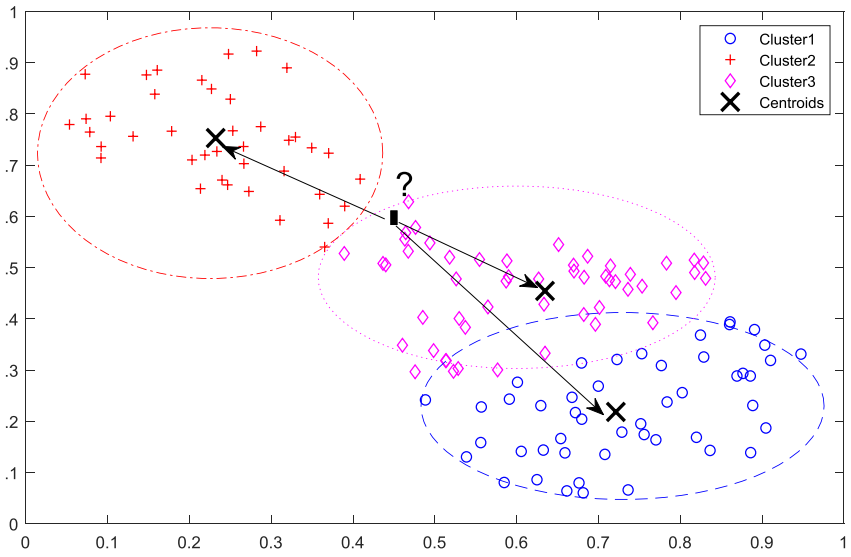

Fig.5. Fuzzy C-Means Parameter-based Imputation

2) Fuzzy C-Means Rough Parameter-based imputation algorithm

The **F**uzzy **C**-**M**eans **R**ough **P**arameter-based (FCMRP) imputation algorithm is proposed by research work [5] to enhance the capabilities of handling vagueness information in missing dataset. The FCMRP algorithm can be summarized into three main steps, which are: 1). Dividing the complete dataset into several clusters based on FCM algorithm; 2). Searching the closest center and the nearest approximation for the incomplete instance in each cluster; 3). Filling the missing values based on the lower and upper approximation of the target objects. The procedures of FCMRP are detailed in the following steps:

Step 1. Splitting the raw matrix $X_{n\times f}$ into two parts *CX* (objects with complete objects) and *MX* (objects with missing value objects). Which has the relation: $CX \bigcup MX = X$.

Step 2. Applying FCM algorithm to *CX* and gathering all *CX* into *C* clusters.

Step 3. Obtaining the centroid value $m_{cluster_i}(i = 1,2,\cdots,C)$ for clusters (complete objects) and figuring out the membership functions of *MX* for all clusters based on Eq. (2).

Step 4. Applying RCM to the clusters based on the previous steps 1-3.

Step 5. The missing attribute value in MX_i is imputed by the related attribute value m_j based on distance analysis, which is shown in the following equation:

$$\begin{aligned} \min(d_{i,j}) &= \text{distance}(MX_i, m_j) \\ &= \min_{c=1,2,\cdots,C} \text{distance}(MX_i, m_c) \end{aligned} \quad (16)$$

Where, distance(MX_i, m_c) represents the Euclidean distance between the object MX_i and the cluster center.

Step 6. Based on the RCM, searching the closest approximation value to cluster $Cluster_i$ and then use that value to fill in the missing value, which is shown in the following equation:

$$MX_i = \begin{cases} \dfrac{\sum_{Cluster_i \in \underline{A}(Cluster)} CX_i}{|Cluster|}, if\ lower \\ \dfrac{\sum_{Cluster_i \in \overline{A}(Cluster)} CX_i}{|Cluster|}, if\ upper \end{cases} \quad (17)$$

Step 7. Repeat step 5 and step 6 for all instances in MX till all missing values have been filled.

Based on above description of FCMRP imputation algorithm, Fig.6 illustrates how it works of FCMRP. All objects in Fig.6 are divided into three clusters based on the fuzzy membership of FCM, which is shown in continuous circles of red, blue, and magenta. And then, all of the instances in one cluster are gathered into two sub-clusters based on RCM. The black continued circles mean upper approximation and the black dotted circles represent lower approximation in each sub-cluster. For instance, in order to fill in the missing instances in sub-cluster C11, the lower approximation of sub-cluster C11 will be selected if the missing instance is present in lower approximation, otherwise, the upper approximation will be signed for current missing instance. And if the missing instances belong to more than one sub-cluster, Eq. (17) will be used to impute the missing value.

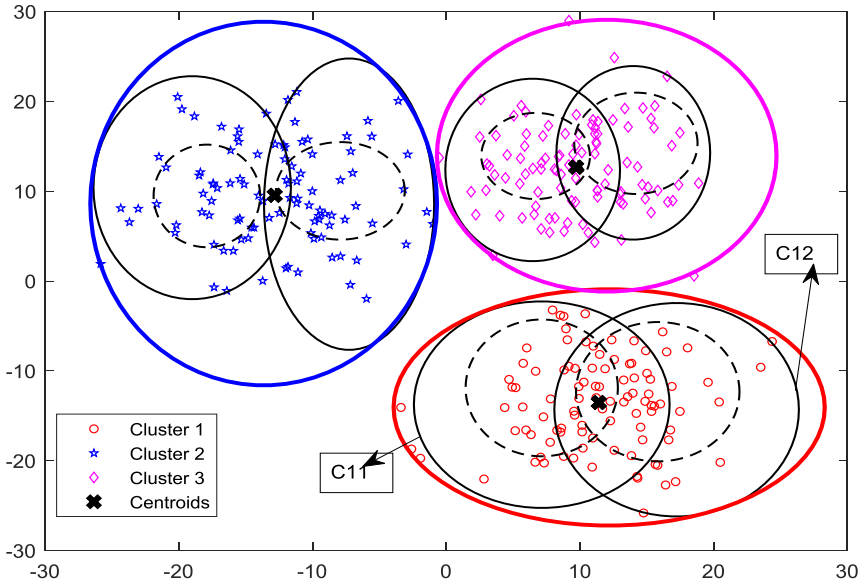

Fig.6. Fuzzy C-Means Rough Parameter-based Imputation

III. Predicting the Missing Values According to the Proposed Algorithms

In this paper, the studied datasets including missing values are partitioned into two parts: complete records (all of the features of an object have been filled by real values) and incomplete (one or more features of an object have missing values) objects. Fig.7 illustrates the missing value imputation procedure in this paper.

In Fig.7, the missing values are filled separately based on four algorithms in terms of Fuzzy C-Means parameter-based imputation, Fuzzy C-Means Rough parameter-based imputation, Jointly Fuzzy C-Means and VQNN Imputation, and Jointly Fuzzy C-Means and Fitted VQNN Imputation. In order to obtain the optimal prediction accuracy, the parameters of optimal cluster and weighting factors in fuzzy C-means clustering should be determined by multiple iterations. The parameters that including evaluation measures for complete dataset and classification accuracy for incomplete dataset are adopted to judge the efficiency of filling. Where, the evaluation measures for complete dataset is used to judge the variation between the filled value and the real value and the classification accuracy is computed by using the filled complete dataset. The imputation procedure is finished until the optimal filling values are achieved.

Next, we will give a clear description of the proposed algorithms of Jointly Fuzzy C-Means and VQNN Imputation and Jointly Fuzzy C-Means and Fitted VQNN Imputation.

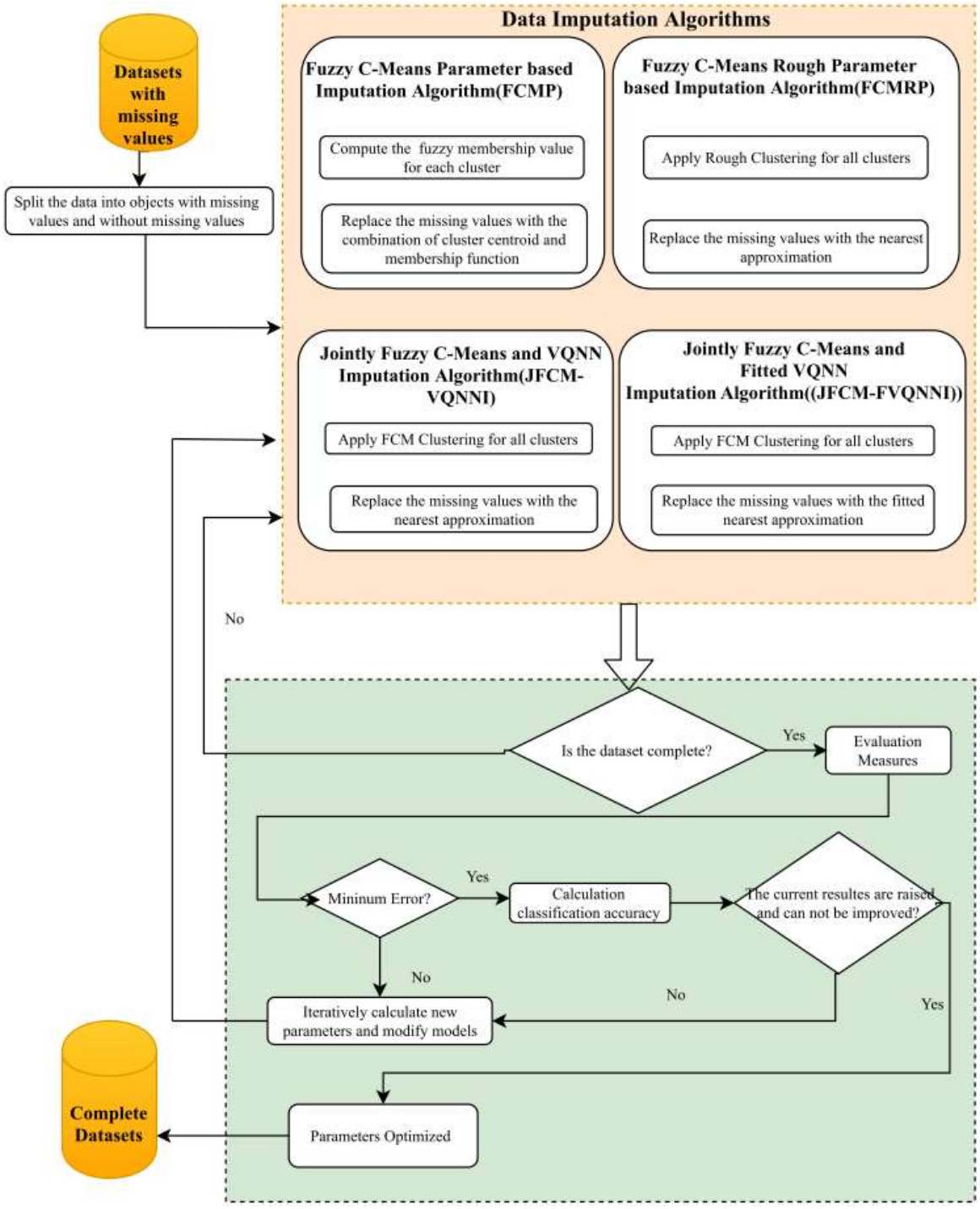

Fig.7. Overview of missing value imputation procedure

A. Jointly Fuzzy C-Means and VQNN Imputation Algorithm

1) Problem Definition

Definition 8 [25]. The incomplete information system (IIS) $<U, A>$ is occurred when some of the values of objects' attributes are missing. The unknown value is expressed by using the symbol "*". For instance, if $f(x, a_i) = *, (x \in U, a_i \in A)$, then the value of object x on attribute a_i is unknown.

The unknown (or missing) value can be obtained based on the comparable analysis with the objects covering all attributes or the defined values in the domain of the related attributes.

Definition 9. For all objects x in $U(\forall x \in U)$, the missing attributes set (*MAS*) is defined based on the following function:

$$MAS(x) = \{\forall a_i \in A \mid a_i(x) = *\} \quad (18)$$

Definition 10. The missing object set (*MOS*) is consisted by objects with at least one missing value on its attributes. The missing object set (*MOS*) is defined based on the following function:

$$MOS = \{\forall x \in U \mid MAS(x) \neq \varnothing\} \quad (19)$$

Definition 11. The complete object set (COS) is consisted by objects without missing value on all attributes. The complete object set (COS) is defined based on the following function:

$$\mathrm{COS} = \{\forall x \in U \mid MAS(x) = \varnothing\} \quad (20)$$

Based on the previous research, toleration and similarity relations are proposed to handle missing values [25]. These theories have been propagated and applied in many research papers [26], [27].

The imputation of missing data is an important task in data processing. This problem can be considered as uncertainty and vagueness analysis. Fuzzy rough set (FRS) has been widely used due to its ability to deal with real-valued data with vague information [28]. Besides, the approaches based on FRS have shown high performance for dealing with uncertainty data and possess excellent properties for robustness and noise tolerance. In this paper, the perspective of FRS is adopted for missing data imputation in IIS.

2) Fuzzy-Rough Set Theory

An FRS is the pair of lower and upper approximations of a fuzzy set A in a universe U on which a fuzzy relation R is defined. The FRS is given based on fuzzifying the definitions of the crisp lower and upper approximation.

Given a fuzzy tolerance relation R and a fuzzy set A in U, the lower and upper approximation of A by R can be constructed in several ways. Suppose ℓ is a fuzzy implication, $\ell(R(x,y), A(y))$ is used to express the extent of an element that is similar to x belongs to A. The lower approximation is defined in the following:

$$(R \downarrow A)(x) = \inf_{y \in U} \ell(R(x,y), A(y)) \quad (21)$$

where the lower approximation is high for the membership value of an element $x \in U$ if these values $\ell(R(x,y), A(y))$ are high for all $y \in U$. Suppose $\mathcal{T}$ is a t-norm, $\mathcal{T}(R(x,y), A(y))$ is used to express to what extent there exist instances that are similar to x and belong to A. The upper approximation is defined as follows:

$$(R \uparrow A)(x) = \sup_{y \in U} \mathcal{T}(R(x,y), A(y)) \quad (22)$$

Fuzzy relation $R(x,y)$ is very important in defining the fuzzy similarity degree among objects with respect to the corresponding attributes. This work is a continuous research of previous published papers [29], [30]. We still follow the same defined fuzzy relation $R(x,y)$. Given a set of attributes C, a common method to construct $R(x,y)$ is shown as follows:

$$R(x,y) = min_{a \in C} R_a(x,y) \quad (23)$$

where $R_a(x,y)$ is the degree to which instances *x* and *y* are similar for attribute a. The definition of $R_a(x,y)$ is given in the following equation:

$$R_a(x,y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} \quad (24)$$

where a_{min} and $a_{\max}$ are the minimal and maximal value of attribute a, respectively.

3) The Proposed Jointly Fuzzy C-Means and VQNN Imputation Algorithm

On the basis of clustering conception, the similarity relations among instances to predict the missing feature values are also adopted in paper due to its capability to handle MVI task [26], [27], [31], [32], [33], [34], [35]. Indistinguishable matrixes and tolerance relations are also used to identify and investigate the potential closest replacement values based on its similarity objects.

Based on similarity relation principle, **F**uzzy-**R**ough **N**earest **N**eighbor **I**mputation algorithms called FRNNI (implicator/t-norm based fuzzy-rough sets) and VQNNI (**V**aguely **Q**uantified rough set based **N**earest **N**eighbor **I**mputation algorithm) (or called FRNN-VQS in some research work) [26] have been proposed. VQNNI is an extension of FRNNI. The experimental evaluation demonstrates that VQS is superior to FRS in handling noise and vagueness datasets in most cases [36]. Therefore, in this paper, we focus on the extension method of VQNNI.

In order to better handle noise issue by comparing with FRNNI, the VQNNI adopts $R \downarrow^{Q_u} C$ and $R \uparrow^{Q_l} C$ to replace $R \downarrow C$ and $R \uparrow C$, respectively. Where, (Q_u, Q_l) is a pair of fuzzy quantifiers that model 'most' and 'some'. In VQNNI, we adopt the default quantifiers $Q_l = Q_{(0.1,0.6)}$ and $Q_u = Q_{(0.2,1.0)}$, which is based on the following Eq. (25).

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \dfrac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \dfrac{\alpha+\beta}{2} \\ 1 - \dfrac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \dfrac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} \quad (25)$$

Algorithm 1: **J**ointly **F**uzzy **C**-**M**eans and **VQNN I**mputation (JFCM-VQNNI) Algorithm

Input: U , nonempty finite set of objects; MOS , missing object set; *COS*, complete object set;
k , nearest neighbor; *C*, the number of clusters

Output: The new dataset U with the imputed values.

Begin
for object $\forall y \in \mathrm{COS}$
Apply Fuzzy C-Means clustering algorithm to COS
Calculate the centroid value for clusters based on Eq. (2)
Figure out membership functions of MOS for all clusters based on Eq. (3).
endfor
for object $\forall y \in MOS$
$N \leftarrow \mathrm{getNearestNeighbors}(y,k)\ in\ cluster\ C_i$
while $MAS(y) \neq \varnothing$ **do**
$\tau_1 \leftarrow 0, \tau_2 \leftarrow 0$
$\forall z \in N$, compute fuzzy similarity relation $R(z,y)$
for $z \in N$ **do**
$M \leftarrow \dfrac{((R \downarrow R_a z)(y) + (R \uparrow R_a z)(y))}{2}$
$\tau_1 \leftarrow \tau_1 + M \times a(z)$
$\tau_2 \leftarrow \tau_2 + M$
endfor
if $\tau_2 > 0$
$MAS(y) \leftarrow ((\tau_1 / \tau_2) + \sum_{i=1}^{C} u(y, cluster_i) \times m_{cluster_i}) \Big/ 2$
else
$MAS(y) \leftarrow (\dfrac{\sum a(z)}{|N|} + \sum_{i=1}^{C} u(y, cluster_i) \times m_{cluster_i})) \Big/ 2$
endif
endwhile
endfor
end

As for MVI task, we still follow the same idea that every missing value is considered as a decision feature and then the prediction is generated for every object containing at least one missing value. The proposed **J**ointly **F**uzzy **C**-**M**eans and **VQNN I**mputation (JFCM-VQNNI) algorithm are given in Algorithm 1. Firstly, JFCM-VQNNI clusters the complete data objects into *C* clusters based on FCM algorithm. Next, for any missing object *y*, the JFCM-VQNNI finds its *k* nearest neighbors in its corresponding cluster C_i and puts these points into set *N*. Then, the lower and upper approximations of object *y* w.r.t the object *z* that is used to produce the membership *M*. This process is conducted for all of the *k* nearest neighbors of *y*. Finally, the missing attribute value is assigned based on the calculation of membership *M* and the summation of the centroid value and the corresponding membership values. In order to deal with the rare case where the denominator is 0, the mean of feature values for all *k* nearest neighbors and the modified fuzzy membership value regarding all clusters are adopted to assign the missing value.

B. The Proposed Fuzzy C-Means and Fitted VQNN Imputation Algorithm

The idea that the nearest neighbor of the target object is treated as a rule and then the strategies related to how to rational assign the activation degree of the lower and upper approximations of the selected test objects is further studied in this paper. In fact, most of the values assigned in the decision attributes are correct and rarely miss. This information can be captured and applied to give more rational degree of the similarity objects. Therefore, a fitted fuzzy-rough method has been proposed to adjust the obtained weights for the nearest neighbor objects.

Definition 12. The fuzzy similarity class $[x]_{R_a}$ is associated with object *x* and R_a, which is the fuzzy neighborhood of object *x*. For instance, $[x]_{R_a}(y) = R_a(x, y), y \in U$.

Suppose the decision attribute set D on sample set U can be partitioned into r equivalence classes, which is $U / D = \{D_1, D_2, \cdots, D_r\}$.

Definition 13. The fuzzy set D in U is defined as $\{\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_r\}$, which is a family of fuzzy set on U and means fuzzy decision of objects induced by decision attribute set D and conditional attribute *C*. If the relation $\sum_{i=1}^{r} \tilde{D}_i(x) = 1, \ \forall x \in U$ is satisfied, and then the $\{\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_r\}$ is defined as fuzzy partition and $\tilde{D}_i(x)$ means the membership degree of the object *x* to D_i or the fuzzy decision of object *x*. The method of calculation $\tilde{D}_i(x)$ is shown as follows:

$$\tilde{D}_i(x) = \frac{\left|[x]_{R_a} \cap D_i\right|}{[x]_{R_a}}, i = 1, 2, ..., r, \forall x \in U \tag{26}$$

The classical way of calculating the lower and upper approximation of the corresponding decision of objects is shown as follows:

$$\underline{R_a}(D_i)(d) = \min_{x \in U} \max\{1 - R_a(x, d), \tilde{D}_i(x)\}, d \in \tilde{D}_i \tag{27}$$

$$\overline{R_a}(D_i)(d) = \max_{x \in U} \min\{R_a(x, d), \tilde{D}_i(x)\}, d \in \tilde{D}_i \tag{28}$$

It should aware that if D_i is a fuzzy set and then Eq. (27) and Eq. (28) belong to the same type with Eq. (21) and Eq. (22), respectively.

The drawbacks of classical fuzzy rough set can conclude in two aspects including it cannot ideally portray the differences of the object classifications and also the membership function cannot fit the dataset well. We give *Example 1* to clarify this situation.

Example 1. Given a decision table $< U, C, D >$ (Table I), $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, conditional attribute set $C = \{a_1, a_2, a_3, a_4\}$, and $D = \{d\}$ is the decision attribute set.

TABLE I
AN EXAMPLE DECISION TABLE

U	***a₁***	***a₂***	***a₃***	***a₄***	***d***
x₁	2	6	6	7	1
x₂	5	2	3	3	1
x₃	6	3	5	9	2
x₄	6	8	4	6	2
x₅	7	8	3	5	3
x₆	8	6	4	7	3

The fuzzy similarity relations between two objects with respect to the corresponding attributes are calculate based on Eq. (24), which is shown as follows:

$$R(x, y) = \begin{bmatrix} 1 & 0 & 0.3333 & 0.3333 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.3333 \\ 0.3333 & 0 & 1 & 0.1667 & 0.1667 & 0.5000 \\ 0.3333 & 0 & 0.1667 & 1 & 0.6667 & 0.6667 \\ 0 & 0 & 0.1667 & 0.6667 & 1 & 0.6667 \\ 0 & 0.3333 & 0.5000 & 0.6667 & 0.6667 & 1 \end{bmatrix}$$

The decision attribute set D splits the objects into three parts, which is $U / D = \{\tilde{D}_1, \tilde{D}_2, \tilde{D}_3\}$, $\tilde{D}_1 = \{x_1, x_2\}$, $\tilde{D}_2 = \{x_3, x_4\}$, $\tilde{D}_3 = \{x_5, x_6\}$. Meanwhile, the fuzzy decision of object *x* can be calculated based on Eq. (26). The fuzzy decision matrix of the objects in Table I are shown as follows:

$$\tilde{D} = [\tilde{D}_1, \tilde{D}_2, \tilde{D}_3] = \begin{bmatrix} 0.6000 & 0.4000 & 0 \\ 0.7500 & 0 & 0.2500 \\ 0.1538 & 0.5385 & 0.3077 \\ 0.1176 & 0.4118 & 0.4706 \\ 0 & 0.3333 & 0.6667 \\ 0.1053 & 0.3684 & 0.5263 \end{bmatrix}$$

Based on Eq. (27) and Eq. (28), we can get the results of the lower and upper approximation, which is described in Table II.

Based on Table I, we can see that the decision value of object x_4 is '2', and yet, the object x_4 should belong to the third category according to the results in Table II. Therefore, the lower and upper approximations of decision attribute set and the maximum membership should be updated to improve the classification accuracy.

In order to guarantee the membership of an object to its own category in a maximal degree and make the fuzzy rough set more suitable and original reflects the diverse interval of dataset, we have studied multiple-granularity fuzzy-rough set in this paper.

TABLE II
CALUATED LOWER AND UPPER APPROXIMATION OF THE DECISION TABLE

U	$\underline{R}(\tilde{D}_1)$	$\overline{R}(\tilde{D}_1)$	$\underline{R}(\tilde{D}_2)$	$\overline{R}(\tilde{D}_2)$	$\underline{R}(\tilde{D}_3)$	$\overline{R}(\tilde{D}_3)$
x_1	0.6	0.6	0.4	0.4	0	0.3333
x_2	0.6667	0.75	0	0.3333	0.25	0.3333
x_3	0.1538	0.3333	0.5	0.5385	0.3077	0.5
x_4	0.1176	0.3333	0.3333	0.4118	0.4706	0.6667
x_5	0	0.1538	0.3333	0.4118	0.4706	0.6667
x_6	0.1053	0.3333	0.3333	0.5	0.4706	0.6667

Most of the current research works employ fuzzy/ rough similarity relations to analyze missing value

imputation only at one level of granularity. However, more practical classification information is obtained when similarity relations function is built at different levels of granularity. In order to better control the noise in sample dataset, the fitted fuzzy neighborhood of object *x* is shown as follows:

$$[x]^{\beta}_{R_a}(y) = R^{\beta}_a(x, y) \quad (29)$$

The parameter β is consided as weight coefficient that can enlarge or reduce the value of $R_a(x,y)$ in different degrees according to actual requirements. Thus, the membership degree of a fuzzy similarity relation is influenced by weight coefficient β and relations with respect to attributes $R_a(x,y)$. The studied fuzzy similarity relation on U is denoted as $R^{\beta}_a(x,y)$ in this paper. Therefore, the fitted lower and upper approximations of decision attribute set D regarding attributes C are defined.

$$\underline{R^{\beta}_a}(D_i)(d) = \min_{x\in U} \max\{1 - R^{\beta}_a(x,d), \tilde{D}^{\alpha}_i(x)\}, d \in \tilde{D}_i \quad (30)$$

$$\overline{R^{\beta}_a}(D_i)(d) = \max_{x\in U} \min\{R^{\beta}_a(x,d), \tilde{D}^{\alpha}_i(x)\}, d \in \tilde{D}_i \quad (31)$$

Where, α is a parameter that used to control the fuzzy decision of objects. It considered as a weight coefficient that is adopted to adjust fuzzy decision of an object according to the accurate fuzzy membership.

For the dataset in *Example 1*, object x_3 and object x_4 have the same category. Set $\beta = 5$, the updated $R^{\beta}(x,y)$ is shown as follows:

$$R^{\beta}(x,y) = \begin{bmatrix} 1 & 0 & 0.3333 & 0.3333 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.3333 \\ 0.3333 & 0 & 1 & 0.1667 & 0.1667 & 0.5000 \\ 0.3333 & 0 & 0.8333 & 1 & 0.6667 & 0.6667 \\ 0 & 0 & 0.1667 & 0.6667 & 1 & 0.6667 \\ 0 & 0.3333 & 0.5000 & 0.6667 & 0.6667 & 1 \end{bmatrix}.$$

Set $\alpha = 1.5$, the updated decision attribute is obtained based on the value of α and the modified $R^{\beta}(x,y)$, as follows:

$$\tilde{D} = [\tilde{D}_1, \tilde{D}_2, \tilde{D}_3] = \begin{bmatrix} 0.6000 & 0.4000 & 0 \\ 0.7500 & 0 & 0.2500 \\ 0.1538 & 0.5385 & 0.3077 \\ 0.0952 & 0.7857 & 0.3810 \\ 0 & 0.3333 & 0.6667 \\ 0.1053 & 0.3684 & 0.5263 \end{bmatrix}$$

The modified lower approximation of object x_4 is $\underline{R}(x_4) = [\underline{R}(\tilde{D}_1), \underline{R}(\tilde{D}_2), \underline{R}(\tilde{D}_3)] = [0.0952\ 0.3333\ 0.3077]$, and the modified upper approximation of object x_4 is $\overline{R}(x_4) = [\overline{R}(\tilde{D}_1), \overline{R}(\tilde{D}_2), \overline{R}(\tilde{D}_3)] = [0.3333\ 0.7857\ 0.6667]$ The maximal membership degree of object x_4 is $< \underline{R}(\tilde{D}_2), \overline{R}(\tilde{D}_2) >$, thus the decision of object x_4 belongs to group '2'.

Next, the proposed **J**ointly **F**uzzy **C**-**M**eans and **F**itted **VQNN I**mputation (JFCM-FVQNNI) Algorithm is given in Algorithm 2. Similarly, the JFCM-FVQNNI algorithm treats the missing value as a unknown decision attribute and then predict the corresponding value based on classcification training results. The decision attribute contains more valuable imformation for grouping objects. More important weight has been assigned based on membership analysis of lower and upper approximations of decision attribute set.

The running mechanism of Algorithm 2 is detailed. The dataset is firstly separated into complete and incompete part. FCM clustering algorithm is applied into complete part. As for the object *y* contains at least one missing value for all attributes, JFCM-FVQNNI algorithm finds its *k* nearest neighbors and storage them into set N . In order to obtain more accurate lower and upper approximations of object *y* by comparing with JFCM-VQNNI algorithm, the modification weight coefficent α and β are adopted to change fuzzy neighborood of decision set *D*. Next, the adjustment weights of fuzzy membership of all neighbors are given based on $\underline{R^{\beta}_a}(D_i)(z)$, $\overline{R^{\beta}_a}(D_i)(z)$ and $\tilde{D}(z)$, where $z \in N$. And then, as for the expected calulcated missing value in object *y*, the information obtained from its neighbor *z* is used to predict the final membership (M). Finally, the final missing value is assigned based on final membership and the summation of the centroid value with its corresponding membership. The whole process iteratively execution for all neighbors. Thus, the obtained value depends on the existing values of all neighbors and the centroid values of all clusters. It is possible, though unlikely, that $\tau_2 = 0$. In this case, τ_1 / τ_2 cannot be calculated and wrong value will be assigned. To handle this issue, the average value of the attributes for the neighbors and the sum of the centroid value with its corresponding fuzzy membership is adopted.

Algorithm 2: **J**ointly **F**uzzy **C**-**M**eans and **F**itted **VQNN I**mputation (JFCM-FVQNNI) Algorithm

Input: U , nonempty finite set of objects; *MOS* , missing object set; *COS*, complete object set; k , nearest neighbor; *C*, the number of clusters; α , weight coefficient for fuzzy decision adjustment; β , weight coefficient for $R_a(x,y)$.

Output: The new dataset U with the imputed values.
Begin

for object $\forall y \in \mathrm{COS}$
Apply Fuzzy C-Means clustering algorithm to COS
Calculate the centroid value for clusters based on Eq. (2)
Figure out membership functions of MOS for all clusters based on Eq. (3).
e**ndfor**
for object $\forall y \in MOS$
$N \leftarrow \mathrm{getNearestNeighbors}(y,k)\ in\ cluster\ C_i$
while $MAS(y) \neq \varnothing$ **do**
$\tau_1 \leftarrow 0, \tau_2 \leftarrow 0$
$\forall z \in N$, compute fuzzy similarity relation $R_a(z,y)$
$\forall z \in N$, compute fuzzy decision $\tilde{D}(z) = [\tilde{D}_1, \tilde{D}_2, \cdots, \tilde{D}_r]$
$\forall z \in N$, compute $\underline{R_a^{\beta}}(D_i)(z)$ and $\overline{R_a^{\beta}}(D_i)(z)$
Compute the modified percentage μ based on $\underline{R_a^{\beta}}(D_i)(z)$, $\overline{R_a^{\beta}}(D_i)(z)$, and $\tilde{D}(z)$
for $z \in N$ **do**
$M \leftarrow \dfrac{((R \downarrow R_a z)(y) + (R \uparrow R_a z)(y))}{2} \times \mu$
$\tau_1 \leftarrow \tau_1 + M \times a(z)$
$\tau_2 \leftarrow \tau_2 + M$
endfor
if $\tau_2 > 0$
$$MAS(y) \leftarrow ((\tau_1 / \tau_2) + \sum_{i=1}^{C} u(MOS, cluster_i) \times m_{cluster_i}) \Big/ 2$$
else
$$MAS(y) \leftarrow (\frac{\sum a(z)}{|N|} + \sum_{i=1}^{C} u(MOS, cluster_i) \times m_{cluster_i})) \Big/ 2$$
endif
endwhile
endfor
end

We give the time complexity analysis of the proposed Algorithm 1 (JFCM-VQNNI) and Algorithm 2 (JFCM-FVQNNI). Suppose the dataset with some missing value has m objects, n features, and k nearest neighbor. And suppose the algorithm divides the dataset into c clusters and algorithm iterated p times, and then the total complexity of JFCM-VQNNI and JFCM-FVQNNI is $O(mn^2k^2) + O(m^2n) + O(mcp)$. The added time complexity by comparing with orginal VQNNI is $O(mn^2k^2) + O(mcp)$. Theoretically, the bigger of parameter k and the higer accuracy, howerever, the larger of time complexity. Meanwhile, when k reaches to a certain level, the accuracy does not increase any more. Thus, the time complexity shows that selecting sensible parameter k is very imporant. Meanwhile, the parameters of cluster number and iteration number are also very important and reasonable values should be given in experiment. It should be pointed out that the running time of JFCM-FVQNNI is higher than that of JFCM-VQNNI Algorithm.

IV. EXPERMENTAL PERFORMANCE

In order to illustrate the imputation performance of the proposed JFCM-VQNNI and JFCM-FVQNNI algorithms, experiments have been done in this section. Firstly, we compare the proposed JFCM-VQNNI and JFCM-FVQNNI with FCMP and FCMRP for complete datasets in three aspects of actual and the filled values comparation, imputation efficiency judgment based on the evaluation measures of RMSE and MAE, and the classification analysis for the imputed datasets based on

different degree of missing rate. And then, the classification results for the incomplete datasets have been further analyzed to evaluate the capability of handling missing values of the proposed algorithms.

A. Experimental Datasets and Environment

The experiments use five benchmark datasets in terms of two complete datasets (Yeast and Gene expression cancer RNA-Seq (Short for Gene expression)) and three incomplete datasets (Cleveland, Pima, and Mice protein expression) to explore the performance of the proposed algorithms. These datasets have between 303 and 1484 objects with the number of attributes ranging from 9 to 20532 as shown below in Table III. In Table III, the column labeled as ‘‘% MV’’ indicates the percentage of all values of the data set which are missing. The column labeled as ‘‘%Ins. with MV’’ refers to the percentage of instances in the data set which have at least one MV. In order to create the missing dataset for complete dataset, part of the dataset is deleted randomly. Randomly made missing ratios are 1%, 3%, 5%, 7%, and 9% for complete dataset.

Stratified 10×10 -Fold Cross Validation ($10\times10-FCV$) is employed for all datasets for all experiments. All the experiments were evaluated in a PC computer by running in Windows 10 system, with a 3.4 GHz Intel Core (TM) i7-6700 CPU, 16GB RAM, and a 2TGB hard disk. The programs were written in Matlab.

TABLE III
DATA SETS USED FOR EXPERIMENTATION

Data set	Ins.	Attribute	Class	% MV	%Ins. with MV
Cleveland	303	14	5	0.14	1.98
Pima	768	9	2	11.04	56.25
Mice protein expression	1080	81	8	1.59	48.89
Yeast	1484	9	10	-	-
Gene expression	801	20532	5	-	-

B. Evaluation Measures

The studied datasets contain complete dataset and incomplete dataset. As for the complete dataset, the imputation efficiency of the algorithms of FCMP, FCMRP, JFCM-VQNNI, and JFCM-FVQNNI are judged by the evaluation measures of root mean squared error (RMSE) and mean absolute error (MAE).

The RMSE is defined in the following equation:

$$RMSE = (\frac{1}{nm}\sum_{i=1}^{nm}(AC_i - IM)^2)^{\frac{1}{2}} \tag{32}$$

And the MAE is defined in the following equation:

$$MAE = \frac{1}{nm}\sum_{i=1}^{nm}\left|AC_i - IM_i\right| \tag{33}$$

Where, the parameter ‘nm’ means the number of missing values, the parameter ‘AC_i’ represents the actual value of the *ith* original value, and ‘IM_i’ indicates the imputed value for the *ith* original value.

The range of the evaluation measures of RMSE and MAE varies from 0 to ∞. Meanwhile, the small of the measures RMSE and MAE, and the better performance of the algorithms.

C. Experimental Performance

1) Parameters Set-up

Initially, the parameter cluster number should be given in cluster imputation algorithms. If the input cluster number is small, the data objects will be gathered in specific clusters. In some cases, some objects with larger difference may be constrained to be in one cluster, which will decrease the imputation accuracy. In contrast, if the input cluster number is big, and then the data objects will be dispersed and even some object will detach from the main points. Thus, the number of clusters should be given reasonable for missing objects imputation. Since the decision attribute in the dataset is objectively classified the entire data set into several categories based on practical meaning, we consider the number of classifications in decision attribute is the number of clusters.

Secondly, the nearest neighbor number is an important parameter in the proposed algorithms of JFCM-VQNNI and JFCM-FVQNNI. We have conducted four times of all algorithms with setting $k=[5:5:60]$ in the first three times and setting k to the full set of training instances in the fourth time for all datasets. A large number of experiments show the experimental results are more significant and stable when $k=[20:5:40]$, at the same time, the running time will increase with the increase of value *k*, while the imputation efficiency does not increase significantly. Therefore, we select $k=20$ to illustrate the classification results.

The weight coefficients of α and β in the proposed algorithm JFCM-FVQNNI aim to modify the decision membership for each object with respect to each class and the related fuzzy similarity relation. It is very difficult to assign the exactly range which is capable to obtain the maximum accuracy. The best situation tends to be obtained in practice and random. Since this paper aims to verify whether the introduce of more correct classification results will affect the performance of MVI or not, we did not find the optimal situation for each studied dataset, and we adopt the common value 1 for parameters of α and β.

2) Experimental Performance with Complete Dataset

The evaluation of the experimental performance for complete datasets consists of three aspects. Firstly, the actual and the filled values by the algorithms of FCMP,

FCMRP, JFCM-VQNNI, and JFCM-FVQNNI are compared for two complete datasets, which is shown in Table IV and Table V. Secondly, the efficiency of imputation algorithms for complete dataset is judged by RMSE and MAE. The comparison results are presented in Fig.8 and Fig.9. Thirdly, the classification results are analyzed for the filled dataset of the missing rate of 1%, 3%, 5%, 7%, and 9%, respectively. The results are shown in Fig.10 and Fig.11.

First of all, the comparison results between actual values and the predicted values for Yeast Dataset are shown in Table II. The 'actual place' means the original real data position. For instance, the actual value of position with row 167 and column 3 is '0.42' in the original dataset. In order to evaluate the experimental performance, this data is deleted based on randomly selection in 1% missing rate. The predicted value of FCMP, FCMRP, JFCM-VQNNI, and JFCM-FVQNNI is '0.3596', '0.3640', '0.3638', and '0.3739', respectively. In Table II, five points are randomly selected for each percentage of missing rate. Based on the presented total number of 25 points for missing rate of 1%, 3%, 5%, 7%, and 9%, the results have shown that 16 points of the proposed JFCM-FVQNNI are closer to the true value and 9 points of the proposed JFCM-VQNNI are closer to the original value. However, only 4 points for FCMP algorithm and 3 points for FCMRP algorithm are superior to other algorithms.

TABLE IV
ACTUAL AND PREDICTED VALUES COMPARISION AMONG IMPUTATION ALGORITHMS: YEAST DATASET

Missing Rate (%)	Actual place		Actual value	Imputation Algorithms			
	Row	Column		FCMP	FCMRP	JFCM-VQNNI	JFCM-FVQNNI
1	167	3	0.42	0.3596	0.3640	0.3638	**0.3739**
	231	1	0.41	0.4564	0.4493	0.4627	**0.4350**
	610	1	0.6	0.5193	0.5298	0.5353	**0.5410**
	623	4	0.44	0.3677	0.3738	0.4070	**0.4580**
	709	8	0.32	0.2381	0.2321	0.2373	**0.2487**
3	21	1	0.45	0.4390	0.4714	**0.4640**	0.4676
	55	2	0.59	0.4708	0.4817	**0.5300**	**0.5300**
	141	4	0.16	0.2347	0.2229	0.2134	**0.2052**
	151	5	0.5	0.4991	0.5011	**0.5000**	**0.5000**
	221	3	0.54	**0.5075**	0.4680	0.4771	0.4645
5	44	2	0.54	0.4408	0.4876	**0.4903**	**0.4903**
	66	1	0.6	**0.5020**	0.4827	0.4929	0.4974
	144	1	0.37	0.4790	**0.4550**	0.4852	0.4778
	144	2	0.46	0.4337	0.4489	**0.4678**	**0.4678**
	150	3	0.5	**0.5108**	0.5280	0.5362	0.5323
7	1	2	0.61	0.5306	0.5136	0.5657	**0.5697**
	6	3	0.56	0.5208	0.5239	0.5429	**0.5448**
	6	7	0.49	0.4785	0.4949	**0.4886**	0.4981
	17	5	0.5	0.5108	0.5040	**0.5000**	**0.5000**
	27	3	0.53	0.5244	**0.5302**	0.5090	0.5157
9	6	2	0.4	**0.4785**	0.4816	0.4852	0.4868
	22	1	0.43	0.4980	**0.4903**	0.5032	0.5156
	26	7	0.54	0.5119	0.5013	0.5105	**0.5284**
	34	1	0.33	0.4559	0.4568	**0.4207**	**0.4207**
	41	6	0	0.0005	0.0009	**0.0000**	**0.0000**

TABLE V
ACTUAL AND PREDICTED VALUES COMPARISION AMONG IMPUTATION ALGORITHMS: GENE EXPRESSION CANCER RNA-SEQ DATASET

Missing Rate (%)	Actual place		Actual value	Imputation Algorithms			
	Row	Column		FCMP	FCMRP	JFCM-VQNNI	JFCM-FVQNNI
1	16	26	0	0.5265	0.5232	0.1446	**0.0622**
	371	1	0	0.0266	0.0268	**0**	**0**
	371	8	0.3977	0.4999	0.4912	0.2910	**0.3395**
	467	2	4.2257	3.0109	3.1214	**5.0195**	**5.0195**
	467	147	8.6685	9.2480	9.3006	**8.7537**	8.8713
3	8	4	7.2267	8.4369	6.6838	**6.8185**	6.7948
	8	11	0.4418	**0**	0.6726	1.4806	1.1672
	181	3	3.6825	3.0953	3.1064	2.8255	**3.3629**
	181	7	7.6703	7.4055	7.3674	7.4851	**7.5148**
	212	2	2.4233	3.0109	3.0224	2.9737	**2.8826**
5	5	13	2.1412	2.6673	**2.6361**	2.8457	2.7246
	30	4	10.1295	6.7223	6.7678	6.9896	**7.1889**

	49	9	0.4348	0.0167	0.0129	0.0435	**0.0725**
	86	4	6.3041	6.7223	6.7565	6.6129	**6.5393**
	92	8	0.5541	0.4999	0.5001	0.4899	**0.5323**
7	6	3	3.5819	3.0953	3.0557	**3.7172**	4.2374
	23	5	9.1969	9.8136	9.6963	9.6340	**9.4863**
	25	4	6.0093	6.7223	6.6251	6.2769	**6.2014**
	26	2	4.3177	3.0109	2.9649	3.1935	**3.4302**
	41	3	2.2939	3.0953	3.0064	2.4487	**2.4212**
9	7	3	1.6912	3.0953	2.9082	2.5894	**2.4875**
	12	17	0	0.0030	0.0010	**0**	**0**
	13	3	2.3643	3.0953	**3.0500**	3.6961	4.0580
	58	11	0	0.6882	0.4981	0.4148	**0.0921**
	161	15	0	0.2146	0.2161	0.2178	**0.0813**

In addition, the measures of RMSE and MAE for the algorithms of FCMP, FCMRP, JFCM-VQNNI, and JFCM-FVQNNI are further discussed in Fig. 8. As for the judgement of RMSE, Fig.8 shows that FCMP and FCMRP have more similar experimental performance and the proposed JFCM-VQNNI and JFCM-FVQNNI algorithm have more closer results. In general, the proposed JFCM-VQNNI and JFCM-FVQNNI algorithm obtain better performance by comparing with the algorithms of FCMP and FCMRP. Moreover, the proposed JFCM-FVQNNI algorithm achieved optimal solutions in all cases. Similarly, we can draw the same conclusion for the measurement indictor of MAE. But the difference is that the gap of four algorithms for the test indicator MAE is larger than the previous RMSE. Overall, the proposed JFCM-VQNNI and JFCM-FVQNNI algorithms are better than the algorithms of FCMP and FCMRP in MAE. And in most cases, the proposed JFCM-FVQNNI algorithm is superior than the proposed JFCM-VQNNI.

Similarly, we have conducted experiment comparison for Gene expression dataset. Since the dataset Gene expression has 20532 attributes that belong to a large-scale dataset, the processing of missing values imputation is more complicated than the other datasets. In general, for all missing rate from 1% to 9%, the ranking of the overall experimental performance of the algorithms (from high to low) is JFCM-FVQNNI, JFCM-VQNNI, FCMRP, and FCMP. In order to demonstrate the comparative effects of four algorithms, we have randomly pick up 5 points in every missing rate, the imputation comparison results are shown in Table II. In detail, the results in Table II have exemplified that 19 out of 25 missing items of the proposed JFCM-FVQNNI can obtain optimal missing imputation. In contrast, only 6 items, 2 items, and 1 item of JFCM-VQNNI, FCMRP, and FCMP are the optimal missing imputation, respectively.

Moreover, the experiment results of measurements of RMSE and MAE of Gene expression dataset for the algorithms of JFCM-FVQNNI, JFCM-VQNNI, FCMRP, and FCMP are shown in Fig.9. According to the overall results of two indicators, Fig.9 represents that the proposed JFCM-FVQNNI is better than the proposed JFCM-VQNNI, JFCM-VQNNI is better than FCMRP, and FCMRP is better than FCMP. Compared to the results of Yeast dataset in Fig.8, the proposed imputation algorithms JFCM-FVQNNI and JFCM-VQNNI have more obvious gaps with FCMRP and FCMP, which is verified that the proposed algorithms in this paper can obtain more advanced experimental performance in processing comparatively larger-scale datasets.

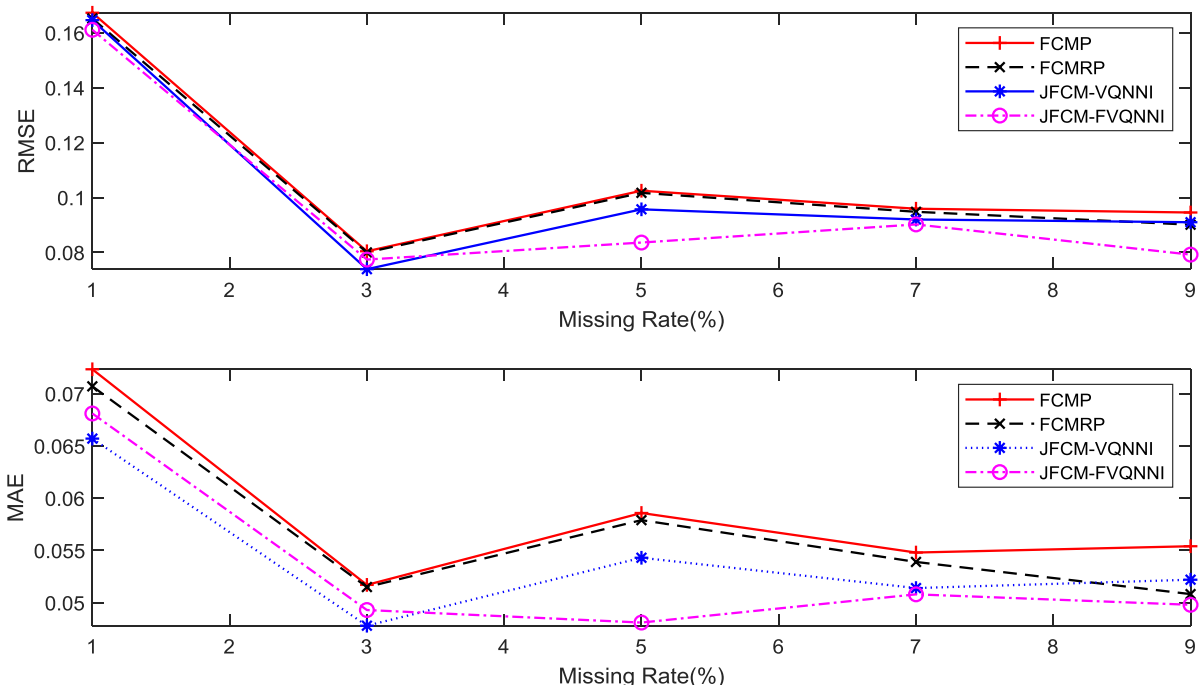

Fig.8. Performance comparison among imputation algorithms for complete dataset: Yeast

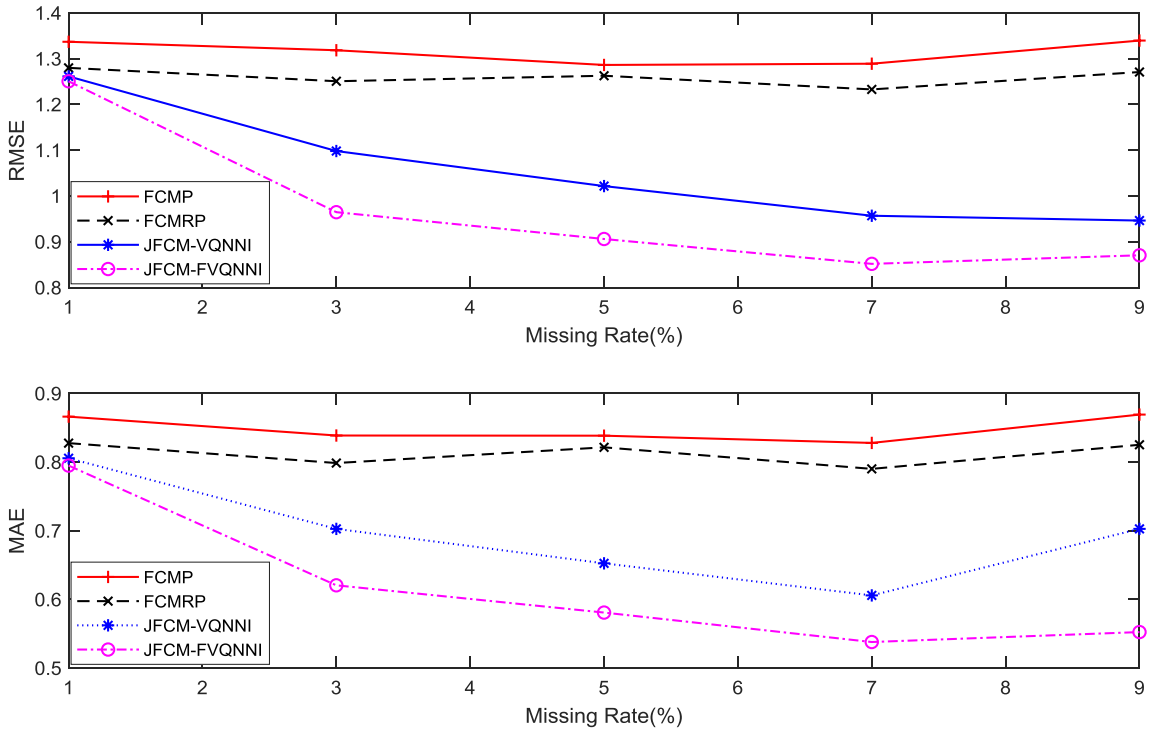

Fig.9. Performance comparison among imputation algorithms for complete dataset: Gene expression

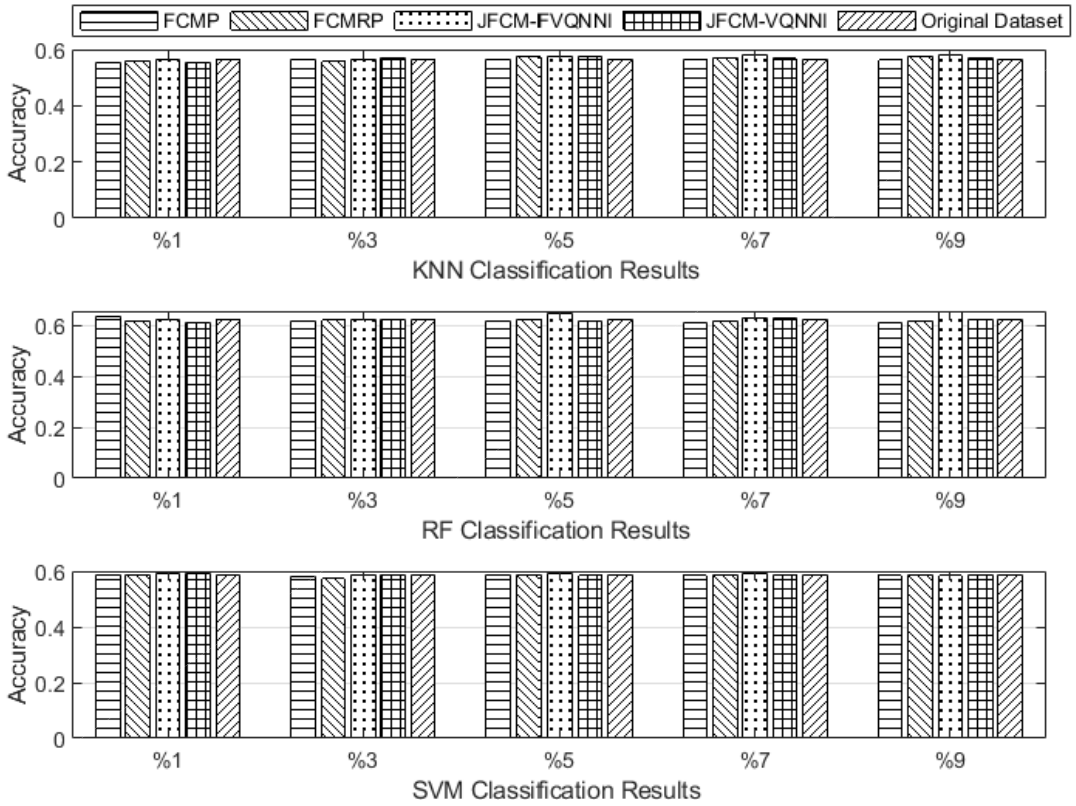

Fig. 10. Classification Results for Yeast Dataset Based on Different Types of Missing Rate

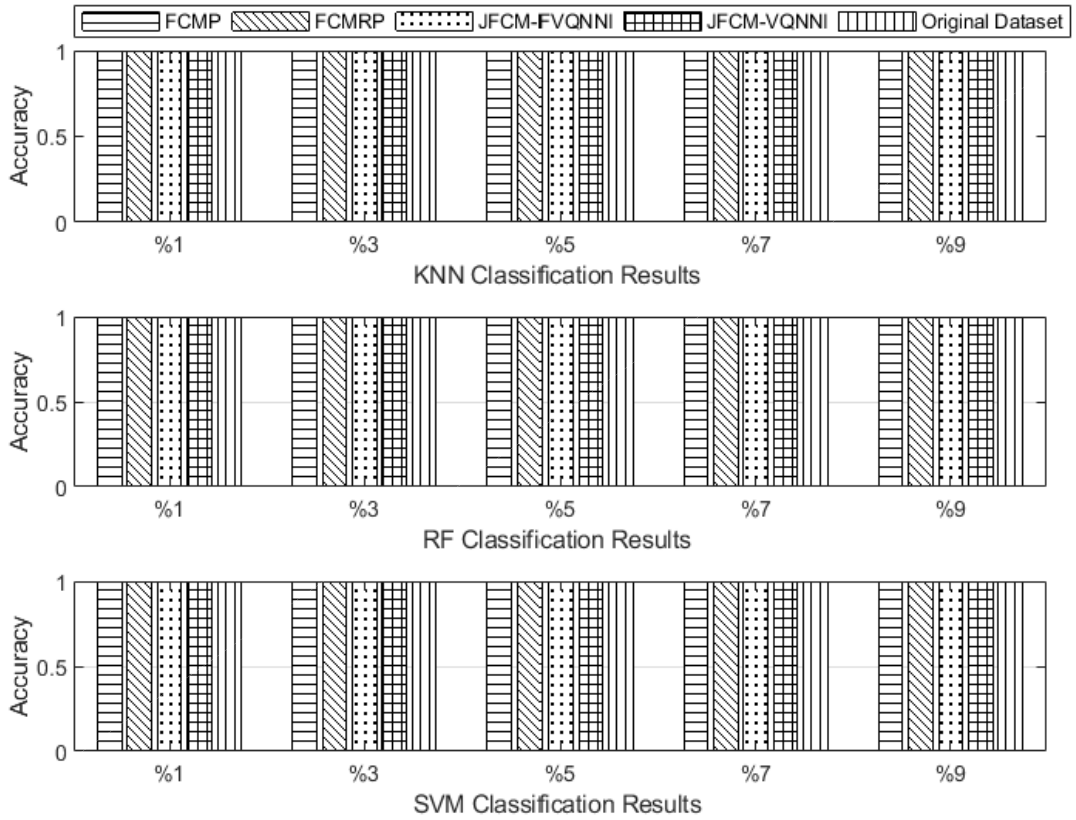

Fig.11. Classification Results for Gene Expression Dataset Based on Different Types of Missing Rate

Finally, the classification results of the filled dataset have been analyzed in Fig.10 and Fig.11. Due to the purpose of the experiment is to observe the effect of the filled data on classification, thus, we adopt the traditional classical classification algorithms named KNN, RF, and SVM. Based on a comparative analysis of original complete dataset, as for the Yeast dataset, from the overall classification results point of view, the overall effect on classification of all imputed dataset from 1% to 9% almost at the same level and the imputed datasets do not reduce the classification accuracy. It should be pointed out that the classification accuracy of the filled dataset by the proposed JFCM-FVQNNI algorithm has been improved by 1%-3% under all classification algorithms and different missing rate by comparing with the original complete dataset. This phenomenon may indicate that the original dataset itself has non-obvious missing values, and we will study it in the follow-up work. As for the large-scale dataset Gene expression, the accuracy of the dataset itself is relatively high, and the classification result based on KNN approach has reached to 99%, thus, the classification results of all algorithms converge to 1 in Fig.11. Similarly, on the whole, the effects of all classification results based on different algorithms on imputed dataset filled by different imputation methods are rarely distinguishable (close to 1). Compared with the Yeast dataset, the classification results of Gene expression are more indistinguishable. Part of the reason for this phenomenon is due to the fact that the classification result of this dataset is very high.

3) Experimental Performance with Incomplete Dataset

Three incomplete datasets in terms of Mice protein expression, Cleveland, and Pima have been analyzed in this paper. These datasets have been imputed by the algorithms of FCMP, FCMRP, JFCM-VQNNI, and JFCM-FVQNNI. We have conducted classification analysis and the results are shown in Fig.12, Fig.13, and Fig.14.

Firstly, the classification results of the imputed Mice protein expression are shown in Fig.12. Based on the classification results of original unfilled dataset in KNN, RF, and SVM, the accuracy rate is already relatively high. However, the classification effect in KNN and SVM of the filled dataset based on imputation algorithms has been slightly improved. In addition, the effect of JFCM-FVQNNI algorithm is more superior. Since the classification results of the imputation dataset under four different imputation algorithms in RF algorithm are all reach to 1, it is difficult to compare the filling efficiency of the imputation algorithms.

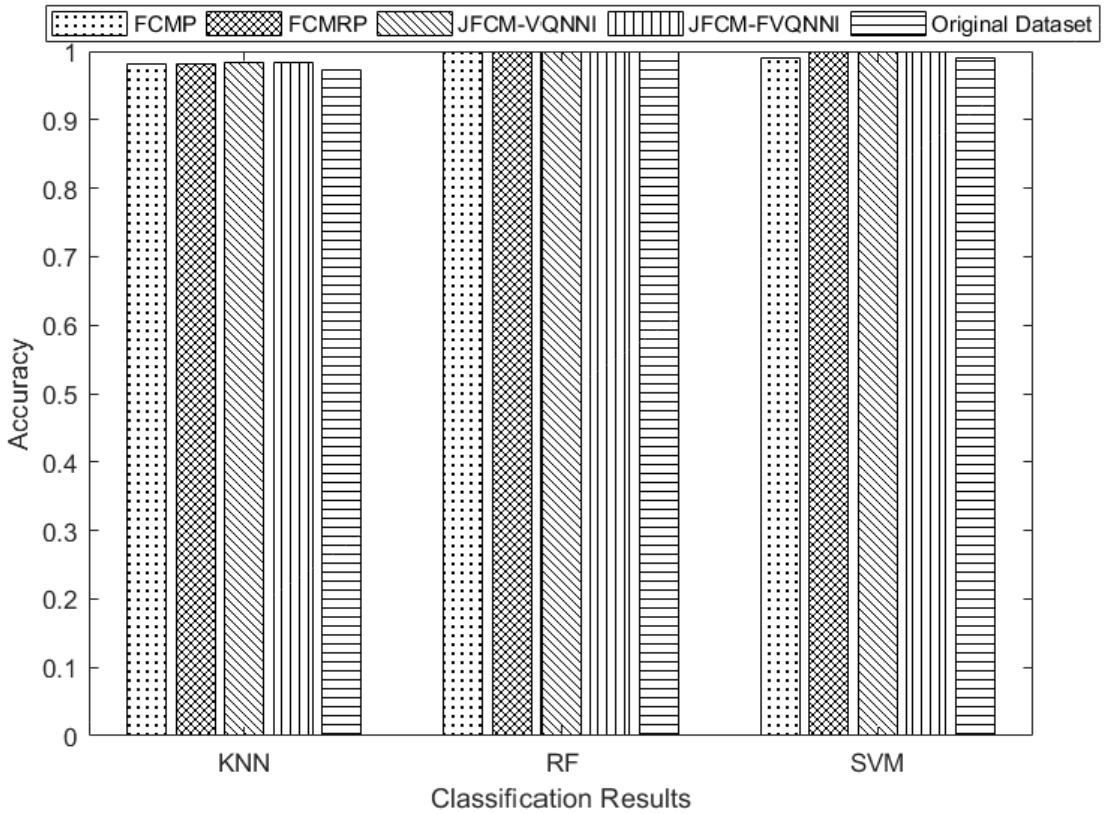

Fig.12. Classification Results for the Imputed Mice protein expression Dataset

Secondly, the classification analysis of Cleveland dataset is illustrated in Fig.13. The classification result of

this dataset is relatively low, after imputation, the imputation algorithms can improve the classification accuracy to varying degrees. From the perspectives of classification results under KNN, RF, and SVM algorithms for the imputed datasets under four types of imputation algorithms, the imputed datasets can achieve much more better results. Moreover, we can draw the conclusion that FCMRP surpass FCMP algorithm, JFCM-VQNNI is superior to FCMRP algorithm, and JFCM-FVQNNI is more advanced than JFCM-VQNNI obviously.

Thirdly, the classification results for Pima dataset after imputation have been studied in Fig.14. Compared with Mice protein expression and Cleveland datasets, the enhancement of the classification effect of the filled datasets is comparatively high. After the missing values imputation strategy, the classification accuracy has been increased varying from 1% to 9%. The reason for this phenomenon is that the original Pima dataset has a higher percentage of missing rate and also the missing types are more complex. This further proves that it is very necessary to perform missing analysis on this type dataset.

In conclusion, combining all of the results of all studied datasets, imputation strategies are inevitable to be conducted for missing datasets especially for the datasets with high missing rate and lower classification accuracy. After comprehensive analysis of the efficiency of KNN, RF, and SVM algorithm after imputation, we can draw the conclusion that FCMRP is better than FCMP, JFCM-VQNNI is excellent than FCMRP, and JFCM-FVQNNI is more higher-ranking than JFCM-VQNNI.

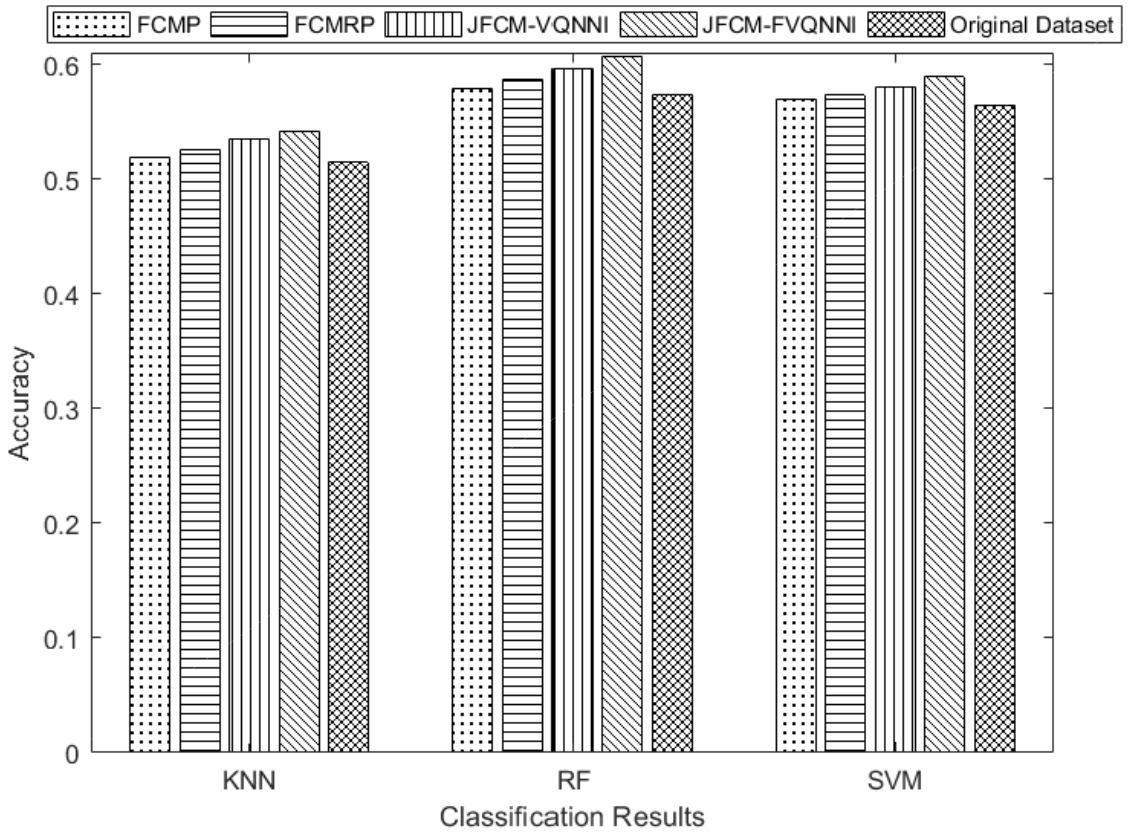

Fig.13. Classification Results for the Imputed Cleveland Dataset

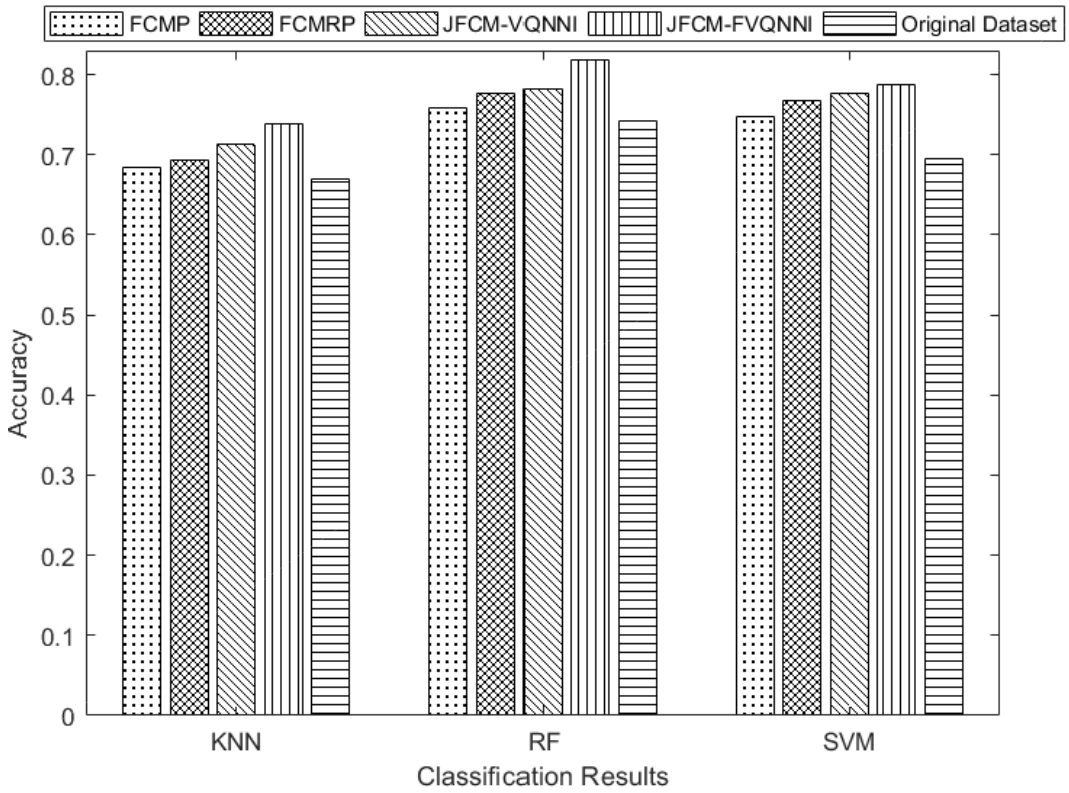

Fig.14. Classification Results for the Imputed Pima Dataset

V. Conclusion

Missing value imputation is a very important task for meaningful information analysis to enhance the performance of the experimental results. In order to better handle the vagueness information, this paper has proposed the imputation algorithms of JFCM-VQNNI and JFCM-FVQNNI by combing fuzzy clustering strategy and vaguely quantified rough set conception. As for JFCM-VQNNI algorithm, initially, fuzzy c-means algorithm is used to cluster the complete objects into several groups, and then, fuzzy similarity relations are implemented to judge the relevance degree of the missing object with its similar records by taking fuzzy nature of clustering into account. JFCM-FVQNNI is an upgrade algorithm of JFCM-VQNNI, which has added the analysis of fuzzy membership of dependent features for instances with the corresponding clusters. In order to increase the efficiency of missing value imputation, JFCM-FVQNNI accelerates fuzzy decision membership adjustment in each instance with respect to the related clusters by considering highly relevant decision attribute.

We have compared the proposed algorithms with two other outstanding existing algorithms of FCMRP that is published in 2019 and FCMP that is firstly published in 2005 and then is improved and concluded in recent research works. The experiments have been conducted on five publicly available datasets. The evaluation criteria in terms of RMSE and MAE, the imputation comparation with actual values, and classification accuracy results have been adopted to judge the effectiveness of the proposed algorithms. Based on the experimental performances, in the aspect of filling efficiency, we can draw the conclusion that FCMRP is better than FCMP, JFCM-VQNNI is superior to FCMRP, and JFCM-FVQNNI can achieve the best performance. Moreover, the experiments have proved that the classification accuracy has been highly improved after conducting filling especially for datasets with a relatively high missing rate, which indicate that it is necessary to fill missing values for missing dataset before performing the other data mining tasks.

It should be pointed out that, especially for large-scale data sets, the proposed JFCM-FVQNNI and JFCM-VQNNI algorithms do have high running time by comparing with FCMRP and FCMP. In the future work, we will study how to ensure imputation efficiency while reducing time complexity. At the same time, the experiments found that using the missing imputation algorithms to fill some complete dataset with different artificial missing ratios, and the experimental performance can even exceed the original complete datasets. This phenomenon indicates that there may be non-obvious missing values in some complete datasets, which will be further analyzed and verified in future work.

References

[1] W.C. Lin and C.F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," Artif Intell Rev, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, DOI: 10.1007/s10462-019-09709-4.

[2] Md. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach," Knowl Inf Syst, vol. 46, no. 2, pp. 389–422, Feb. 2016, DOI: 10.1007/s10115-015-0822-y.

[3] C. Abhishek and T. Cai, "Efficient and Adaptive Linear Regression in Semi-Supervised Settings," Annals of Statistics, vol. 46, no. 4, pp. 1541–1572, 2018, DOI: 10.1214/17-AOS1594.

[4] R. Wei et al., "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data," Scientific Reports, vol. 8, no. 1, p. 663, 2018, DOI: 10.1038/s41598-017-19120-0.

[5] P. S. Raja, K. Sasirekha, and K. Thangavel, "A Novel Fuzzy Rough Clustering Parameter-based missing value imputation," Neural Comput & Applic, Oct. 2019, DOI: 10.1007/s00521-019-04535-9.

[6] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," Information Sciences, vol. 233, pp. 25–35, Jun. 2013, DOI: 10.1016/j.ins.2013.01.021.

[7] P. S. Raja and K. Thangavel, "Soft Clustering Based Missing Value Imputation," in Digital Connectivity – Social Impact, vol. 679, S. Subramanian, R. Nadarajan, S. Rao, and S. Sheen, Eds. Singapore: Springer Singapore, pp. 119–133, 2016, DOI: 10.1007/978-981-10-3274-5_10.

[8] W.Y. Loh, Q. Zhang, W. Zhang, and P. Zhou, "MISSING DATA, IMPUTATION AND REGRESSION TREES," MISSING DATA, p. 30.

[9] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," Neurocomputing, vol. 205, pp. 152–164, Sep. 2016, DOI: 10.1016/j.neucom.2016.04.015.

[10] D. J. Stekhoven and P. Buhlmann, "MissForest--non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, Jan. 2012, DOI: 10.1093/bioinformatics/btr597.

[11] J. Zhou, Z. Lai, D. Miao, C. Gao, and X. Yue, "Multigranulation rough-fuzzy clustering based on shadowed sets," Information Sciences, vol. 507, pp. 553–573, Jan. 2020, DOI: 10.1016/j.ins.2018.05.053.

[12] A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model," Expert Systems with Applications, vol. 115, pp. 68–94, Jan. 2019, DOI: 10.1016/j.eswa.2018.07.057.

[13] J. Xia et al., "Adjusted weight voting algorithm for random forests in handling missing values," Pattern Recognition, vol. 69, pp. 52–60, Sep. 2017, DOI: 10.1016/j.patcog.2017.04.005.

[14] J. Huang, B. Mao, Y. Bai, T. Zhang, and C. Miao, "An Integrated Fuzzy C-Means Method for Missing Data Imputation Using Taxi GPS Data," Sensors, vol. 20, no. 7, p. 1992, Jan. 2020, DOI: 10.3390/s20071992.

[15] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, no. 2–3, pp. 191–203, 1984, DOI: 10.1016/0098-3004(84)90020-7.

[16] Z. Pawlak, "Rough sets," International journal of computer & information sciences, vol. 11, no. 5, pp. 341–356, 1982.

[17] P. Lingras and C. West, "Interval Set Clustering of Web Users with Rough K-Means," Journal of Intelligent Information Systems, vol. 23, no. 1, pp. 5–16, Jul. 2004, DOI: 10.1023/B:JIIS.0000029668.88665.1a.

[18] S. Mitra, H. Banka, and W. Pedrycz, "Rough–Fuzzy Collaborative Clustering," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 4, pp. 795–805, Aug. 2006, DOI: 10.1109/TSMCB.2005.863371.

[19] I. Saha, J. P. Sarkar, and U. Maulik, "Ensemble based rough fuzzy clustering for categorical data," Knowledge-Based Systems, vol. 77, pp. 114–127, Mar. 2015, DOI: 10.1016/j.knosys.2015.01.008.

[20] Z. Ji, Q. Sun, Y. Xia, Q. Chen, D. Xia, and D. Feng, "Generalized rough fuzzy c-means algorithm for brain MR image segmentation," 2012, DOI: 10.1016/j.cmpb.2011.10.010.

[21] F. Cai and F. J. Verbeek, "Rough Fuzzy C-means and Particle Swarm Optimization Hybridized Method for Information Clustering Problem," Journal of Communications, vol. 11, no. 12, pp. 1106–1113, 2016, DOI: 10.12720/jcm.11.12.1106-1113.

[22] P. Maji and S. K. Pal, "RFCM: A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets," Fundamenta Informaticae, vol. 80, no. 4, pp. 475–496, Jan. 2007.

[23] P. S. Raja and K. Thangavel, "Missing value imputation using unsupervised machine learning techniques," Soft Comput, vol. 24, no. 6, pp. 4361–4392, Mar. 2020, DOI: 10.1007/s00500-019-04199-6.

[24] G. Peters and F. Crespo, "An Illustrative Comparison of Rough k-Means to Classical Clustering Approaches," in Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, vol. 8170, D. Ciucci, M. Inuiguchi, Y. Yao, D. Ślęzak, and G. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 337–344, 2013, DOI: 10.1007/978-3-642-41218-9_36.

[25] M. Kryszkiewicz, "Rough set approach to incomplete information systems," Information Sciences, vol. 112, no. 1–4, pp. 39–49, Dec. 1998, DOI: 10.1016/S0020-0255(98)10019-1.

[26] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," Neurocomputing, vol. 205, pp. 152–164, Sep. 2016, DOI: 10.1016/j.neucom.2016.04.015.
[27] J. Prieto-Cubides and C. Argoty, "Dealing with Missing Data using a Selection Algorithm on Rough Sets," International Journal of Computational Intelligence Systems, vol. 11, no. 1, pp. 1307–1321, 2018, DOI: 10.2991/ijcis.11.1.97.
[28] S. Vluymans, L. D'eer, Y. Saeys, and C. Cornelis, "Applications of Fuzzy Rough Set Theory in Machine Learning: a Survey," Fundamenta Informaticae, vol. 142, no. 1–4, pp. 53–86, Dec. 2015, DOI: 10.3233/FI-2015-1284.
[29] H. Zhang, D. Li, T. Wang, T. Li, X. Yu, and A. Bouras, "Hesitant extension of fuzzy-rough set to address uncertainty in classification," Journal of Intelligent & Fuzzy Systems, vol. 34, no. 4, pp. 2535–2550, Jan. 2018, DOI: 10.3233/JIFS-17415.
[30] H. Zhang, D. Li, T. Wang, T. Li, and A. Bouras, "Uncertainty and Equivalence Relation Analysis for Hesitant Fuzzy-Rough Sets and Their Applications in Classification," Computing in Science Engineering, vol. PP, no. 99, pp. 1–1, 2018, DOI: 10.1109/MCSE.2018.110150747.
[31] T. Li, D. Ruan, W. Geert, J. Song, and Y. Xu, "A rough sets based characteristic relation approach for dynamic attribute generalization in data mining," Knowledge-Based Systems, vol. 20, no. 5, pp. 485–494, Jun. 2007, DOI: 10.1016/j.knosys.2007.01.002.
[32] S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg, "Combining Fourier and Lagged k-Nearest Neighbor Imputation for Biomedical Time Series Data," J Biomed Inform, vol. 58, pp. 198–207, Dec. 2015, DOI: 10.1016/j.jbi.2015.10.004.
[33] V. Prasad, T. S. Rao, and M. S. P. Babu, "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms," Soft Comput, vol. 20, no. 3, pp. 1179–1189, Mar. 2016, DOI: 10.1007/s00500-014-1581-5.
[34] A. H. Attia, A. S. Sherif, and G. S. El-Tawel, "Maximal limited similarity-based rough set model," Soft Comput, vol. 20, no. 8, pp. 3153–3161, Aug. 2016, DOI: 10.1007/s00500-016-2243-6.
[35] C. Luo, T. Li, and Y. Yao, "Dynamic probabilistic rough sets with incomplete data," Information Sciences, vol. 417, pp. 39–54, Nov. 2017, DOI: 10.1016/j.ins.2017.06.040.
[36] Y. Qu, Q. Shen, N. M. Parthaláin, C. Shang, and W. Wu, "Fuzzy similarity-based nearest-neighbour classification as alternatives to their fuzzy-rough parallels," International Journal of Approximate Reasoning, vol. 54, no. 1, pp. 184–195, Jan. 2013, DOI: 10.1016/j.ijar.2012.06.008.

Daiwei Li received his B.S. degree from Sichuan Normal University, China in 1999 and M.S. degree from the Southwest Jiaotong University, China in 2003. He is presently associate professor in Chengdu University of Information Technology (CUIT), China. He has won 3 Awards of the Sichuan Provincial Science and Technology Progress Award and authorized more than 20 patents. Chaired and participated more than ten projects in the Sichuan Provincial Science and Technology Department, Department of Education and other vertical and horizontal scientific research projects. His research interests include Data Mining and Knowledge Discovery, Cloud Computing and Big Data, Artificial Intelligence, Rough Sets and Agriculture & meteorological Information Technology.

Haiqing Zhang received her Ph.D. in the DISP (Decision & Information Sciences for Production Systems) laboratory of University Lyon 2. Currently, she is an Associate researcher in Chengdu University of Information Technology. She has published more than 30 research papers in this field. She also holds one National Natural Science Funds and one project supported by Science and Technology of Sichuan Province. Her research interests include Data Mining and Knowledge Discovery, Cloud Computing and Big Data, Artificial Intelligence, Rough Sets, Decision-making Methodology and PLM Maturity Models.

Tianrui Li received his B.S. degree, M.S. degree and Ph.D. degree from the Southwest Jiaotong University, China in 1992, 1995 and 2002 respectively. He was a Post-Doctoral Researcher at Belgian Nuclear Research Centre (SCK • CEN), Belgium from 2005-2006, a visiting professor at Hasselt University, Belgium in 2008, the University of Technology Sydney, Australia in 2009 and the University of Regina, Canada in 2014. And, he is presently a Professor, Deputy Director of National Engineering Laboratory of Integrated Transportation Big Data Application Technology, and the Director of the Key Lab of Cloud Computing and Intelligent Technique of Sichuan Province, Southwest Jiaotong University, China. He currently serves as Director of Professor Committee, School of Information Science and Technology, Southwest Jiaotong University, China.

Since 2000, he has co-edited 6 books, 11 special issues of international journals, 18 proceedings, received 6 Chinese invention patents and published over 360 research papers (e.g., Artificial Intelligence, IEEE Transaction on Knowledge and Data Engineering, IEEE Transactions on Evolutionary Computation, IEEE Transactions on Image Processing, IEEE Transactions on Fuzzy Systems, IEEE Transactions on Information Forensics & Security, IEEE Transactions on Industrial Electronics, IEEE Transactions on Cybernetics, ACM/IEEE Transactions on Audio Speech and Language Processing, IEEE Transactions on Communications, IEEE Transactions on Cloud Computing) in refereed journals and conferences (e.g., ACL, IJCAI, KDD, UbiComp, WWW, ICDM, CIKM, EMNLP). 3 papers were ESI Hot Papers and 15 papers was ESI Highly Cited Papers. His Google H-index is 45. He was recognized as the Top 1% Scientists (ranked 397/3472) in the field of Computer Science based on Thomson Reuter's Essential Science Indicators (ESI) in May 2020. His research interests include Data Mining and Knowledge Discovery, Cloud Computing and Big Data, Artificial Intelligence, Granular Computing and Rough Sets.

Abdelaziz Bouras is Professor in the Computer Science and Engineering Department at Qatar University which he joined as Chair of the ictQatar (Ministry). He is currently the Chair of the IFIP WG5.1 on "Global Product development for the whole life-cycle". His current research interests focus on distributed systems for lifecycle engineering, including ontologies and fuzzy approaches for lifecycle modeling and intelligent products. Abdelaziz was professor at the University of Lyon (France) where he leaded a research team on Information and Decision Systems. He has been conferred the HONORIS-CAUSA honorary Doctoral Degree in Science from Chiang Mai University (Thailand) by Her Royal Highness Princess Maha Chakri Sirindhorn. He co-founded several international journals and conferences and International Journals on product development and lifecycle management (IJPLM, IJPD, IFIP PLM, SKIMA, etc).

Xi Yu received his Ph.D. in the DISP (Decision & Information Sciences for Production Systems) laboratory of University Lyon 2. Currently, he is an Associate Professor in Chengdu University. His research interests lie in the fields of environmental impact assessment, decision making, and deep learning. He is also the leader of the Key Laboratory of Pattern Recognition and Intelligent Information Processing. He had joined two National Natural Science Funds and holds one Sichuan Province project.

Tao Wang is associate professor on Industrial Engineering and Computer Science at University of Saint Etienne, researcher and doctoral advisor in the Decision & Information Sciences for Production Systems laboratory at National Institute of Applied Sciences of Lyon (INSA Lyon). His research interests lie in the interdisciplinary field of Operations Research, Artificial Intelligence, and Healthcare Engineering, especially in the mathematical modeling and multi-agent simulation. Since 2004, he continuously works in several national and regional research projects on health care reform, covering operating theatre, bed management, emergency network, cancer treatment, drug delivery, home care services and medical data mining.