Neural Processing Letters (2022) 54:3537–3550
https://doi.org/10.1007/s11063-020-10298-5

An Improved Mean Imputation Clustering Algorithm for Incomplete Data

Hong Shi[1] · Pingxin Wang[2,3] · Xin Yang[1] · Hualong Yu[1]

Published online: 2 July 2020

Abstract

There are many incomplete data sets in all fields of scientific studies due to random noise, data lost, limitations of data acquisition, data misunderstanding etc. Most of the clustering algorithms can not be used for incomplete data sets directly because objects with missing values need to be preprocessed. For this reason, this paper presents an improved mean imputation clustering algorithm for incomplete data based on partition clustering algorithm. In the proposed method, we divide the universe into two sets: the set of objects with non-missing values and the set of objects with missing values. Firstly, the objects with non-missing values are clustered by traditional clustering algorithm. For each object with missing values, we use the mean attribute's value of each cluster to fill the missing attribute's value based on the cluster results of the objects with non-missing values, respectively. Perturbation analysis of cluster centroid is applied to search the optimal imputation. The experimental clustering results on some UCI data sets are evaluated by several validity indexes, which proves the effectiveness of the proposed algorithm.

Keywords Incomplete data · Mean imputation · K-means · Validity index

✉ Pingxin Wang
pingxin_wang@hotmail.com

Hong Shi
RainbowOneHong@163.com

Xin Yang
yangxinzero@163.com

Hualong Yu
yuhualong@just.edu.cn

[1] School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212003, People's Republic of China

[2] School of Science, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, People's Republic of China

[3] College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050024, People's Republic of China

1 Introduction

Cluster analysis plays an indispensable role in data mining and machine learning [1–4]. It is widely used in different fields such as information granulation [5–7], image processing [8], bioinformatics [9], security assurance [10] etc. The primary task of clustering is to group a set of objects into multiple clusters, which can identify the internal structure of massive data. In this way, the dissimilarity of samples in the same cluster is lower than that of samples in different clusters. There are many different clustering algorithms in cluster analysis, and these existing algorithms can be roughly divided into hierarchical clustering and partition clustering [11]. In this paper, we mainly focus on partition clustering. K-means clustering algorithm [12] is one of typical partition clustering, which was introduced by Macqueen [13] in 1967. In k-means algorithm, the distance is used as the evaluation index of similarity. The closer the distance between two objects is, the greater the similarity. It is not stranger to use the k-means algorithm to deal with the clustering of complete datasets. How to handle a specific data set containing missing data based on the k-means algorithm, this is still an urgent problem to be solved. Therefore, in this paper, we propose an algorithm for processing the clustering problem of incomplete data sets.

The traditional clustering algorithm like k-means can not deal with the datasets containing missing values straightly. However, in the actual scenario, some values in the data set are missing due to random noise, data lost, limitations of data acquisition, data misunderstanding etc. These objects with missing values in a specific data set are generally referred to as incomplete data set. According to the theory proposed by Rubin et al. [20,21], the types of missing data can be classified as missing completely at random(MCAR), missing at random(MAR) and not missing at random(NMAR). Due to the emergence of missing data, k-means can not be applied to cluster the incomplete data sets directly. Therefore, how to deal with incomplete data sets is a problem to be solved in cluster analysis. In this paper, we consider the incomplete information system with missing completely at random(MCAR).

In the study of incomplete data clustering, the effective imputation method of missing values is the key to improve the accuracy of clustering result. There are many ways to fill incomplete data, for example, mean imputation, regression imputation, multiple imputation, hot-deck, cold-deck ect. [22]. EM algorithm [23] produces the maximum likelihood estimate value through iteration. The Most Common Attribute Value method [24] replaces the missing value with the most frequent attribute value. G. Doquire et al. [25] proposed the nearest-neighbor method based on mutual information to evaluate the missing data. J. Van Hulse et al. [26] used the mean values of k neighbors(complete data or incomplete data that have been imputed) with incomplete data to interpolate the missing data.

In order to solve the problem of incomplete data clustering, many scholars have proposed different clustering strategies. Hathaway and Bezdek [27] put forward four incomplete data clustering methods based on the fuzzy C-means clustering algorithm(FCM), which are Whole Data Strategy(WDS), Partial Distance Strategy(PDS), Optimal Completion Strategy(OCS) and Nearest Prototype Strategy(NPS), respectively. Zhang and Chen [28], in 2003, came up with some sorts of methods to solve the problem of clustering incomplete data by using kernel-based fuzzy c-means algorithm. On the basis of FCM algorithm, Li et al. [29] used the FCM clustering algorithm to process the incomplete data, and the premise is that the missing data is estimated by nearest neighbor interval. In Li et al. [30] proposed the attribute weighted FCM algorithm to solve incomplete data. In Li et al. [31] introduced the combination of genetic algorithm and FCM algorithm to get the clustering result and the estimated value of missing data. In Li et al. [32] studied a robust FCM clustering algorithm to deal with incomplete data.

In addition, besides FCM algorithm, there are other ways for incomplete data clustering. Su et al. [33] put forward the three-way decision clustering algorithm for incomplete data based on q-nearest neighbors. Shi et al. [34] introduced a clustering ensemble algorithm for mixed data. Rencently, Mesquita et al. [35] used a new technique called artificial neural networks with random weights to deal with incomplete data clustering. All the above results enrich the theories and models of incomplete data clustering.

This paper proposes the KM-IMI algorithm, which is a kind of clustering algorithm for incomplete data sets. Due to the k-means algorithm can't directly deal with incomplete data sets, the method of adding sample weights and analyzing perturbation distance of cluster centroid are introduced to achieve clustering result of incomplete data sets. Firstly, the incomplete data set was given, where the object with missing values is constrained by two conditions: (1) each original feature vector x_i retains at least one component; (2) each feature has at least one value present in the incomplete data set. Secondly, the k-means is used to process the set of objects with non-missing values to get clustering result. Thirdly, the set of objects with missing values are searching for the optimal imputation by adding sample weights and analyzing the change of cluster centroid. Finally, a kind of partition clustering algorithm is used to obtain the final clustering result.

The rests of this paper are organised as follows. Section 2 reviews k-means algorithm and incomplete information system. Section 3 introduces the KM-IMI algorithm. Section 4 reviews clustering performance measurement. Experiment results are reported in section 5.

2 Preliminaries

To facilitate the description of the proposed method, we introduce some basic concepts related to this paper, which include the k-means clustering algorithm and incomplete information system.

2.1 K-means Clustering Algorithm

The k-means algorithm was introduced by Macqueen [13] in 1967, which is a commonly cluster analysis method in data mining. It has been successfully applied in many fields like computer vision, market segmentation, astronomy, agriculture and geostatistics [19]. It has several advantages like: (1) the principle is uncomplicated and easy to implement. (2) the classic algorithm for solving clustering problems is simple and fast.(3) maintain scalability and high efficiency when dealing with large data sets. There are certain limitations for k-means algorithm: (1) the value of k has to be given in advance. (2) different initial clustering centroids lead to different clustering results. (3) it can easily fall into local optimal rather than global optimal results. (4) it is sensitive to noise and outliers. Some scholars focus on solving the shortcomings of k-means algorithm. Such as Yu et al. [16] proposed two improved algorithms, mainly aiming at the fact that the k-means is vulnerable to outliers and noisy data and also susceptible to initial cluster centers. Franti et al. [17] discussed the problem of reasonable initialization and iteration times of the k-means to improve its performance. It is the most widely used algorithm in clustering algorithms due to its simplicity and efficiency, although it has some drawbacks.

At present, many clustering algorithms are improved on the basis of the k-means, like k-means++ [14], k-prototype and k-mediods [15]. Chao [18] put forward an algorithm, which primarily combined the discrimination k-means and the spectral clustering to improve clus-

Algorithm 1. k-means Clustering Algorithm

```
   Input: Dataset: $\mathbf{U} = \{x_1, x_2, \dots, x_n\} \in \mathbf{R}^m$, number of clusters: $k$.
   Output: Clustering result: $\mathbb{C} = \{C_1, C_2, \dots, C_k\}$.
1. Randomly select $k$ objects from the dataset, where $k < n$ . Generally view these objects as the initial
   centroids $z_1, z_2, z_3, \dots, z_k$.
2. Assign each samples to one of $k$ centroids, according to the shortest distance principle. i.e., $C_i =
   \{x_j | d(x_j, z_i) \leq d(x_j, z_l), (l \neq i), j = 1, 2, \dots, n\}$.
3. When all samples have been assigned, recalculate the value of centroids. i.e., $z_i = \frac{\sum_{x_j \in C_i} x_j}{|C_i|}$, $(i =
   1, 2, \dots, k)$.
4. Repeat Step 2 and Step 3 until the centroids no longer change or satisfy some stop conditions.
5. Return $\mathbb{C} = \{C_1, C_2, \dots, C_k\}$.
```

tering performance and deal with high dimensional problem. The standard k-means algorithm has four steps: given that the data set $\mathbf{U} = \{x_1, x_2, \dots, x_n\}$ contains n objects, these objects have m attributes and the number of clusters is k; initialize the value of cluster center (specified or random); find the cluster of each object based on the shortest distance principle; recalculate the cluster centroid until the given convergence condition is met. Algorithm 1 gives the detailed process of k-means clustering algorithm.

2.2 An Incomplete Information System

The information system also known as the knowledge representation system. It can be represented as $S = \{U, A, V, f\}$ or abbreviated as $S = \{U, A\}$. $U = \{x_1, x_2, \dots, x_n\}$ is a finite non-empty set, called a universe, where n denotes the number of samples. $A = \{a_1, a_2, \dots, a_m\}$ is a finite set of non-empty attributes, where m represents the number of object features. $V = \{V_1, V_2, \dots, V_m\}$ is the set of object attribute values, V_i is the possible feature values of a_i. f is an information function, $f : V_{ik} = f(x_i, a_k) \in V_k$, V_{ik} represents the value of the sample x_i on the feature a_k. For example, the x_i is the ith object with m features, namely, $x_i = \{x_i^1, x_i^2, \dots, x_i^m\}$, where x_i^l $(l < m, i \leq n)$represents the value of lth feature of sample x_i.

When some attribute values are missing, the information system S is called as the incomplete information system. This paper mainly discusses the incomplete information system with missing completely at random (MCAR). An example of the incomplete information system is shown in Table 1, which includes 6 samples and each sample has 4 attributes. Missing values are represented by *.

Table 1 An example of the incomplete information system

Objects	Attributes			
	a_1	a_2	a_3	a_4
x_1	3	2	5	9
x_2	4	3	$*$	7
x_3	$*$	$*$	8	11
x_4	6	3	4	7
x_5	12	$*$	$*$	11
x_6	14	3	$*$	$*$

3 An Improved Mean Imputation Incomplete Data Clustering Algorithm

Suppose that $\mathbf{U} = \{x_1, x_2, \ldots, x_n\}$ is an incomplete data set with n objects. The object with missing values is constrained by two conditions: each original feature vector x_i retains at least one component and each feature has at least one value present in the incomplete data set. Naturally, the data set $\mathbf{U}$ can be classified into two disjoint subsets. One set $\mathbf{U_W}$ requires that each object x with non-missing values called complete data set. In contrast, another set $\mathbf{U_M}$ is called an incomplete data set, where $\mathbf{U_W} \cup \mathbf{U_M} = \mathbf{U}$. Algorithm 1 processes the object in set $\mathbf{U_W}$ to get result $\mathbb{C}$. Algorithm 2 is used to fill the object with missing values in $\mathbf{U_M}$ to obtain the result $\mathbb{C}'$. Finally, the final clustering result $\mathbb{C}_{final}$ acquired by the k-means algorithm. The specific process of our algorithm is shown in Fig. 1.

In the study of incomplete data clustering, effectively fill in missing values of object is the key to improve the accuracy of clustering result. There are many methods to impute missing values, like mean imputation, regression imputation, multiple imputation etc. In this section, we present a kind of an improved mean imputation clustering algorithm for incomplete data. It can be briefly divided into four phases: (1) classifying the data set $\mathbf{U}$ into two disjoint subsets: the set of objects with non-missing values $\mathbf{U_W}$ and the set of objects with missing values $\mathbf{U_M}$. (2) clustering the object in set $\mathbf{U_W}$ through Algorithm 1. (3) filling the object with missing values in $\mathbf{U_M}$ through Algorithm 2. (4) getting the final clustering result by the k-means. The improved mean imputation incomplete data clustering algorithm based on k-means, shorted by KM-IMI, is described in Algorithm 2.

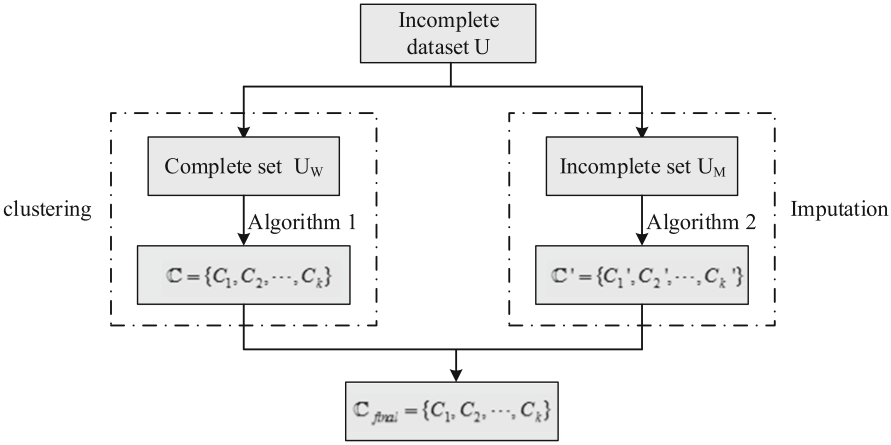
Incomplete dataset U
clustering
Complete set U_W
Algorithm 1
$\mathbb{C}=\{C_1, C_2, \cdots, C_k\}$
Incomplete set U_M
Algorithm 2
$\mathbb{C}'=\{C_1', C_2', \cdots, C_k'\}$
Imputation
$\mathbb{C}_{final}=\{C_1, C_2, \cdots, C_k\}$

Fig. 1 Procedure diagram of our algorithm

Algorithm 2. An Improved Mean Imputation Incomplete Data Clustering Algorithm

```
    Input: Dataset: $\mathbf{U} = \{x_1, x_2, \ldots, x_n\}$, Clustering number: $k$.
    Output: Clustering result: $\mathbb{C} = \{C_1, C_2, \ldots, C_k\}$.
 1. Obtaining an incomplete data set $\mathbf{U_M}$ by randomly selecting some missing feature values according to two restrictive conditions.
 2. Classifying the data set $\mathbf{U}$ into two disjoint subsets. One set $\mathbf{U_W}$ requires that each object $x$ with non-missing values and the other set $\mathbf{U_M}$, Where $\mathbf{U_W} \cup \mathbf{U_M} = \mathbf{U}$.
 3. Executing Algorithm 1 to deal with the set $\mathbf{U_W}$, and getting the clustering result $\mathbb{C}_W = \{C_W^1, C_W^2, \ldots, C_W^k\}$.
 4. For each object $x$ in set $\mathbf{U_M}$, impute the object with missing values by using Equation (1) to get the imputation result $x_l = \{x_l^1, x_l^2, \ldots, x_l^k\}(l \in \{1, 2, 3, ..., m\})$.
 5. Add $x$ with $\frac{|C_W^i|}{k}(i = 1, 2, \ldots, k)$ times into $C_W^i$ and receive the new cluster $C_W^{i\prime}$.
 6. Recalculate the cluster centroid using Equation (2).
 7. Compute the difference between $z_i$ and $z_i^*$. i.e., $d_i = |z_i - z_i^*|$ $(i = 1, 2, \ldots, k)$.
 8. Select the minimum value among $d_i$, and assign $x$ to $C_W^i$, the value of these missing values of object $x$ can determined at the same time.
 9. Repeat step 4-8 until $\mathbf{U_M} = \emptyset$.
10. Using Algorithm 1 to cluster the complete dataset $\mathbf{U}$ and obtaining the final clustering result.
11. Return $\mathbb{C} = \{C_1, C_2, \ldots, C_k\}$.
```

The first phase, in this paper, an incomplete data set was obtained by setting the missing rate range from 5% to 30%. Naturally, it is not complicated to distinguish the object with non-missing values or missing values according to the definition in Sect. 2.2. And the set of objects with non-missing values and the set of objects with missing values are represented by $\mathbf{U_W}$ and $\mathbf{U_M}$, respectively, where $\mathbf{U_W} \cup \mathbf{U_M} = \mathbf{U}$.

The second phase, the Algorithm 1 is used to cope with the object in set $\mathbf{U_W}$ and get the clustering result $\mathbb{C}_W = \{C_W^1, C_W^2, \ldots, C_W^k\}$. The object with missing values in the set of $\mathbf{U_M}$ need to be filled based on the clustering result $\mathbb{C}_W$.

The third phase, the method of mean attribute's value of each cluster C_W^i is used to fill the missing value, respectively. The perturbation analysis of cluster center is applied to search the optimal imputation value. For example, the lth attribute value of the object x in set of $\mathbf{U_M}$ is missing, ie $x_l = *$. The Equation (1) is used to impute x_l, and the k interpolation result $x_l = \{x_l^1, x_l^2, \ldots, x_l^k\}$ is acquired spontaneously. The imputation formula is defined as follows.

$$x_l^i = \frac{1}{|C_W^i|} \sum_{x \in C_W^i} x_l \tag{1}$$

where $x_l = *$ is the lth attribute of the object x and $x \in \mathbf{U_M}$, x_l^i represents the i-th imputation result of the l-th attribute of the object x, $|C_W^i|(i = 1, 2, \ldots, k)$ is the cardinality of the ith cluster.

Meanwhile, the method for disturbing distance of cluster centroid is applied to search the optimal imputation value. Each filled object x on corresponding cluster centroid by adding x with $\frac{|C_W^i|}{k}(i = 1, 2, \ldots, k)$ times into C_W^i and receive the new cluster C_W^{i*}. The new cluster centroid was recalculated by Equation (2). Calculating the difference between the old and new cluster centroids, and assigning x to the certain cluster with the smallest difference. And then, the optimal imputation value of the object x with missing values is determined. The formula for calculating the new cluster centroid is as follows.

$$z_i^* = \frac{1}{|C_W^{i*}|}\left(\sum_{x \in C_W^{i*}} x\right) \tag{2}$$

where z_i^* represents the new cluster center and $|C_W^{i*}|$ is the cardinality of the $(i*)$th cluster.

The fourth phase, via the third phase, the incomplete data set $\mathbf{U}$ is transformed into complete data set. Using the k-means algorithm to achieve the final clustering result $\mathbb{C} = \{C_1, C_2, \ldots, C_k\}$.

4 Clustering Performance Measurement

Generally, clustering performance measurement also known as "validity index". It is similar to the effect of performance metrics of supervised learning. As for validity index, we need to adopt it to evaluate the clustering result on one side. On the other side, it can be viewed as an optimization goal during the process of clustering when the validity index is determined.

The clustering performance measurement can be roughly divided into two types. One kind index named external index, which is compare the clustering result with a certain reference model. Another kind index is to directly examine the clustering result without using any reference models, which is called internal index.

4.1 Accuracy

The Accuracy is a frequently-used evaluation index and easy to understand. It represents the ratio between the number of correctly partitioned objects and the total number of samples. The greater value of Accuracy means the more objects are correctly divided. Otherwise, the fewer objects are correctly partitioned.

Definition 1 Accuracy(ACC hereafter).

$$ACC = \frac{1}{n}\sum_{i=1}^{k} \theta_i$$

where the symbol n denotes the total number of samples in a dataset, θ_i represents the amount of objects that are exactly divided into the i-th cluster and the letter k represents clustering number.

4.2 Davies-Bouldin Index

Davide L.davies and Donald W.bouldin [36] proposed the Davies-Bouldin Index is called DBI or DB for short. It mainly compute the distance between clusters and within the cluster. The smaller DBI means the farther distance between clusters and the closer distance within the cluster. Otherwise, the distance among different clusters is close and within the same cluster is far.

Definition 2 Davies-Bouldin Index(DBI hereafter).

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(\frac{S_i + S_j}{M_{i,j}})$$

Where the symbol k is the number of clusters, S_i and S_j represent within the cluster scatter for cluster i and j respectively, which has to be as low as possible. $M_{i,j}$ denote the separation between the cluster i and the cluster j, which ideally has to be as large as possible. Hence, the DBI is defined as the ratio of $M_{i,j}$ and the sum of S_i and S_j. With this formulation is a measure of how good the clustering scheme is.

4.3 Silhouette Coefficient

The Silhouette Coefficient [37] was introduced by Peter J. Rousseeuw. It refers to a method of interpretation and validation of consistency within clusters of data. Meanwhile it is a measure of how similar an object is to its own cluster compared to other clusters. The value of Silhouette Coefficient ranges from -1 to +1, where a high value implies that the sample is well matched to its own cluster and poorly matched to neighboring clusters.

Definition 3 Silhouette Coefficient of single object.

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Where $a(i)$ is the average distance between i and all other data within the same cluster, we can interpret $a(i)$ as a measure of how well i is assigned to its cluster. $b(i)$ is the smallest average distance of i to all points in any other clusters, of which i is not a member. Furthermore, a large $b(i)$ indicates that i is badly matched its neighbouring cluster. Thus an $S(i)$ close to positive one means that the object is appropriately clustered. If $S(i)$ is close to negative one, then by the same logic we see that is would be more appropriate if it was cluster in its neighbouring cluster.

Definition 4 Average Silhouette Coefficient(AS hereafter).

$$AS = \frac{1}{n}\sum_{i=1}^{n} S(i)$$

Where the symbol n represent the total number of objects. $S(i)$ denote the Silhouette Coefficient of single object i. The average $S(i)$ over all points of a cluster is a measure of how tightly grouped all the points in the clusters are. Thus the average $S(i)$ over all data of the entire dataset is a measure of how appropriately the data have been clustered.

5 Experimental Illustration

To illustrate the effectiveness of Algorithm 2, the eight UCI [38]data sets employed in this subsection are Iris, Wine, Glass Identification (Glass), Wisconsin Diagnostic Breast Cancer (WDBC), Banknote, Contraceptive Method Choice (CMC), Pendigits and Page Blocks, respectively. Table 2 shows the details of these data sets. The purpose of our experiment is to verify the performance of the proposed algorithm for incomplete data. On one hand,

Table 2 UCI Datasets used in the experiments

Datasets	Instances	Attributes	Classes
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
WDBC	569	30	2
Banknote	1372	4	2
CMC	1473	9	3
Pendigits	3498	16	10
Page Blocks	5473	10	5

Table 3 Experimental results on UCI datasets

Datasets	Algorithm	Miss rate%	Average value			Best value		
			DBI	AS	ACC	DBI	AS	ACC
Iris	KM–CD	0	0.773	0.686	0.884	0.761	0.696	0.887
		5	0.756	0.695	0.881	0.709	0.719	0.900
		10	0.736	0.709	0.881	0.661	0.736	0.913
	KM–IMI	15	0.732	0.709	0.864	0.625	0.767	0.927
		20	0.745	0.693	0.849	0.661	0.741	0.907
		25	0.735	0.693	0.834	0.659	0.750	0.927
		30	0.728	0.704	0.824	0.628	0.771	0.887
Wine	KM–CD	0	1.317	0.474	0.947	1.305	0.476	0.966
		5	1.265	0.491	0.945	1.232	0.504	0.966
		10	1.235	0.502	0.943	1.089	0.521	0.966
	KM–IMI	15	1.229	0.503	0.931	1.166	0.528	0.966
		20	1.190	0.520	0.923	1.101	0.555	0.955
		25	1.176	0.528	0.912	1.092	0.564	0.972
		30	1.195	0.517	0.892	1.076	0.578	0.955
Glass	KM–CD	0	1.126	0.510	0.632	0.870	0.572	0.841
		5	1.136	0.515	0.652	0.813	0.715	0.883
		10	1.155	0.513	0.651	0.869	0.683	0.874
	KM–IMI	15	1.100	0.523	0.640	0.789	0.692	0.842
		20	1.113	0.528	0.654	0.780	0.642	0.846
		25	1.144	0.523	0.640	0.754	0.698	0.842
		30	1.148	0.518	0.641	0.734	0.621	0.846
WDBC	KM–CD	0	1.136	0.577	0.928	1.136	0.577	0.928
		5	1.099	0.596	0.926	1.084	0.606	0.937
		10	1.079	0.606	0.925	1.050	0.625	0.935
	KM–IMI	15	1.045	0.624	0.922	1.013	0.641	0.937
		20	1.023	0.634	0.921	0.969	0.667	0.933
		25	1.000	0.646	0.918	0.933	0.685	0.938
		30	0.980	0.657	0.912	0.923	0.686	0.938

Table 4 Experimental results on UCI datasets

Datasets	Algorithm	Miss rate%	Average value			Best value		
			DBI	AS	ACC	DBI	AS	ACC
Banknote	KM–CD	0	1.191	0.500	0.574	1.190	0.501	0.576
		5	1.154	0.515	0.576	1.142	0.521	0.600
		10	1.115	0.528	0.597	1.090	0.539	0.620
	KM–IMI	15	1.075	0.551	0.579	1.041	0.559	0.619
		20	1.036	0.572	0.579	1.014	0.582	0.622
		25	0.995	0.587	0.585	0.873	0.616	0.745
		30	0.968	0.596	0.588	0.857	0.624	0.765
CMC	KM–CD	0	1.542	0.373	0.489	1.342	0.480	0.648
		5	1.500	0.388	0.499	1.297	0.497	0.756
		10	1.492	0.398	0.510	1.259	0.511	0.718
	KM–IMI	15	1.429	0.420	0.526	1.213	0.525	0.697
		20	1.406	0.431	0.515	1.167	0.542	0.709
		25	1.421	0.422	0.533	1.169	0.542	0.737
		30	1.438	0.416	0.508	1.171	0.536	0.707
Pendigits	KM–CD	0	1.259	0.466	0.737	1.120	0.509	0.770
		5	1.258	0.472	0.735	1.122	0.519	0.762
		10	1.251	0.477	0.731	1.116	0.524	0.756
	KM–IMI	15	1.257	0.475	0.731	1.113	0.526	0.758
		20	1.254	0.477	0.731	1.120	0.525	0.758
		25	1.260	0.472	0.731	1.122	0.523	0.754
		30	1.261	0.475	0.731	1.130	0.520	0.759
Page Blocks	KM–CD	0	1.021	0.481	0.482	0.886	0.532	0.553
		5	1.010	0.485	0.474	0.873	0.546	0.547
		10	1.007	0.487	0.475	0.874	0.546	0.547
	KM–IMI	15	1.008	0.484	0.473	0.872	0.546	0.548
		20	1.011	0.493	0.484	0.871	0.546	0.558
		25	1.004	0.495	0.485	0.871	0.547	0.551
		30	1.006	0.485	0.472	0.870	0.548	0.556

comparing the average and best values of DBI, AS and ACC of the algorithm KM-IMI and KM-CD (the k-means clustering algorithm under complete datasets). On the other hand, for purpose of better show the superiority of the proposed method, we compare the results between the KM-IMI and the OCS-FCM, NPS-FCM, which are two kind of classical clustering algorithms for incomplete data. The experimental results are shown in Table 3, Table 4 and Table 5, respectively. And the detailed analysis of the experimental results is given below.

An incomplete data set $\mathbf{U_M}$ was got by selecting some missing feature values randomly. The object with missing values is constrained by two restrictive conditions: (1) each original feature vector x retains at least one component;(2) each feature has at least one value present in the incomplete data set. That is to say, an object can not to be missing all attribute values and all objects can not be missing the same attribute. In most cases, the higher the missing rate in the dataset, the lower the accuracy of the clustering results. Because the higher the

Table 5 Experimental results of ACC on UCI datasets

Datasets	Miss rate%	Average value			Best value		
		KM-IMI	OCS-FCM	NPS-FCM	KM-IMI	OCS-FCM	NPS-FCM
Iris	5	**0.883**	0.833	0.715	**0.913**	0.873	0.860
	10	**0.874**	0.833	0.649	**0.900**	0.840	0.700
	15	**0.857**	0.819	0.660	**0.900**	0.833	0.840
	20	**0.835**	0.751	0.669	**0.893**	0.787	0.860
	25	**0.825**	0.787	0.644	**0.900**	0.807	0.787
	30	**0.807**	0.720	0.678	**0.827**	0.767	0.787
Wine	5	**0.944**	0.933	0.788	**0.978**	0.955	0.933
	10	**0.942**	0.910	0.749	**0.983**	0.927	0.865
	15	**0.926**	0.899	0.862	**0.978**	0.899	0.966
	20	**0.916**	0.890	0.802	**0.966**	0.899	0.883
	25	**0.892**	0.860	0.849	**0.962**	0.865	0.889
	30	**0.868**	0.801	0.739	**0.962**	0.831	0.889
Glass	5	**0.645**	0.562	0.504	**0.883**	0.593	0.575
	10	**0.642**	0.628	0.512	**0.869**	0.696	0.631
	15	0.645	0.654	**0.672**	**0.862**	0.682	0.702
	20	0.652	**0.677**	0.539	**0.854**	0.696	0.785
	25	0.639	**0.677**	0.539	**0.859**	0.687	0.721
	30	0.642	**0.677**	0.540	**0.838**	0.710	0.776
WDBC	5	**0.926**	0.919	0.919	**0.938**	0.926	0.919
	10	**0.925**	0.917	0.891	**0.938**	0.917	0.891
	15	**0.922**	0.917	0.858	**0.938**	0.917	0.858
	20	**0.919**	0.917	0.796	**0.944**	0.917	0.796
	25	0.914	**0.916**	0.735	**0.948**	0.917	0.735
	30	0.906	**0.910**	0.707	**0.947**	0.914	0.707
Pendigits	5	**0.733**	0.611	0.703	**0.759**	0.697	0.741
	10	**0.729**	0.558	0.642	**0.753**	0.607	0.734
	15	**0.734**	0.511	0.681	**0.755**	0.557	0.674
	20	**0.731**	0.529	0.584	**0.751**	0.571	0.674
	25	**0.732**	0.578	0.643	**0.758**	0.586	0.705
	30	**0.729**	0.600	0.534	**0.756**	0.607	0.703
Page Block	5	**0.466**	0.441	0.427	**0.534**	0.461	0.466
	10	**0.491**	0.427	0.350	**0.540**	0.433	0.364
	15	**0.495**	0.429	0.365	**0.537**	0.432	0.387
	20	**0.490**	0.431	0.360	**0.540**	0.436	0.369
	25	**0.490**	0.445	0.347	**0.540**	0.456	0.361
	30	**0.489**	0.463	0.348	**0.538**	0.469	0.363

missing rate, the accuracy of incomplete data filling will also decrease, which directly leads to the performance degradation of the clustering algorithm. Therefore, in this paper, we set the missing rate range from 5% to 30%.

Table 3 and Table 4 present the experimental results on eight UCI data sets, mainly make a comparison between the KM-IMI and KM-CD in the average and best value of DBI, AS and ACC. The underlined data indicates that the KM-IMI is not as good as the KM-CD. Based on this, we can find that the KM-IMI is superior to the KM-CD on most data sets in the average and best value of DBI and AS. Like Iris, Wine, WDBC, CMC and Page Blocks. However, the effect of the average and best value of Pendigits and Page Blocks are not so good as the KM-CD but the gap is only between 0.01 and 0.02. One of the reasons is that there is a proportional relationship between the missing rate and the accuracy rate, which directly leads to the performance degradation of the clustering algorithm. The average value of ACC in most data sets does not perform well, but the best value of ACC is effectiveness. It is worth mentioning that Banknote and CMC have performed well in the average and best value of DBI, AS and ACC. Therefore, we can conclude that the Algorithm 2 can be used to some extent to fill and cluster incomplete data sets. The results of several data sets in the experimental results are poor performance, indicating that the Algorithm 2 needs to be further improved and perfected.

In Table 5, the experimental results of the average and best value of ACC on six UCI data sets of the KM-IMI and OCS-FCM and NPS-FCM are shown, where the bold data represents the best results. We get the average and best ACC value through 100 times experimental test under different missing rates range from 5% to 30%. Table 5 includes KM-IMI, OCS-FCM and NPS-FCM, respectively. The OCS-FCM and NPS-FCM are two classic algorithms for clustering incomplete data set, most methods based on FCM are implemented on the basis of [27]. By comparing the experimental results of the three algorithms on the average and best ACC values, the superiority of the proposed algorithm KM-IMI can be highlighted. From the experimental results recorded in Table 5, it can be seen that the average and best ACC experimental results of KM-IMI are better than OCS-FCM and NPS-FCM, like Iris, wine, Pendigits and Page Blocks. Only on a few data sets, KM-IMI is not better than OCS-FCM, like Glass and WDBC but the difference is between 0.01 and 0.04. After analysis, it is not difficult to find that the comparison method is based on the fuzzy C-means method, so it is difficult to obtain good results on non-spherical data sets. At the same time, on the Pendigits and Page Blocks datasets, the accuracy of this method is significantly higher than the comparison method. Therefore, the algorithm KM-IMI in this paper is better than the comparison algorithms OCS-FCM and NPS-FCM in the average and best ACC value.

In the experimental part, we not only compared our method with the original k-means algorithm, but also with the algorithms OCS-FCM and NPS-FCM. On one hand, the difference between the filling result and the actual value of the incomplete data can be obtained from the clustering result. On the other hand, the effectiveness of the algorithm can be verified by clustering performance measurement. Because the size of the missing rate will affect the accuracy of the algorithm, this paper has certain requirements on the range of missing rate. All these works have proved that our algorithm has better performance.

6 Conclusion

In this paper, our work focuses on how to impute and cluster incomplete data set. An improved mean imputation clustering algorithm for incomplete data based on k-means algorithm is proposed. In the proposed algorithm, we cluster the objects with non-missing values and use the mean attribute's value of each cluster to fill the corresponding missing value, respectively. The method of perturbation analysis of cluster centroid is used to find the optimal filling value. Finally, the k-means algorithm clusters all objects including the original object and the filled object. In order to verify the effectiveness of our algorithm, on the one hand, the average and best value of DBI, AS and ACC of algorithms KM-IMI and KM-CD are compared. On the other hand, in order to better demonstrate the superiority of the proposed method, we compare the effects of the algorithm KM-IMI and two classical incomplete data clustering algorithms OCS-FCM and NPS-FCM. By comparing these experimental results, we can summarize that our algorithm is effective in filling incomplete data. However, the algorithm does not perform well in some data sets, which indicates that further improvements are needed.

There are some shortcomings in our algorithm. In the following work, we will consider how to improve and refine the interpolation and clustering algorithm for incomplete data, as well as think about how to combine incomplete data with the three-way decision clustering algorithm.

Acknowledgements The authors would like to thank the editor and the anonymous reviewers for their constructive and valuable comments. This work was supported in part by National Natural Science Foundation of China (Nos. 61503160 and 61773012), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 15KJB110004), Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX19_1699).

References

1. Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16:645–678
2. Aggarwal CC, Reddy CK (2013) Data clustering: algorithms and applications. Chapman & Hall/CRC Press, Boca raton
3. Wang PX, Yao YY (2018) CE3: a three-way clustering method based on mathematical morphology. Knowl-Based Syst 155:54–65
4. Wang PX, Shi H, Yang XB, Mi JS (2019) Three-way k-means: integrating k-means and three-way decision. Int J Mach Learn Cybernet 10:2767–2777
5. Yang XB, Qi YS, Song XN, Yang JY (2013) Test cost sensitive multigranulation rough set: model and minimal cost selection. Inf Sci 250:184–199
6. Qian YH, Cheng HH, Wang JT, Liang JY, Pedrycz W, Dang CY (2017) Grouping granular structures in human granulation intelligence. Inf Sci 382–383:150–169
7. Yang XB, Yao YY (2018) Ensemble selector for attribute reduction. Appl Soft Comput 70:1–11
8. Elalami ME (2011) Supporting image retrieval framework with rule base system. Knowl-Based Syst 24:331–340
9. Sebiskveradze D, Vrabie V, Gobinet C (2011) Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections. Lab Investig 91:799–811
10. Kalyani S, Swarup KS (2011) Particle swarm optimization based k-means clustering approach for security assessment in power systems. Expert Syst Appl 38:10839–10846
11. Wu YH (2015) General overview on clustering algorithms. Comput Sci 42:491–499
12. Jain AK (2008) Data clustering: 50 years beyond k-means. In: 2008 European conference on machine learning and principles and practice of knowledge discovery in databases. Springer, Berlin, pp 3–4
13. Macqueen J (1967) Some methods for classification and analysis of multi-variate observations. In: 1967 Proceeding of Berkeley symposium on mathematical statistics and probability conference, pp 281–297
14. Arthur D, Vassilvitskii S (2007) K-Means++: the advantages of careful seeding. In: ACM-SIAM symposium on discrete algorithms (SODA'07), New Orleans, LA, pp 1027–1035
15. Park HS, Jun CH (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341
16. Yu SS, Chu SW, Wang CM, Chan YK, Chang TC (2018) Two improved k-means algorithms. Appl Soft Comput 68:747–755

17. Franti P, Sieranoja S (2019) How much can k-means be improved by using better initialization and repeats? Pattern Recogn 93:95–112
18. Chao GQ (2019) Discriminative k-means laplacian clustering. Neural Process Lett 49:393–405
19. Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modeling. Math Geosci 42:487–517
20. Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
21. Little RJA, Rubin DB (2002) Statistical analysis with missing data. Technometrics 45:364–365
22. Pang XS (2012) Comparative study on interpolation processing method of missing data. Stat Decis 24:18–22
23. Dempster A (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38
24. Grzymala-Busse JW, Fu M (2000) A comparison of several approaches to missing attribute values in data mining. In: Proceedings of the 2nd international conference on rough sets and current trends in computing. Springer, Berlin, pp 378–385
25. Doquire G, Verleysen M (2012) Feature selection with missing data using mutual information estimators. Neurocomputing 90:3–11
26. Jason VH, Khoshgoftaar TM (2014) Incomplete-case nearest neighbor imputation in software measurement data. Inf Sci 259:586–610
27. Hathaway RJ, Bezdek JC (2001) Fuzzy c-means clustering of incomplete data. IEEE Trans Syst Man Cybern Part B Cybern 31:735–744
28. Zhang DQ, Chen SC (2003) Clustering incomplete data using kernel-based fuzzy c-means algorithm. Neural Process Lett 18:155–162
29. Li D, Gu H, Zhang LY (2010) A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. Expert Syst Appl 37:6942–6947
30. Li D, Zhang LY, Gu H (2012) An attribute weighted fuzzy c-means algorithm for incomplete data clustering. J Dalian Univ Technol 52:449–453
31. Li D, Gu H, Zhang LY (2013) A hybrid genetic algorithm fuzzy c-means approach for incomplete data clustering based on nearest-neighbor intervals. Soft Comput 17:1787–1796
32. Li JH, Song SJ, Zhang YL, Li K (2017) A robust fuzzy c-means clustering algorithm for incomplete data. In: 2017 International conference on life system modeling and simulation & 2017 international conference on intelligent computing for sustainable energy and environment, vol 762, pp 3–12 (2017)
33. Su T, Yu H (2016) Three-way decision clustering algorithm for incomplete data based on q-nearest neighbors. J Frontiers Comput Sci Technol 10:875–883
34. Shi QY, Liang JY, Zhao XW (2016) A clustering ensemble algorithm for incomplete mixed data. J Comput Res Develop 53:1979–1989
35. Mesquita DPP, Gomes JPP, Rodrigues LR (2019) Artificial neural networks with random weights for incomplete datasets. Neural Process Lett. https://doi.org/10.1007/s11063-019-10012-0
36. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1:224–227
37. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
38. Bache K, Lichman M (2013) UCI machine learning repository. University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.