A generalized fuzzy clustering framework for incomplete data by integrating feature weighted and kernel learning

Ying Yang Corresp., 1 , **Haoyu Chen** 2 , **Haoshen Wu** 3

1 College of Information and Interlligence, Hunan Agricultural University, Changsha, China

2 New Energy College,, Xi'an Shiyou University, Xi'an, China

3 College of Management, Guangdong University of Technology, Guangzhou, China

Corresponding Author: Ying Yang
Email address: 1254809709@stu.hunau.edu.cn

Missing data presents a challenge to clustering algorithms, as traditional methods tend to pad incomplete data first before clustering. To cluster the two processes and improve the clustering accuracy, a generalized fuzzy clustering framework is proposed based on optimal completion strategy (OCS) and nearest prototype strategy (NPS) with four improved algorithms developed. Feature weights were introduced to reduce outliers' influence on the cluster centers, and kernel functions were used to solve the linear indistinguishability problem. The proposed algorithms were evaluated regarding correct clustering rate, iteration number, and external evaluation indexes with nine datasets from the UCI database. The results of the experiment indicate that the clustering accuracy of the feature weighted fuzzy C - means algorithm with NPS (NPS - WFCM) and the feature weighted fuzzy C - means algorithm with OCS (OCS - WFCM) under varying missing rates is superior to that of seven conventional algorithms. Meanwhile, feature weighted kernel fuzzy C - means algorithm with NPS (NPS - WKFCM) and feature weighted kernel fuzzy C - means algorithm with OCS (OCS - WKFCM) are better than OCS - WFCM and NPS - WFCM in all indexes. Experiments demonstrate that the enhanced algorithm proposed for clustering incomplete data is superior.

A generalized fuzzy clustering framework for incomplete data by integrating feature weighted and kernel learning

Ying Yang[1,*], Haoyu Chen[2], Haoshen Wu[3]
[1]College of Information and Interlligence, Hunan Agricultural University, 410125, Changsha, China
[2] New Energy College, Xi'an Shiyou University, Xi'an, 710065, China
[3] College of Management, Guangdong University of Technology, 510000, Guangzhou, China
***Correspondence:** Ying Yang, 1254809709@stu.hunau.edu.cn

Abstract

Missing data presents a challenge to clustering algorithms, as traditional methods tend to pad incomplete data first before clustering. To cluster the two processes and improve the clustering accuracy, a generalized fuzzy clustering framework is proposed based on optimal completion strategy (OCS) and nearest prototype strategy (NPS) with four improved algorithms developed. Feature weights were introduced to reduce outliers' influence on the cluster centers, and kernel functions were used to solve the linear indistinguishability problem. The proposed algorithms were evaluated regarding correct clustering rate, iteration number, and external evaluation indexes with nine datasets from the UCI database. The results of the experiment indicate that the clustering accuracy of the feature weighted fuzzy C - means algorithm with NPS (NPS - WFCM) and the feature weighted fuzzy C - means algorithm with OCS (OCS - WFCM) under varying missing rates is superior to that of seven conventional algorithms. Meanwhile, feature weighted kernel fuzzy C - means algorithm with NPS (NPS - WKFCM) and feature weighted kernel fuzzy C - means algorithm with OCS (OCS - WKFCM) are better than OCS - WFCM and NPS - WFCM in all indexes. Experiments demonstrate that the enhanced algorithm proposed for clustering incomplete data is superior.

keywords: Incomplete data; Fuzzy C - Means; Kernel function; Feature weights

1. Introduction

In entering the information society, people have also entered the data society. All areas are flooded with massive amounts of data with complex trends. Clustering analysis[1] is an unsupervised learning technique, which can autonomously classify data without a priori knowledge. Additionally, it is one of the effective tools to fully exploit the value present in the data. The traditional hard clustering approach considers that data objects can be grouped entirely into a certain category. However, in real life, there are no clear boundaries for many things. Some scholars introduced the fuzzy set theory[2] into the clustering algorithm and proposed the FCM algorithm. The algorithm represents the relationship between data and clusters with an affiliation value of 0 - 1, which is more suitable for practical clustering problems. Whereas, the FCM algorithm cannot directly cluster incomplete datasets. But missing datasets are more prevalent in real - world fields such as industry, medicine, business and scientific research[5]. Nearly 45% of the datasets in the UCI database are missing relevant data. Not only do missing data result in the loss of a substantial quantity of valuable information, but they also present difficulties for cluster analysis. Therefore, it is of great practical importance to investigate fuzzy clustering algorithms for incomplete data.

Numerous researchers have proposed enhanced algorithms to address the issue of FCM clustering of insufficient data. The most classic of these are the four improved fuzzy clustering algorithms for incomplete data proposed by Bezdek and Hathaway[6]. Based on whole data strategy (WDS), partial distance strategy (PDS), optimized complete strategy (OCS), and nearest prototype strategy (NPS), four algorithms are enhanced. (NPS). The WDS - FCM algorithm is a rounding method that discards missing values. The PDS-FCM algorithm improves the formulation of the FCM clustering algorithm by introducing the local distance introduced by Dixon[7] without considering missing values in the calculation to fulfill incomplete data clustering. The OCS - FCM algorithm continuously interpolates absent values as updateable variables. In addition, the NPS - FCM algorithm replaces absent values with attribute values corresponding to clustering centers closest to the incomplete data. The four algorithms provide effective ideas for the interpolation of incomplete data.

Among the four strategies, the OCS and the NPS are more widely adopted and continuously improved by researchers. Li et al.[8] proposed an interval kernel fuzzy C-means clustering method for incomplete data by converting the incomplete data set into an interval data set and introducing the NPS-based kernel method. Najib[9] modified the NPS-FCM algorithm based on the continuous mechanism so that it can be used to aggregate incomplete data streams with high error rates. Meng[10] applied the OCS-FCM algorithm to incomplete spectral data to calculate galaxy abundances at high redshifts. Villuendas[11] presented a cluster intelligence-based framework for clustering incomplete data using a swarm intelligence algorithm to determine cluster centers and hyperparameters. Shi et al.[12] proposed a clustering algorithm based on the relationship between attributes, which combines support vector machines with the four clustering strategies mentioned above.

In addition, another solution for clustering incomplete data is to first interpose the missing values by evaluation and then cluster the completed dataset. Due to the few parameters and straightforward principle of the K - Nearest Neighbor (KNN) algorithm, it is gaining popularity for interpolating incomplete data[13]. Doquire and Veleysen[14] estimated the missing values of fragmentary data using a KNN method based on mutual information.Tutz and Ramzan[15] proposed The weighted KNN imputation method, which uses a kernel function to generate weights and achieves a reduction in interpolation error. Tsai[16] introduced a missing value interpolation method based on the class center. The method classifies the dataset, calculates the distance between different classes, and determines the threshold to be filled according to the magnitude of the distance. Williams et al. [17] put forward the Bayesian comparative compression prediction and empirical modal decomposition algorithm. It has a significant filling advantage for signal-type data. Baligh et al. [18] presented a novel genetic programming and weighted KNN-based interpolation method for incomplete data regression.

Based on the idea of Expectation - Maximization(EM), the corresponding incomplete data processing and clustering methods are proposed. Eirola et al.[19] fitted the Gaussian mixture model with the EM algorithm, which was then used to estimate the distance between incomplete data and to cluster the incomplete dataset. The vector autoregressive model - imputation algorithm proposed by Faraj[20] is used to deal with incomplete data. When the data are missing randomly, the EM method cannot achieve good results. Using the EM algorithm, Hung[21] estimated the parameters of the absent values.

With intensive research and development, neural networks are also used to process incomplete data. Vadlamani[22] introduced automatic associative neural networks for valuation of incomplete data. Rancoita[23] suggested using Bayesian networks to model the dependencies between data variables and perform data valuation. Kancherla et al. [24] put forward a probabilistic neural network - based algorithm for incomplete data estimation. Dušan[25] introduced a multiple valuation algorithm for incomplete data based on the Gaussian mixture model and extreme learning machine.

After filling the incomplete dataset with various interpolation methods, the second step is to perform clustering. Several experts have improved the clustering algorithm from the perspective of dataset attributes[26]. The idea of feature weights was first introduced into the clustering algorithm by Desarbo[27]. The core of the algorithm is to determine the weights of the features using K - means clustering. Makarenkov[28] extended the clustering algorithm and selected the optimal feature weights for K - means clustering. In order to solve the clustering of complex data, Zhang[29] introduced the kernel method into the clustering algorithm and proposed the k - medoids cluster algorithm. Modha et al. [30] investigated a new method for determining the feature weights by minimizing the generalized Fisher ratio for feature - weighted K - means clustering algorithm, which leads to better clustering results.

The three interpolation methods mentioned above all present different disadvantages. KNN filling - based clustering methods can achieve better results only in large - scale sparse data with few values of missing attributes. The EM - based clustering methods often fail to obtain the desired filling effect when there is a large amount of missing data, or a certain large class of values is missing. The neural network - based clustering methods require a large amount of model training to estimate the missing values of individual missing instances, which greatly increases the computational cost. Although the clustering improvement methods that introduce feature weighting and kernel functions[31] are effective, methods that split the interpolation and clustering ultimately lead to a secondary reduction of computational accuracy.

So far, it is still an open issue how to effectively solve the clustering task for incomplete data. To enhance the performance of incomplete data clustering tasks, we therefore propose a generalized fuzzy clustering framework integrating feature weights and kernel learning. Currently, a number of experiments conducted on public data sets demonstrate the efficacy and superiority of the proposed method. The following are the primary contributions of this work:

1. On the basis of OCS and NPS in literature[6], we unify imputation, feature learning and

clustering as one optimization objective, and propose OCS - WFCM and NPS - WFCM, respectively.

2. In order to better adapt to incomplete data clustering in complex cases (e.g., non-linear data), we further propose kernel-based OCS - WKFCM and NPS - WKFCM methods.

3. An alternate optimization method is used to solve the objective functions of the above methods, and the optimal solutions are obtained by iterative updating of variables.

The research is structured as follows. In section 2, the FCM algorithm theory and four strategies for incomplete data are analyzed in detail. In section 3, four improved algorithms based on the established framework are introduced. In section 4, comparing the four algorithms proposed in this research to other fragmentary data clustering algorithms verifies the framework's efficacy. Finally, the summaries and optimizations are given in section 5.

2. Analysis of incomplete data clustering algorithm

2.1 Fuzzy C - means algorithm

FCM algorithm's fundamental concept is to minimize objective function to solve clustering center and membership matrix. The primary implementation process is to establish the objective function formula based on the data sample's proximity to the clustering centroid. Iteratively updating the membership moment clustering center matrix, the algorithm determines the objective function's extreme point. Finally, the category of the data sample is determined according to the size of the membership value[32].

Let $U_{(c\times n)}$ represent the membership matrix, and V represent the cluster center matrix. Suppose a dataset $X=\{x_1,x_2,\cdots,x_n\}$ exists in s dimensions and n samples. The dataset can be represented as $x_k=\left[x_{1k},x_{2k},\cdots,x_{sk}\right]^T$, and the samples can be defined as x_{ik}. The number of sample clusters in the dataset is set to c, the membership value of data x_j to category i is expressed as $u_{ij}\in U_{(c\times n)}$. The sample x_k is characterized by different affiliation values for different clusters, and the sum of c categories' membership values is 1. That is, u_{ij} is shown in the constraint formula (2.1).

$$\begin{aligned}&\sum_{i=1}^{c}u_k=1,k=1,2,\cdots,n \text{ and } \mu_{ij}\in[0,1]\\&0<\sum_{k=1}^{n}u_{ik}<1,\ i=1,2,\cdots,c\end{aligned}\tag{2.1}$$

The objective function formula established by FCM is shown in (2.2).

$$\min J_m(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\left\|x_k-v_i\right\|_2^2\tag{2.2}$$

Where, m is the fuzzy weighting coefficient, $\|\cdot\|_2$ is normal form, cluster center $V=\{v_1,v_2,\cdots,v_c\}$, and the membership matrix $U_{(c\times n)}$, $J(U,V)$ equals the sum of the sample cluster squares and the cluster center.

Lagrange multiplier method is used to solve the multivariate function's extreme value, which is used to solve membership function matrix and clustering center function matrix of FCM algorithm. The membership updating formula is shown in (2.3).

$$u_{ik} = \left[\sum_{t=1}^{c} \left(\frac{\|x_k - v_i\|_2^2}{\|x_k - v_t\|_2^2} \right)^{\frac{1}{\text{m-1}}} \right]^{-1}, \ i = 1, 2, \cdots, c; \ k = 1, 2, \cdots, n \quad (2.3)$$

The cluster center update formula is shown in (2.4).

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}{}^{m} x_k}{\sum_{k=1}^{n} u_{ik}{}^{m}}, i = 1, 2, \cdots, c \quad (2.4)$$

2.2 Improved FCM algorithm for incomplete data

Four classical FCM for incomplete data that Hathaway and Bezdek[6] proposed are well used. The data set information is described as follows :

$\bar{X} = \{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n\}$ is an incomplete data set, the single sample data in the data set is expressed as $\bar{x}_i = [\bar{x}_{1i}, \bar{x}_{2i}, \cdots, \bar{x}_{si}]^T \ (1 \le k \le n)$, and the number of attribute values is s.

$\bar{X}$ will be divided into two types of data sample sets: complete data set $X_W = \{x_k \in X | x_k \text{ is the complete data sample}\}$ and incomplete data sample set $\tilde{X}_N = \{\tilde{x}_k \in \tilde{X} | \tilde{x}_k \text{ is an incomplete data sample}\}$. The attribute information set is divided into two categories : $1 \le j \le s$, $1 \le k \le n$, complete data set $X_P = \{x_{jk} | x_{jk} \text{ is the complete attribute}\}$, and missing attribute set $X_M = \{x_{jk} = ? | x_{jk} \text{ is the missing attribute}\}$.

2.2.1 FCM algorithm with whole data strategy

In the WDS-FCM algorithm, a simple method is used to directly discard the samples with missing attributes. Then, the data samples in the sample set $X_P = \{x_{jk} | x_{jk} \text{ is the complete attribute}\}$ are directly clustered by FCM.

The dealing strategy of WDS - FCM algorithm will cause data samples with missing attributes to discard other complete attributes. This can result in a large amount of wasted data information. When the missing rate in the dataset is low, it has little effect on the overall dataset. With the remaining complete sample for fuzzy clustering, the calculated clustering center is not much different from the original data clustering center. Due to the absence of a large number of attributes, the clustering accuracy will be significantly impacted by an increase in the missing rate. Therefore, Hathaway and Bezdek[6] suggest that WDS - FCM algorithm is more suitable for clustering analysis of datasets, as the proportion of missing attribute information in incomplete datasets is less than 0.25.

The WDS-FCM algorithm proceeds as follows:

(1) Split data X : The incomplete data set is separated into two sections: the complete part X_P, the missing part X_M, and $X = X_P \cup X_M$. In the experiment, X_P instead of X, X_M in FCM algorithm does not participate in the calculation.

(2) Initialization : iterative convergence threshold ε, fuzzy parameter m, cluster number $c\left(2 \le c \le \sqrt{n}\right)$, maximum number of iterations G, initial membership matrix $U^{(0)}$.

(3) Updating the cluster center : when the algorithm performs L ($L = 1, 2, \ldots$) iterations, cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and the cluster center calculation formula (2.4).

(4) Calculation of membership matrix : according to $V^{(l)}$ and (2.3), solve membership matrix $U^{(l)}$.

(5) Iteration termination : when the iteration count approaches $L = G$, or $\forall i$, k, $\max\left|u_{ik}^{(l)} - u_{ik}^{(l-1)}\right| < \varepsilon$, WDS - FCM algorithm iteration stops, the algorithm ends, the output membership U and cluster center V ; or else $L = L + 1$, return (3) to continue.

2.2.2 FCM algorithm with partial distance strategy

On the basis of WDS - FCM, PDS - FCM in terms of attributes, the attributes participate in calculating local distances as long as they exist. When the attribute is missing, the complete attribute participation is converted. The distance between missing data sample x_k and cluster center v_i is determined according to attribute ratio.

$$D_{ik} = \frac{s}{\sum_{j=1}^{s} I_{jk}} \sum_{j=1}^{s} \left(x_{jk} - v_{ji}\right)^2 I_{jk} \tag{2.12}$$

Among them,

$$I_{jk} = \begin{cases} 0, & if\ x_{jk} \in \tilde{X}_M \\ 1, & if\ x_{jk} \in \tilde{X}_P \end{cases}, \ 1 \le j \le s,\ 1 \le k \le n. \tag{2.13}$$

The clustering center at the extremum point is as follows.

$$v_{ji} = \frac{\sum_{k=1}^{n} \mu_{ik}^{m} I_{jk} x_{jk}}{\sum_{k=1}^{n} \mu_{ik}^{m} I_{jk}},\ 1 \le j \le s,\ 1 \le i \le c \tag{2.14}$$

The membership formula is shown as (2.15).

$$u_{ik} = \left[\sum_{t=1}^{c} \left(\frac{\left\|x_k - v_i\right\|_2^2}{\left\|x_k - v_t\right\|_2^2}\right)^{\frac{1}{\text{m-1}}}\right]^{-1},\ i = 1, 2, \cdots, c;\ k = 1, 2, \cdots, n \tag{2.15}$$

The PDS - FCM algorithm proceeds as follows:

(1) Initialization : iterative convergence threshold ε, fuzzy parameter m, cluster number $c\left(2 \le c \le \sqrt{n}\right)$, maximum number of iterations G, initial membership matrix $U^{(0)}$.

(2) Updating the cluster center : when the algorithm performs L (L = 1, 2, …) iterations, the cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (2.14).

(3) Calculating the membership matrix : according to $V^{(l)}$ and (2.15), solving membership matrix $U^{(l)}$.

(4) Iteration termination : when the iteration count approaches $L = G$, or $\forall i, \quad k$, $\max\left|u_{ik}^{(l)} - u_{ik}^{(l-1)}\right| < \varepsilon$, PDS - FCM algorithm iteration stops, the algorithm ends, the output membership U and cluster center V; otherwise $L = L + 1$, return (3) to continue.

2.2.3 FCM algorithm with optimal completion strategy

The OCS - FCM algorithm assigns the lacking attributes as variables and incorporates variables into the objective function calculation of the FCM algorithm. Iterative clustering is performed with variables instead of missing attributes.

The variable membership U and the cluster center V are iteratively updated in the clustering iteration process to find the optimal value. The objective function formula established by OCS - FCM is (2.16).

$$J\left(U,V,\tilde{X}_M\right) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\left\|\tilde{x}_k - v_i\right\|_2^2 \qquad (2.16)$$

Using the Lagrange multiplier method to locate the extremum of objective function (2.16), the missing attribute update formula (2.17) is obtained.

$$\mathrm{x}_{jk} = \frac{\sum_{i=1}^{c} u_{ik}^{m} v_{ji}}{\sum_{i=1}^{c} u_{ik}^{m}} \qquad (2.17)$$

The main steps of OCS - FCM algorithm are :

(1) Initialization : Set the fuzzy parameter m, number of clusters $c\left(2 \le c \le \sqrt{n}\right)$, utmost allowed iterations G, iterative convergence threshold ε, the missing attribute matrix $\tilde{X}_M^{(0)}$, and the membership matrix $U^{(0)}$ combined with the constraint conditions.

(2) Updating the cluster center matrix : when the algorithm performs L (L = 1, 2, …) iterations, the cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (2.3).

(3) Calculate the membership matrix : according to $V^{(l)}$, and (2.4) solving membership matrix $U^{(l)}$.

(4) Update the missing value : calculate the missing value $\tilde{X}_M^{(l)}$ according to the membership partition matrix $U^{(l)}$ and cluster center matrix $V^{(l)}$ and (2.17).

(5) Iteration termination : when the iteration count approaches $L = G$, or $\forall i, \quad k$, $\max\left|u_{ik}^{(l)} - u_{ik}^{(l-1)}\right| < \varepsilon$, OCS - FCM algorithm iteration stop, the output U and V.

2.2.4 FCM algorithm with the nearest prototype strategy

The NPS - FCM algorithm is an estimation method. In the NPS - FCM algorithm, the missing data attributes in the NPS - FCM algorithm participate in clustering with the nearest neighbor center instead. The missing data no longer remain constant after pre - population. During the iterative process, the corresponding attribute values of the clustering centers are continuously followed and adjusted. The filling method for missing attributes is as follows (2.18).

$$x_{jk}^{(l)} = v_{ji}, D_{ik} = \min\left\{D_{1k}, D_{2k}, \cdots, D_{ck}\right\} \tag{2.18}$$

The NPS - FCM algorithm is based on the OCS - FCM algorithm. In the process of iteration, the missing data attribute is replaced by (2.18), and then the clustering analysis is performed according to the implementation steps of the OCS - FCM algorithm.

The main steps of NPS - FCM algorithm are :

(1) Initialization : Set the fuzzy parameter m, the number of clusters $c\left(2 \le c \le \sqrt{n}\right)$, the maximum number of iterations G, the iterative convergence threshold ε, the missing attribute matrix $\tilde{X}_M^{(0)}$, and the membership matrix $U^{(0)}$ combined with the constraint conditions.

(2) Updating the cluster center matrix : when the algorithm performs L ($L = 1, 2, \ldots$) iterations, the cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (2.3).

(3) Calculate the membership matrix : according to $V^{(l)}$, and (2.4) solving membership matrix $U^{(l)}$.

(4) Update the missing value : calculate the missing value $\tilde{X}_M^{(l)}$ according to the membership partition matrix $U^{(l)}$ and cluster center matrix $V^{(l)}$ and (2.18).

(5) Iteration termination : when the iteration count approaches $L = G$, or $\forall i$, k, $\max\left|u_{ik}^{(l)} - u_{ik}^{(l-1)}\right| < \varepsilon$, NPS - FCM algorithm iteration stop, the output U and V.

3. Feature weighted kernel function FCM of incomplete data

3.1 Feature weighted FCM of incomplete data

3.1.1 Feature weighted FCM algorithm with OCS

In order to solve the defects of FCM in practical application, the different contributions of FCM and sample attribute vectors to classification are considered. The sample attribute weight is introduced into the objective function, which can obtain more effective clustering analysis results. This method is called the feature weighted FCM algorithm (WFCM).

In the optimization of the complete strategy, the sample data x_{jk} is composed of two segments, the complete attribute part $x_{jk}(o_{jk})$, and the missing attribute part $x_{jk}(m_{jk})$. Then $x_{jk}(o_{jk}) \cup x_{jk}(m_{jk}) = x_{jk}$, $x_{jk}(o_{jk})$ remain unchanged in the clustering process. Assuming that u_{ij} represents the degree of the j sample data x_j belonging to the i cluster (the cluster center is v_i), v_{ik} represents the i feature of the k cluster center, w_{ik} represents the weight of the i feature of the k cluster center, the objective function that OCS - WFCM needs to minimize is :

$$\begin{aligned}
&\min \sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l} u_{ij}^{m}\omega_{ik}^{\beta}\left\|x_{jk}-v_{ik}\right\|^{2}\\
&s.t.\sum_{i=1}^{c} u_{ij}=1, u_{ij}\in[0,1],\\
&\quad \sum_{k=1}^{l} w_{ik}=1, w_{ik}\in[0,1],\\
&i=1,2,\cdots,c\\
&j=1,2,\cdots,n\\
&k=1,2,\cdots,l
\end{aligned} \tag{3.1}$$

Furthermore, because of $x_{jk}=\left[x_{jk}\left(o_{jk}\right), x_{jk}\left(m_{jk}\right)\right]$, (3.1) is equivalent to

$$\min \sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l} u_{ij}^{m}\omega_{ik}^{\beta}\left(\left\|x_{jk}\left(o_{jk}\right)-v_{ik}\right\|^{2}+\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}\right) \tag{3.2}$$

Because the complete attribute $x_{jk}(o_{jk})$ remains unchanged during the clustering process and is a fixed constant, the minimum value of (3.2) can be simplified as

$$\min \sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l} u_{ij}^{m}\omega_{ik}^{\beta}\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2} \tag{3.3}$$

The optimal solution of (3.3) can be further analyzed as

$$x_{jk}\left(m_{jk}\right)=\frac{\sum_{i=1}^{c} u_{ij}^{m}\omega_{ik}^{\beta} v_{ik}}{\sum_{i=1}^{c} u_{ij}^{m}\omega_{ik}^{\beta}} \tag{3.4}$$

In order to obtain the membership degree, cluster center and weight matrix, the Lagrange method is used to solve (3.3).

If x is known, then

$$\sum_{j=1}^{n}\lambda_{j}\left(\sum_{i=1}^{c} u_{ij}-1\right)=0 \tag{3.5}$$

where λ is the Lagrange multiplier, and λ is a vector composed of the Lagrange multiplier $\lambda_1, \lambda_2, \cdots \lambda_n$.

Combining (3.3) and (3.5), we can get

$$\begin{aligned}
&\sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l} u_{ij}^{m}\omega_{ik}^{\beta}\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}\\
&=\sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l} u_{ij}^{m}\omega_{ik}^{\beta}\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}-\sum_{j=1}^{n}\lambda_{j}\left(\sum_{i=1}^{c} u_{ij}-1\right)
\end{aligned} \tag{3.6}$$

Let $Q_{ij}=\sum_{k=1}^{l}\omega_{ik}^{\beta}\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}$, further obtain

$$J_{OCS-WFCM}=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}Q_{ij}-\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right). \tag{3.7}$$

Get the partial derivative of u_{ij} and get

$$\frac{\partial J_{OCS-WFCM}}{\partial u_{ij}}=mu^{m-1}_{ij}Q_{ij}-\lambda_j=0. \tag{3.8}$$

Therefore,

$$u_{ij}=\left(\frac{\lambda_j}{mQ_{ij}}\right)^{\frac{1}{m-1}}. \tag{3.9}$$

And $\sum_{i=1}^{c}u_{ij}=1$ is known, combined with (3.9), we get

$$\lambda_j^{\frac{1}{m-1}}=\sum_{i=1}^{c}\left(\frac{1}{mQ_{ij}}\right)^{\frac{-1}{m-1}} \tag{3.10}$$

Further obtained

$$u_{ij}=\frac{\sum_{r=1}^{c}\left(\frac{1}{mQ_{rj}}\right)^{\frac{-1}{m-1}}}{mQ_{ij}^{\frac{1}{m-1}}}=\left(\sum_{i=1}^{c}\frac{Q_{ij}}{Q_{rj}}\right)^{\frac{1}{1-m}}$$
$$=\left(\sum_{i=1}^{c}\frac{\sum_{k=1}^{l}\omega_{ik}^{\beta}\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^2}{\sum_{k=1}^{l}\omega_{rk}^{\beta}\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^2}\right)^{\frac{1}{1-m}}. \tag{3.11}$$

Similarly, one can obtain

$$\omega_{ik}=\left(\sum_{t=1}^{l}\frac{\sum_{j=1}^{n}u_{ij}^{m}\cdot\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^2}{\sum_{j=1}^{n}u_{ij}^{m}\cdot\left\|x_{jt}\left(m_{jt}\right)-v_{ik}\right\|^2}\right)^{\frac{1}{1-\beta}}. \tag{3.12}$$

Next, take the partial derivative of v_{ik} in Equation (3.3) to get

$$\frac{\partial J_{OCS-WFCM}}{\partial v_{ik}}=-2\sum_{j=1}^{n}u^{m}_{ij}\omega_{\mathrm{ik}}^{\beta}\cdot\left(x_{jk}\left(m_{jk}\right)-v_{ik}\right)=0. \tag{3.13}$$

Further obtained

$$v_{ik}=\frac{\sum_{j=1}^{n}u_{ij}^{m}\omega_{ik}^{\beta}x_{jk}\left(m_{jk}\right)}{\sum_{j=1}^{n}u_{ij}^{m}\omega_{ik}^{\beta}} \tag{3.14}$$

It is observed by (3.14) that when $\omega_{ik}^{\beta}=0$, there is $v_{ik}=0$. The formula for v_{ik} is

$$v_{ik}=\begin{cases}0 & ,if\ \ \omega_{ik}^{\beta}=0\\ \dfrac{\sum_{j=1}^{n}u_{ij}^{m}x_{jk}\left(m_{jk}\right)}{\sum_{j=1}^{n}u_{ij}^{m}} & ,if\ \ \omega_{ik}^{\beta}\neq 0\end{cases} \tag{3.15}$$

The main steps of OCS - WFCM algorithm are :

(1) Initialization : Set the fuzzy parameter m, number of clusters $c\left(2\le c\le\sqrt{n}\right)$, utmost allowed iterations G, iterative convergence threshold ε, the missing attribute matrix $\tilde{X}_M^{(0)}$, and the membership matrix $U^{(0)}$ combined with the constraint conditions.

(2) Updating the cluster center matrix : when the algorithm performs L ($L = 1, 2, \ldots$) iterations, cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (3.15).

(3) Calculate the membership matrix : according to $V^{(l)}$, and (3.11) solving membership matrix $U^{(l)}$.

(4) Calculate the weight matrix : according to $V^{(l)}$, and (3.12) to solve the weight matrix.

(5) Update the missing value : calculate the missing value $\tilde{X}_M^{(l)}$ according to the membership partition matrix $U^{(l)}$ and cluster center matrix $V^{(l)}$ and (3.4).

(6) Iteration termination : when the iteration count approaches $L = G$, or $\forall i$, k, $\max\left|u_{ik}^{(l)}-u_{ik}^{(l-1)}\right|<\varepsilon$, OCS - WFCM algorithm iteration stop, the output U and V.

3.1.2 Feature weighted FCM algorithm with NPS

In the interpolation of NPS - WFCM, the sample data x_{jk} is also divided into two parts, the complete attribute part $x_{jk}(o_{jk})$, and the missing attribute part $x_{jk}(m_{jk})$. Then, $x_{jk}(o_{jk}) \cup x_{jk}(m_{jk}) = x_{jk}$, $x_{jk}(o_{jk})$ remain unchanged in the clustering process. The filling method of missing attributes in NPS - WFCM is as follows (3.16).

$$x_{jk}\left(o_{jk}\right)=v_{ik}=\begin{cases}0 & ,\ if\ \ \omega_{ik}^{\beta}=0\\ \dfrac{\sum_{j=1}^{n}u_{ij}^{m}x_{jk}}{\sum_{j=1}^{n}u_{ij}^{m}} & ,\ if\ \ \omega_{ik}^{\beta}\neq 0\end{cases},\ D_{ij}=\min\left\{D_{1j},D_{2j},\ldots,D_{cj}\right\} \tag{3.16}$$

Similar to OCS - WFCM, only (3.15) needs to be replaced with (3.16) when updating the missing attributes.

The main steps of NPS - WFCM algorithm are :

(1) Initialization : Set the fuzzy parameter m, number of clusters $c\left(2 \le c \le \sqrt{n}\right)$, utmost allowed iterations G, iterative convergence threshold ε, the missing attribute matrix $\tilde{X}_M^{(0)}$, and the membership matrix $U^{(0)}$ combined with the constraint conditions.

(2) Updating the cluster center matrix : when the algorithm performs L ($L = 1, 2, \ldots$) iterations, cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (3.15).

(3) Calculate the membership matrix : according to $V^{(l)}$, and (3.11) solving membership matrix $U^{(l)}$.

(4) Calculate the weight matrix : according to $V^{(l)}$, and (3.12) to solve the weight matrix.

(5) Update the missing value attribute: calculate the missing value $\tilde{X}_M^{(l)}$ according to the membership partition matrix $U^{(l)}$ and cluster center matrix $V^{(l)}$ and (3.16).

(6) Iteration termination : when the iteration count approaches $L = G$, or $\forall i$, k, $\max\left|u_{ik}^{(l)} - u_{ik}^{(l-1)}\right| < \varepsilon$, NPS - WFCM algorithm iteration stop, the output U and V.

3.2 Feature weighted kernel FCM of incomplete data

3.2.1 Feature weighted kernel FCM clustering with OCS

In this section, the kernel function is introduced into the OCS - WFCM in the previous section. Clustering is performed in the kernel space, and the observed data is mapped to a higher dimensional feature space in a nonlinear way to achieve nonlinear classification technology. It is assumed that φ is a nonlinear mapping function, $\varphi : x \to \varphi(x) \in$ maps the high characteristic space, where $x \in X = \{x_1, x_2, \cdots, x_n\}$. $\varphi(x_{jk})$ is the mapping of the jth sample data point to the kth feature in the feature space. The optimization objective function of feature weighted kernel FCM (WKFCM) with OCS is as follows.

$$\min \sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l} u_{ij}^{m}\omega_{ik}^{\beta}\left\|\phi\left(x_{jk}\left(m_{jk}\right)\right)-\phi\left(v_{ik}\right)\right\|^2 \tag{3.17}$$

Expanding $\left\|\phi\left(x_{jk}\left(m_{jk}\right)\right)-\phi\left(v_{ik}\right)\right\|^2$ in (3.17), we can get

$$\begin{aligned}&\left\|\phi\left(x_{jk}\left(m_{jk}\right)\right)-\phi\left(v_{ik}\right)\right\|^2\\ &=\phi\left(x_{jk}\left(m_{jk}\right)\right)\cdot\phi\left(x_{jk}\left(m_{jk}\right)\right)-2\phi\left(x_{jk}\left(m_{jk}\right)\right)\cdot\phi\left(v_{ik}\right)+\phi\left(v_{ik}\right)\cdot\phi\left(v_{ik}\right)\\ &=K\left(x_{jk}\left(m_{jk}\right),x_{jk}\left(m_{jk}\right)\right)-2K\left(x_{jk}\left(m_{jk}\right),v_{ik}\right)+K\left(v_{ik},v_{ik}\right)\end{aligned} \tag{3.18}$$

Where, $K(x,y)=\phi(x)\cdot\phi(y)$ represents the kernel function, which can be used to represent the dot product in the high - dimensional feature space. The kernel function used in this work is the Gaussian kernel function, that is, $K(x,y)=\exp\left(\frac{-\|x-y\|^2}{\sigma^2}\right)$, then $K(x,x)=1$.

Simplifying (3.17) relative to (3.18) yields

$$\begin{aligned}
&\sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l}u_{ij}^{m}\omega_{ik}^{\beta}\left\|\phi\left(x_{jk}\left(m_{jk}\right)\right)-\phi\left(v_{ik}\right)\right\|^{2} \\
&=2\sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l}u_{ij}^{m}\omega_{\mathrm{ik}}^{\beta}\left(1-K\left(x_{jk}\left(m_{jk}\right),v_{ik}\right)\right) \\
&=2\sum_{i=1}^{c}\sum_{j=1}^{n}\sum_{k=1}^{l}u_{ij}^{m}\omega_{ik}^{\beta}\left(1-\exp\left(\frac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)\right).
\end{aligned} \tag{3.19}$$

The optimal solution of (3.19) can be further analyzed as

$$x_{jk}\left(m_{jk}\right)=\frac{\sum_{i=1}^{c}u_{ij}^{m}\omega_{ik}^{\beta}\exp\left(\frac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)v_{ik}}{\sum_{i=1}^{c}u_{ij}^{m}\omega_{ik}^{\beta}\exp\left(\frac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)}. \tag{3.20}$$

Through the Lagrange multiplier method, on the basic of on the objective function (3.19), following updated formulas of membership degree, clustering center and weight matrix can be obtained :

$$u_{ij}=\left(\sum_{i=1}^{c}\frac{\sum_{k=1}^{l}\omega_{ik}^{\beta}\left(1-\exp\left(\frac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)\right)}{\sum_{k=1}^{l}\omega_{rk}^{\beta}\left(1-\exp\left(\frac{-\left\|x_{jk}\left(m_{jk}\right)-v_{rk}\right\|^{2}}{\sigma^{2}}\right)\right)}\right)^{\frac{1}{1-m}}, \tag{3.21}$$

$$\omega_{ik}^{\beta}=\left[\sum_{t=1}^{l}\frac{\sum_{j=1}^{n}u_{ij}^{m}\cdot\left(1-\exp\left(\frac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)\right)}{\sum_{j=1}^{n}u_{ij}^{m}\cdot\left(1-\exp\left(\frac{-\left\|x_{jt}\left(m_{jt}\right)-v_{rt}\right\|^{2}}{\sigma^{2}}\right)\right)}\right]^{\frac{1}{1-\beta}}, \tag{3.22}$$

$$v_{ik}=\begin{cases}0 & ,\ if\ \ \omega_{ik}^{\beta}=0\\ \dfrac{\sum_{j=1}^{n}u_{ij}^{m}\exp\left(\dfrac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)\cdot x_{jk}}{\sum_{j=1}^{n}u_{ij}^{m}\exp\left(\dfrac{-\left\|x_{jk}\left(m_{jk}\right)-v_{ik}\right\|^{2}}{\sigma^{2}}\right)} & ,\ if\ \ \omega_{ik}^{\beta}\neq 0\end{cases} \tag{3.23}$$

The main steps of OCS - WKFCM algorithm are :

(1) Initialization: Set the fuzzy parameter m, number of clusters $c\left(2\le c\le\sqrt{n}\right)$, utmost allowed iterations G, iterative convergence threshold ε, missing attribute matrix $\tilde{X}_M^{(0)}$, membership matrix $U^{(0)}$ combined with the constraint conditions.

(2) Updating the cluster center matrix: when the algorithm performs L (L = 1, 2, …) iterations, cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (3.23).

(3) Calculate the membership matrix: according to $V^{(l)}$, and (3.21) solving membership matrix $U^{(l)}$.

(4) Calculate the weight matrix: according to $V^{(l)}$, and (3.22) to solve the weight matrix.

(5) Update the missing value: calculate the missing value $\tilde{X}_M^{(l)}$ according to the cluster center matrix $V^{(l)}$ and membership partition matrix $U^{(l)}$and (3.20).

(6) Iteration termination: when the iteration count approaches $L = G$, or $\forall i$, k, $\max\left|u_{ik}^{(l)}-u_{ik}^{(l-1)}\right|<\varepsilon$, OCS - WKFCM algorithm iteration stop, the output U and V.

3.2.2 Feature weighted kernel FCM clustering with NPS

NPS - WKFCM divides the sample data x_{jk} into two parts, the complete attribute part $x_{jk}(o_{jk})$ and the missing attribute part $x_{jk}(m_{jk})$, then $x_{jk}(o_{jk})\cup x_{jk}(m_{jk}) = x_{jk}$ and $x_{jk}(o_{jk})$ remain unchanged in the clustering process. The filling method of missing attributes in NPS - WKFCM is as follows (3.24).

$$x_{jk}\left(m_{jk}\right)=v_{ik}=\begin{cases}0 & ,if\ \ \omega_{ik}^{\beta}=0\\ \dfrac{\sum_{j=1}^{n}u_{ij}^{m}x_{jk}}{\sum_{j=1}^{n}u_{ij}^{m}} & ,if\ \ \omega_{ik}^{\beta}\neq 0\end{cases},\ D_{ij}=\min\left\{D_{1j},D_{2j},\cdots,D_{cj}\right\} \tag{3.24}$$

Similar to OCS - WFCM, only (3.20) needs to be replaced with (3.24) when updating the missing missing attributes.

The main steps of NPS - WFCM algorithm are :

(1) Initialization: Set the fuzzy parameter m, number of clusters $c\left(2\le c\le\sqrt{n}\right)$, utmost allowed iterations G, iterative convergence threshold ε, missing attribute matrix $\tilde{X}_M^{(0)}$, membership matrix $U^{(0)}$ combined with the constraint conditions.

(2) Updating the cluster center matrix: when the algorithm performs L ($L = 1, 2, \ldots$) iterations, cluster center $V^{(l)}$ is updated according to $U^{(l-1)}$ and (3.23).

(3) Calculate the membership matrix: according to $V^{(l)}$, and (3.21) solving membership matrix $U^{(l)}$.

(4) Calculate the weight matrix: according to $V^{(l)}$, and (3.22) to solve the weight matrix.

(5) Update the missing value: calculate the missing value $\tilde{X}_M^{(l)}$ according to cluster center matrix $V^{(l)}$ and membership partition matrix $U^{(l)}$ and (3.24).

(6) Iteration termination: when the iteration count approaches $L = G$, or $\forall i, k$, $\max\left|u_{ik}^{(l)} - u_{ik}^{(l-1)}\right| < \varepsilon$, NPS - WKFCM algorithm iteration stop, the output U and V.

3.2.3 The complexity of WKFCM

An algorithm requires analysis of time complexity and space complexity. The complexity of OCS - WKFCM and NPS - WKFCM is mainly generated by clustering . In the clustering process, the number of iterations t, the number of clusters c, the dimension of sample data l, the number of data samples n will affect the time complexity of the algorithm. Considering the worst case, the time complexity of FCM clustering algorithm is $O(Tcnl)$. In the actual calculation process, a certain amount of storage space is needed to store data needed for clustering center matrix, weight matrix, the distance between sample data points, etc. Therefore, in order to store sample data, clustering center, weight matrix, and membership matrix, the space complexity is $O(nc + nl + 2cl)$.

4. Experimental evaluation

In order to verify the superiority of the proposed OCS - WFCM, NPS - WFCM, OCS - WKFCM, and NPS - WKFCM algorithms in clustering incomplete data, experiments are conducted in this section to validate them in several datasets, respectively. The dataset description and experimental steps design are described in the following subsections.

4.1 Dataset

The UCI database is a proposed database for machine learning by the University of California Irvine (UCI)[33]. The UCI dataset is a commonly used standard test dataset. Nine real datasets were selected from them as experimental datasets, and their details are shown in Table 1.

4.2 Experimental settings

For different datasets, the number of categories for clustering of WFCM and WKFCM models is different and needs to be determined according to the relevant attributes in different datasets. The parameters of the clustering algorithm are set uniformly. The maximum number of iterations is 200, the termination threshold is 0.0001, and the fuzzy index is 2.

To make the incomplete data generated in the experiments closer to reality, the data are processed by the random discard method, which uses different proportions set manually for the complete data to be lost randomly. Thus, an incomplete data set was generated. In this research, the missing proportions are taken as 5%, 10%, 15% and 20%. The rules for generating missing data attributes for incomplete datasets are as follows,

(1) In an incomplete dataset, the attribute values of sample data cannot all be lost. If the

dataset is n - dimensional, then at most n - 1 attributes are lost from the incomplete data, and at least one attribute must be present in the incomplete data.

(2) In an incomplete dataset, at least one complete attribute value exists for any one - dimensional attribute, i.e., the attribute column of the dataset cannot be empty to ensure the reliability of the valuation.

Each clustering algorithm performs 100 simulation experiments in each dataset with different missing proportions, and the obtained experimental results are averaged, thus reducing the chance of the experiments and the experimental errors.

4.3 Evaluation Criteria

Currently, there is no uniform evaluation index for the degree of merit of clustering algorithms. Therefore, in this work, the experimental algorithm is chosen to be evaluated from three perspectives: accuracy (Acc), iteration number, and external evaluation indexes concerning relevant literature. Among them, the external evaluation indexes are Normalized Mutual information (NMI), Rand Index (RI) and F_1 - score. The formulas are shown in Table 2.

In Table 2, matrix G represents the actual classification of the samples and T represents the fuzzy division of the clustering algorithm. $MI(G,T)$ is the mutual information of matrices G and T, $H(G)$ and $H(T)$ are the information entropy of matrices G and T, respectively. The set of sample pairs in G that are in the same cluster is denoted by X, and the set of sample pairs in G that are not in the same cluster is denoted by Z. The fuzzy set of sample pairs in T that are in the same cluster is denoted by Y, and the fuzzy set of sample pairs in T that are not in the same cluster is denoted by V. Then, in the above equation, $a=|X \cap Y|$, $b=|X \cap V|$, $c=|Z \cap Y|$, $d=|Z \cap V|$.

4.4 Experimental analysis

The missing treatment is performed on the nine datasets mentioned in Section 4.1, and the four optimized improvement algorithms proposed in this research are run. The results are experimentally compared with seven classical incomplete data clustering algorithms[34] and analyzed and described based on evaluation criteria.

To evaluate the advantages and disadvantages of the algorithms from an overall perspective, the mean values of the evaluation indexes of the 11 algorithms under the four missing ratios are taken, and the results are shown in Tables 3, 4, 5, 6, and 7.

The average ACC of the 11 algorithms with different missing rates in different datasets is reflected in Table 3. The table shows that the OCS - WFCM and NPS - WFCM algorithms proposed in this work based on feature weighting improvement have higher accuracy than the seven classical clustering algorithms under different missing rates in each dataset. The proposed OCS - WKFCM and NPS - WKFCM algorithms based on feature weighting and kernel function improvement have the highest accuracy in all datasets. The accuracy of the clustering algorithms is the most direct representation of the accuracy. This result shows that the incorporation of feature weighting and kernel methods can improve the clustering performance of the FCM algorithm for incomplete data and make it have higher clustering accuracy.

Tables 4, 5, and 6 show the calculation of three external evaluation metrics, NMI, F - score, and RI. The four optimization algorithms achieved the optimum in all datasets. Among them, the OCS - WFCM and NPS - WFCM algorithms are only slightly worse than the others in Bupa and Haberman datasets, and the OCS - WKFCM and NPS - WKFCM are better than the OCS - WFCM and NPS - WFCM algorithms in all datasets. Due to the random nature of missing processing, it may make too many missing features of a certain attribute, which is not conducive to updating the feature weights of OCS - WFCM and NPS - WFCM algorithms. Therefore, on the whole, the clustering accuracy of OCS - WFCM and NPS - WFCM algorithms is still better than that of the seven classical algorithms. Meanwhile, the introduction of the kernel method will alleviate the influence of feature attributes on the clustering accuracy and improve the prediction accuracy, which makes the external evaluation indexes of OCS - WFCM and NPS - WFCM algorithms better than those of OCS - WFCM and NPS - WFCM algorithms.

Table 7 shows the average number of iterations of 11 algorithms. This index mainly reflects the convergence speed of the algorithms. From the table, it can be obtained that all algorithms can reach a stable convergence state. However, in about 2/3 of the datasets, the iterations of OCS - WFCM and NPS - WFCM is significantly higher than that of the seven classical algorithms. In all the datasets, the iterations of OCS - WKFCM and NPS - WKFCM are lower than that of OCS - WFCM and NPS - WFCM. The results show that the feature weights will increase the number of iterations in some datasets, while the kernel method will significantly reduce the number of iterations. While the kernel method will significantly reduce the number of iterations and improve the solving speed of the algorithm.

Compared with AVER - FCM, ZERO - FCM, and KNN - FCM algorithms, the four algorithms proposed in this research are superior. AVER - FCM, ZERO - FCM, and KNN - FCM fill the missing attributes with 0 values, sample mean values, and mean values of K neighboring samples, respectively, and then run the FCM algorithm. 0 - value interpolation and mean interpolation will make the samples lose a large amount of data information, which is the most basic interpolation strategy. The KNN algorithm is extremely data - dependent, and individual data anomalies will affect the effect of the whole clustering. The traversal mechanism of the KNN algorithm is prone to dimensional disasters on large datasets. At the same time, the above algorithms fill in the missing data in the sample and then perform clustering. The data filling algorithm will have certain errors in filling accuracy and cannot accurately represent the missing data, and then clustering on the filled data set will have even lower clustering accuracy. The four improved algorithms are based on OCS - FCM and NPS - FCM algorithms, which dynamically update the incomplete data during the clustering iterations and organically combine clustering and interpolation. This avoids the secondary accuracy reduction caused by the algorithms to some extent and has better robustness.

Compared with the WDS - FCM, PDS - FCM, OCS - FCM, and NPS - FCM algorithms, the OCS - WFCM and NPS - WFCM algorithms are superior. The WDS - FCM algorithm discards incomplete data samples, which will have a greater impact on the clustering results in the case of high missing data samples and reduce the overall sample size. The PDS - FCM algorithm is an improvement of the WDS - FCM algorithm but does not deal with missing attributes. Both algorithms do not treat missing attributes, and the data information is wasted. Its information value is not maximized, and the clustering results are unsatisfactory. The traditional OCS - FCM and NPS - FCM do not consider the role played by different features in the clustering process and treat all features equally. In contrast, the OCS - WFCM and NPS - WFCM algorithms assign weights to different features on this basis. At the same time, dynamic adjustments are made during the iterative process to minimize the influence of outlier points in the sample on the clustering center. This results in a better clustering effect in most of the datasets.

Based on the OCS - WFCM and NPS - WFCM algorithms, a greater improvement is made in this work. The OCS - WKFCM and NPS - WKFCM algorithms are proposed. The above modification introduces the kernel method into the FCM algorithm for incomplete data and solves the nonlinear separable problem between clusters and clusters in complex data. The number of iterations of the algorithm is substantially reduced based on the improved clustering, which makes the algorithm perform better.

Figures 1, 2, 3, and 4 show the specific performance of the evaluation criteria, ACC, NMI, F - score, and RI, respectively, in different datasets and missing proportions. Among them, ZERO - FCM, AVER - FCM, KNN - FCM, WDS - FCM, and PDS - FCM only have good accuracy in partial datasets and fluctuate greatly in some missing proportions. Compared with the above five algorithms, OCS-FCM and NPS-FCM algorithms are not optimal in all cases, but the clustering accuracy starts to maintain stability. Compared with the OCS - FCM and NPS - FCM algorithms, the proposed four algorithms all showed significant improvement in clustering accuracy. This indicates that the optimization algorithms continue the advantages of the original algorithms and still have better robustness. Meanwhile, the histogram distribution in the figure shows that the OCS - WKFCM algorithm possesses higher evaluation criteria values and better clustering accuracy for low missing rates of only 5% - 10%. The NPS - WKFCM algorithm provides higher accuracy for high missing rates of 15 - 20%.

Considered from the perspective of interpolation methods, the OCS - WKFCM algorithm takes into account the information of missing data attributes. It can still maintain the excellent performance of the FCM algorithm as the missing rate increases and keep the clustering accuracy stable. However, the OCS - WKFCM algorithm requires repeated iterations to update the missing attribute values, which can make the number of iterations of the algorithm increase significantly. The NPS - WKFCM algorithm updates the missing values by comparing them with the clustering centers derived from the current iteration. It no longer requires repeated iterations and reduces the difficulty of solving. The experimental comparison reveals that its accuracy is better with a high missing rate.

5. Conclusion

For incomplete data clustering, a new generalized fuzzy clustering framework incorporating feature weights and kernel methods is developed in this work. The four improved algorithms specifically involved are WFCM - OCS, WFCM - NPS, WKFCM - OCS, and WKFCM - NPS. The experimental results validate the effectiveness of the proposed framework and show that the optimized algorithms are superior in the clustering of incomplete data. Meanwhile, the following conclusions are drawn:

(1) The improvement based on feature weights can improve the clustering precision of the FCM algorithm in most incomplete datasets.However, it also dramatically raises the iteration number and increase the complexity of the algorithm.

(2) On the basis of the OCS - WFCM and NPS - WFCM algorithms, the data are mapped by the kernel method for high latitude mapping can effectively improve the clustering accuracy, and does not influence iteration number significantly.

(3) The OCS - WKFCM algorithm has higher clustering precision at low missing rate of 5% - 10%, while the NPS - WKFCM performs better at high missing rate of 15 - 20%.

(4) In the future, the thoughts of intelligent optimization and neural networks can be applied to the incomplete data clustering to obtain better clustering performance.

Conflicts of Interest

The authors declare that they have no competing interests.

Funding

The authors received no funding for this work.

References

[1] Sinaga K P, Yang M S. Unsupervised K-means clustering algorithm[J]. IEEE access, 2020, 8: 80716-80727.

[2] Zadeh L A, Klir G J, Yuan B. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers[M]. World Scientific, 1996.

[3] Z. Ma, Z. Liu, C. Luo, L. Song, Evidential classification of incomplete instance based on k-nearest centroid neighbor[J]. Journal of Intelligent & Fuzzy Systems, 2021, 41(6): 7101-7115.

[4] Hayati Rezvan P, Lee K J, Simpson J A. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research[J]. BMC medical research methodology, 2015, 15(1): 1-14.

[5] Z. Ma, Z. Liu, Y. Zhang, L. Song, J. He, Credal transfer learning with multi-estimation for missing data[J]. IEEE Access, 2020, 8, 70316 - 70328.

[6] Hathaway R J, Bezdek J C. Fuzzy c-means clustering of incomplete data [J]. IEEE transactions on systems, man, and, cybernetics-part B: Cybernetics, 2001, 31(5):735-744.

[7] J. K. Dixon, "Pattern Recognition with Partly Missing Data," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 10, pp. 617-621, Oct. 1979, doi: 10.1109/TSMC.1979.4310090.

[8] Li T , Zhang L , Lu W , et al. Interval kernel Fuzzy C-Means clustering of incomplete data[J]. Neurocomputing, 2017, 237(MAY10):316-331.

[9] Najib F M , Ismail R M , Badr N L , et al. Clustering based approach for incomplete data streams processing[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(1):1-15.

[10] Meng J , Li C , Mo H , et al. Measuring galaxy abundance and clustering at high redshift from incomplete spectroscopic data: Tests on mock catalogs and application to zCOSMOS[J]. 2020.

[11] Villuendas-Rey Y , Barroso-Cubas E , Camacho-Nieto O , et al. A General Framework for Mixed and Incomplete Data Clustering Based on Swarm Intelligence Algorithms[J]. Mathematics, 2021, 9.

[12] Shi, Maolin,Wang, Zihao.A Study of Support Vector Regression-Based Fuzzy c-Means Algorithm on Incomplete Data Clustering[J].Journal of Advanced Computatioanl Intelligence and Intelligent Informatics,2022,26(4 TN.157):483-494
[13] Dan Li,Hong Gu,Liyong Zhang. A fuzzy c -means clustering algorithm based on nearest-neighbor intervals for incomplete data[J]. Expert Systems With Applications,2010,37(10).
[14] G. Doquire, M. Verleysen, Feature selection with missing data using mutual information estimators, Neurocomputing. 90 (2012) 3-11.
[15] G. Tutz, S. Ramzan, Improved methods for the imputation of missing data by nearest neighbor methods, Comput. Stat. Data Anal. 90 (2015) 84-99.
[16] C. F. Tsai, M. L. Li, W. C. Lin, A class center based approach for missing value imputation, Knowledge-Based Systems.151(2018) 124-135.
[17] D. Alexandra Williams, Benjamin Nelsen, Candace Berrett, et al. A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data, Environmental Modelling & Software,102(2018):172-184.
[18] Baligh A H, Chen Q, Xue B, Zhang M G. A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data[J]. Soft Computing,2021(prepublish).
[19] E. Eirola, A. Lendasse, V. Vandewalle, C. Biernacki, Mixture of Gaussians for distance estimation with missing data, Neurocomputing. 131 (2014) 32-42.
[20] Bashir F, Wei H L. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm [J]. Neurocomputing, 2018, 276(1): 23-30.
[21] Hung T, Cristina T. Cluster analysis and outlier detection with missing data. 2020.
[22] Ravi V, Krishna M. A new online data imputation method based on general regression autoassociative neural network [J]. Neurocomputing, 2014, 138(1):106-113.
[23] Rancoita P M V, Zaffalon M, Zucca E, et al. Bayesian network data imputation withapplication to survival tree analysis [J]. Computational Statistics and Data Analysis, 2016, 93(1):373-387
[24] Nishanth K J, Ravi V. Probabilistic neural network based categorical data imputation [J]. Neurocomputing, 2016, 218(1):17-25.
[25] Sovilj D, Eirola E, Miche Y, et al. Extreme learning machine for missing data using multiple imputations [J]. Neurocomputing, 2019, 174(1):220-231.
[26] Cao Truong Tran,Mengjie Zhang,Peter Andreae,Bing Xue,Lam Thu Bui. Improving performance of classification on incomplete data using feature selection and clustering[J]. Applied Soft Computing Journal,2018,73.
[27] Desarbo W.S., Carroll J.D., Clark L.A., et al. Synthe-sized Clustering: A method for amalgamating clustering bases with differential weighting variables[J]. Psychometrika, 1984,49:57-78.
[28] Makarenkov V., Legendre P.. Optimal variable weighting for ultrametric and additive trees and k-means partitioning: methods and software[J]. J. Classification, 2001, 18:245-271.
[29] Zhang L, Da ZW, Jiao LC (2002) Kernel clustering algorithm. Chinese J Comput 25:587–590.
[30] Modha D.S., Spangler W.S. Feature weighting in k-means clustering[J]. Machine Learning, 2003,52:217-237.
[31] Thi Ngoc Chau Vo,Hua Phung Nguyen,Thi Ngoc Tran Vo. Making kernel-based vector quantization robust and effective for incomplete educational data clustering[J]. Vietnam Journal of Computer Science,2016,3(2).
[32] Askari S. Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development[J]. Expert Systems With Applications,2021,165:1-27
[33] Bache K, Lichman M (2013) UCI machine learning repository. University of California, School of Information and Computer

Science. http://archive.ics.uci.edu/ml

[34] Hong Shi,Pingxin Wang,Xin Yang,Hualong Yu. An Improved Mean Imputation Clustering Algorithm for Incomplete Data[J]. Neural Processing Letters,2020(prepublish).

Figure 1

Figure. 1. Histogram of ACC averages in 9 datasets with different missing values

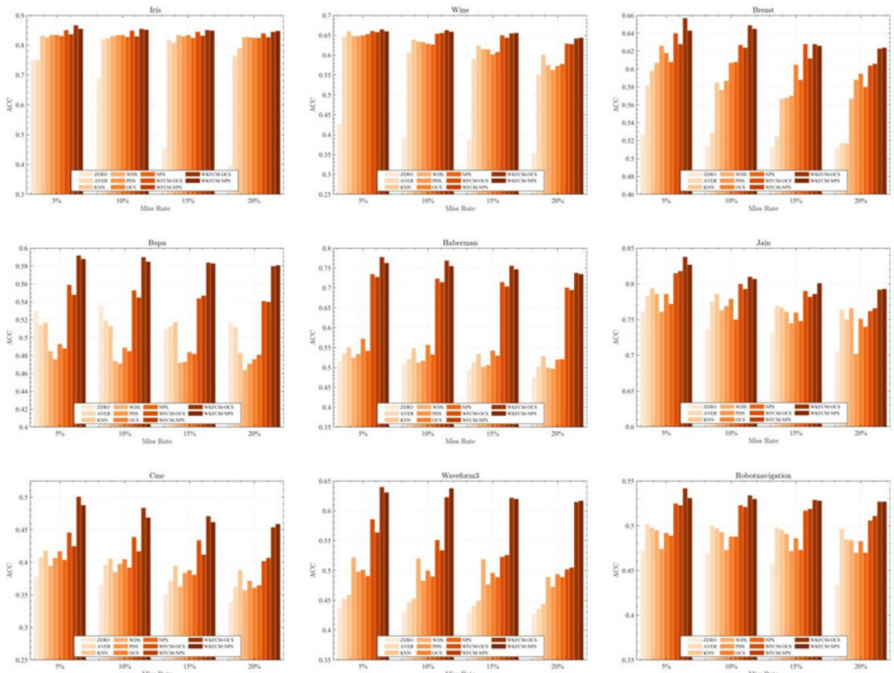

Figure. 1. Histogram of ACC averages in 9 datasets with different missing values

Figure 2

Figure. 2. Histogram of NMI averages in 9 datasets with different missing values

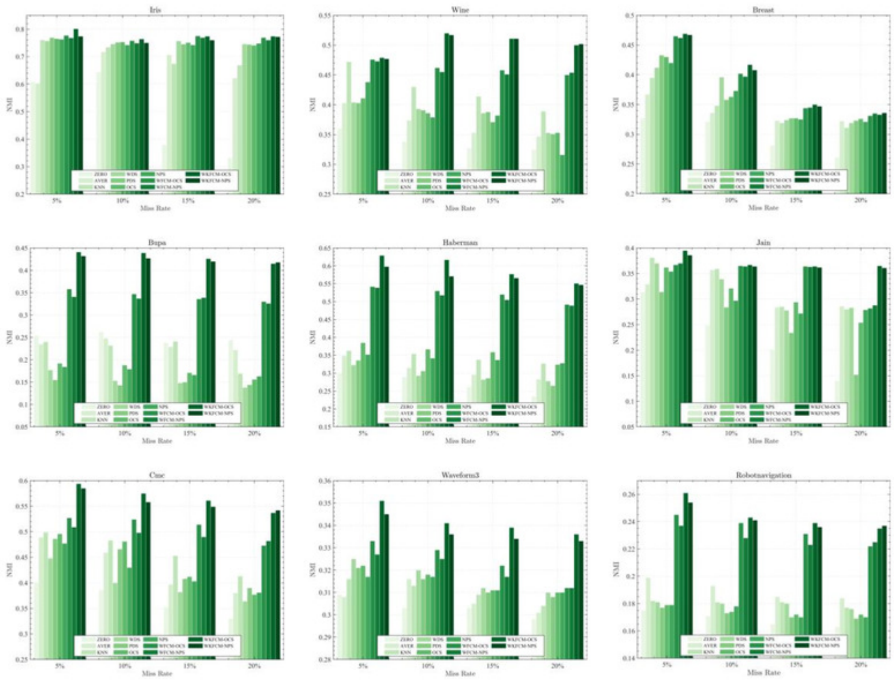

Figure. 2. Histogram of NMI averages in 9 datasets with different missing values

Figure 3

Figure. 3. Histogram of F - score averages in 9 datasets with different missing values

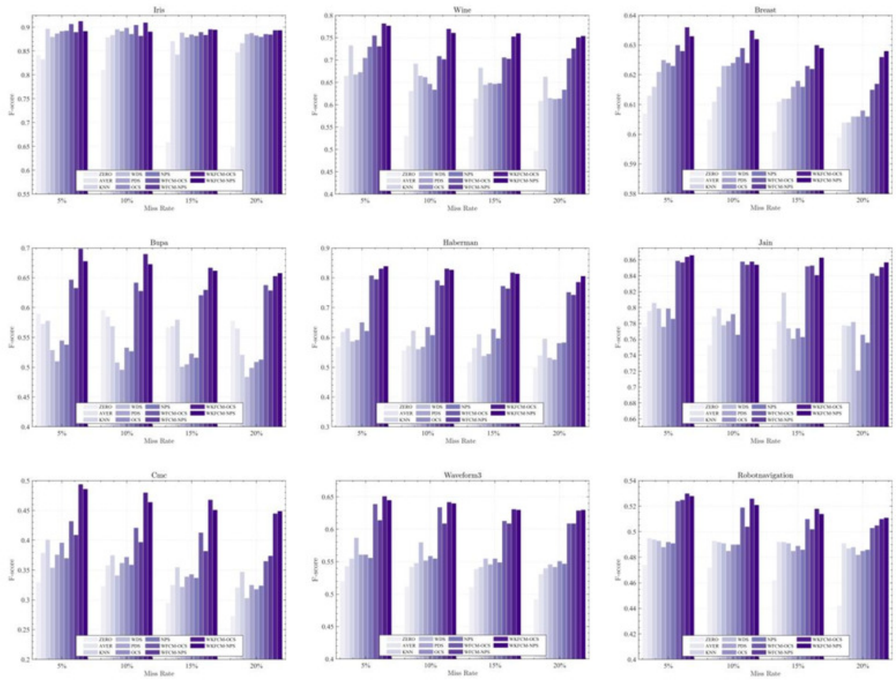

Figure. 3. Histogram of F - score averages in 9 datasets with different missing values

Figure 4

Figure. 4. Histogram of RI averages in 9 datasets with different missing rates

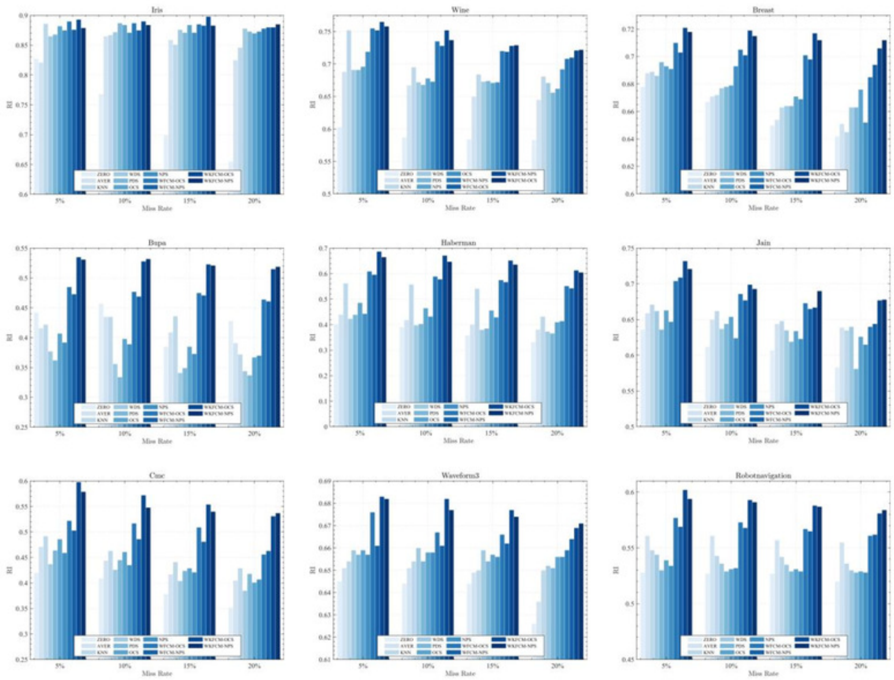

Figure. 4. Histogram of RI averages in 9 datasets with different missing rates

Table 1(on next page)

Table 1. Datasets used in our experiments

Table 1. Datasets used in our experiments

Dataset	Instance	Features	Classes
Iris	150	4	3
Wine	178	13	3
Breast	277	9	2
Bupa	345	6	2
Haber Man	306	3	2
Jain	373	2	2
Cmc	1473	9	3
Waveform3	5000	21	3
Robotnavigation	5456	24	4

Table 2(on next page)

Table 2. external evaluation indicators and formulas

Table 2. external evaluation indicators and formulas

External evaluation indicators	Formula
NMI	$NMI(G,T) = \frac{2MI(G,T)}{H(G)+H(T)}$
F_1 - Score	$F_1 - score = 2\frac{a \times c}{a+c}$
RI	$RI = \frac{a+b}{a+b+c+d}$

Table 3(on next page)

Table 3. ACC averages of different algorithms in 9 datasets with different missing rates

1 **Table 3. ACC averages of different algorithms in 9 datasets with different missing rates**

Dataset / Methods	ACC								
	Iris	Wine	Breast	Bupa	Haber man	Jain	Cmc	Wave form3	Robot navigation
ZERO	0.574	0.390	0.517	0.523	0.498	0.734	0.358	0.431	0.458
AVER	0.789	0.598	0.538	0.514	0.519	0.773	0.385	0.444	0.499
KNN	0.813	0.631	0.567	0.507	0.540	0.774	0.402	0.452	0.494
WDS	0.830	0.618	0.580	0.474	0.510	0.767	0.375	0.513	0.491
PDS	0.832	0.615	0.592	0.473	0.514	0.744	0.390	0.483	0.472
OCS	0.827	0.617	0.596	0.484	0.532	0.753	0.385	0.490	0.480
NPS	0.832	0.613	0.606	0.486	0.548	0.769	0.393	0.498	0.487
WFCM-OCS	0.832	0.646	0.618	0.545	0.711	0.790	0.415	0.532	0.519
WFCM-NPS	0.846	0.648	0.625	0.549	0.719	0.792	0.430	0.541	0.518
WKFCM-OCS	0.851	0.655	0.635	0.584	0.750	0.807	0.469	0.626	0.529
WKFCM-NPS	0.855	0.656	0.639	0.587	0.760	0.807	0.477	0.625	0.533

2

Table 4 (on next page)

Table 4. NMI averages of different algorithms in 9 datasets with different missing rates

Table 4. NMI averages of different algorithms in 9 datasets with different missing rates

Dataset / Methods	NMI								
	Iris	Wine	Breast	Bupa	Haber man	Jain	Cmc	Wave form3	Robot navigation
ZERO	0.491	0.337	0.298	0.250	0.271	0.226	0.367	0.303	0.169
AVER	0.662	0.369	0.362	0.234	0.311	0.314	0.431	0.308	0.190
KNN	0.709	0.426	0.368	0.221	0.346	0.326	0.462	0.311	0.180
WDS	0.752	0.384	0.388	0.154	0.294	0.318	0.399	0.317	0.179
PDS	0.753	0.383	0.385	0.148	0.298	0.246	0.437	0.314	0.172
OCS	0.749	0.379	0.385	0.173	0.340	0.308	0.423	0.314	0.174
NPS	0.753	0.380	0.386	0.177	0.359	0.300	0.442	0.315	0.174
WFCM-OCS	0.761	0.458	0.410	0.336	0.513	0.344	0.495	0.320	0.228
WFCM-NPS	0.766	0.461	0.411	0.343	0.521	0.346	0.510	0.324	0.234
WKFCM-OCS	0.777	0.502	0.414	0.424	0.571	0.372	0.559	0.337	0.242
WKFCM-NPS	0.775	0.502	0.417	0.430	0.594	0.368	0.567	0.341	0.245

Table 5 (on next page)

Table 5. F - score averages of different algorithms in 9 datasets with different missing rates

Table 5. F - score averages of different algorithms in 9 datasets with different missing rates

Dataset / Methods	F - score								
	Iris	Wine	Breast	Bupa	Haber man	Jain	Cmc	Wave form3	Robot navigation
ZERO	0.740	0.528	0.603	0.583	0.536	0.750	0.305	0.509	0.462
AVER	0.858	0.630	0.610	0.573	0.574	0.787	0.346	0.539	0.493
KNN	0.873	0.693	0.612	0.562	0.615	0.800	0.395	0.546	0.491
WDS	0.888	0.648	0.616	0.506	0.555	0.783	0.330	0.567	0.491
PDS	0.886	0.649	0.618	0.503	0.558	0.760	0.351	0.550	0.485
OCS	0.886	0.661	0.618	0.524	0.603	0.783	0.348	0.552	0.488
NPS	0.890	0.653	0.619	0.528	0.624	0.768	0.357	0.557	0.489
WFCM-OCS	0.885	0.716	0.623	0.630	0.769	0.853	0.391	0.610	0.509
WFCM-NPS	0.897	0.718	0.624	0.637	0.781	0.851	0.408	0.624	0.514
WKFCM-OCS	0.893	0.763	0.630	0.668	0.822	0.854	0.463	0.636	0.519
WKFCM-NPS	0.903	0.764	0.632	0.677	0.817	0.860	0.472	0.638	0.521

Table 6 (on next page)

Table 6. RI averages of different algorithms in 9 datasets with different missing rates

Table 6. RI averages of different algorithms in 9 datasets with different missing rates

Dataset / Methods	RI								
	Iris	Wine	Breast	Bupa	Haber man	Jain	Cmc	Wave form3	Robot navigation
ZERO	0.737	0.589	0.659	0.428	0.371	0.609	0.390	0.640	0.525
AVER	0.843	0.662	0.666	0.413	0.410	0.648	0.434	0.647	0.559
KNN	0.862	0.703	0.667	0.416	0.523	0.654	0.456	0.652	0.542
WDS	0.873	0.677	0.673	0.355	0.394	0.644	0.413	0.658	0.536
PDS	0.875	0.672	0.675	0.346	0.398	0.620	0.438	0.654	0.529
OCS	0.873	0.689	0.676	0.381	0.430	0.644	0.431	0.657	0.531
NPS	0.880	0.677	0.680	0.389	0.454	0.627	0.444	0.658	0.533
WFCM-OCS	0.878	0.727	0.699	0.469	0.571	0.676	0.483	0.662	0.566
WFCM-NPS	0.885	0.729	0.700	0.475	0.581	0.674	0.501	0.667	0.569
WKFCM-OCS	0.883	0.737	0.714	0.526	0.638	0.693	0.551	0.676	0.589
WKFCM-NPS	0.890	0.741	0.716	0.525	0.656	0.695	0.564	0.678	0.591

Table 7 (on next page)

Table 7. Iterations averages of different algorithms in 9 datasets with different missing rates

Table 7. Iterations averages of different algorithms in 9 datasets with different missing rates

Dataset / Methods	Iterations								
	Iris	Wine	Breast	Bupa	Haberman	Jain	Cmc	Waveform3	Robot navigation
ZERO	60.20	68.90	27.94	51.45	57.48	31.47	37.49	28.96	27.97
AVER	33.05	43.02	28.67	36.38	24.41	25.24	24.54	27.15	27.06
KNN	41.33	43.70	28.63	36.64	37.79	29.54	35.51	25.37	29.24
WDS	26.40	47.39	33.13	37.48	25.65	19.72	23.48	26.99	25.88
PDS	26.75	41.49	28.04	38.30	26.24	29.94	24.40	31.40	24.79
OCS	33.63	55.56	25.09	43.85	27.20	29.14	26.38	34.98	27.37
NPS	28.75	52.67	27.32	39.65	26.66	27.98	25.74	35.63	25.37
WFCM - OCS	36.23	46.18	30.90	35.43	37.34	22.05	28.05	30.37	30.69
WFCM - NPS	31.20	42.75	29.10	34.35	36.93	19.02	27.33	26.61	28.00
WKFCM-OCS	34.85	42.30	29.23	34.00	30.64	18.94	26.85	28.59	27.05
WKFCM-NPS	29.25	41.13	27.08	32.25	26.99	17.74	25.31	24.49	24.45