Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 535932, 9 pages
http://dx.doi.org/10.1155/2015/535932

Research Article

K-Nearest Neighbor Intervals Based AP Clustering Algorithm for Large Incomplete Data

Cheng Lu,[1,2] Shiji Song,[1] and Cheng Wu[1]

[1]*Department of Automation, Tsinghua University, Beijing 100084, China*
[2]*Army Aviation Institute, Beijing 101123, China*

Correspondence should be addressed to Shiji Song; shijis@mail.tsinghua.edu.cn

Received 15 January 2015; Accepted 2 March 2015

Academic Editor: Hui Zhang

The Affinity Propagation (AP) algorithm is an effective algorithm for clustering analysis, but it can not be directly applicable to the case of incomplete data. In view of the prevalence of missing data and the uncertainty of missing attributes, we put forward a modified AP clustering algorithm based on *K*-nearest neighbor intervals (KNNI) for incomplete data. Based on an Improved Partial Data Strategy, the proposed algorithm estimates the KNNI representation of missing attributes by using the attribute distribution information of the available data. The similarity function can be changed by dealing with the interval data. Then the improved AP algorithm can be applicable to the case of incomplete data. Experiments on several UCI datasets show that the proposed algorithm achieves impressive clustering results.

1. Introduction

With the developments of sensors and database technology, people get more focus on the Big Data issue [1]. But too often the data is difficult to analyze. Cluster analysis is one of the common methods for analyzing data, which is to partition a set of objects into different groups, so that the data in each cluster share some common traits. Affinity Propagation (AP) is a relatively new clustering algorithm that has been introduced by Frey and Dueck [2], which can handle large datasets in a relatively short period to obtain more satisfactory results. AP algorithm has superiority over other clustering algorithms in terms of processing efficiency and quality of clustering, and AP algorithm does not require the prespecified number of clusters and the initial cluster centers. Thus, AP algorithm has attracted the attention of many scholars, and various improvements have emerged [3–5].

As a common and effective clustering algorithm, the original AP clustering algorithm is only applicable to complete data like other traditional clustering algorithms. However, in practice, many datasets suffer from incompleteness due to various reasons, such as bad sensors, mechanical failures to collect data, illegible images due to low pixels and noises, and unanswered questions in surveys. Therefore, some strategies should be employed to make AP applicable to such incomplete datasets.

In the literature, several approaches to handle incomplete data have been proposed, including listwise deletion (LD), imputation, model-based method, and direct analysis [6]. There is a strong connection between these methods on the concrete implementation algorithm. LD ignores those samples with missing values, which may lose a lot of sample information. Imputation and model-based method are usually based on the assumption that data attributes are missing at random. They substitute the missing values with appropriate estimates and construct a complete dataset. However, it is inefficient to perform imputation, and they usually lead to results far from satisfactory. For incomplete data, many methods have been proposed to reduce the impact of the presence of the missing values on the clustering performance in pattern recognition. An important empirically oriented study was done by Dixon [7]. The expectation-maximization (EM) algorithm [8] is a commonly used iterative algorithm based on maximum likelihood estimation in missing data analysis. Neither statistical methods nor machine learning method for dealing with missing data meets the actual needs of current. Various

methods for handling missing data remain to be further optimized.

Though incomplete data appears everywhere, principled clustering methods for such data still deserve further research. The existing research on improved methods for clustering model is mainly concentrated on the fuzzy C-means clustering (FCM) algorithm (Bezdek, 1981) [9]. In 1998, imputation and discarding/ignoring were proposed by Miyamoto et al. [10] for handling missing values in FCM. In 2001, Hathaway and Bezdek proposed four strategies to improve the FCM clustering of incomplete data and proved the convergence of the algorithms [11]. These strategies are whole data strategy (WDS), partial distance strategy (PDS), optimal completion strategy (OCS), and nearest prototype strategy (NPS). In addition, Hathaway and Bezdek used triangle inequality-based approximation schemes (NERFCM) to cluster incomplete relational data [12]. Li et al. [13] put forward a FCM algorithm based on the nearest neighbor intervals and solved the case of incomplete data. Zhang and Chen [14] introduced a kernel method into the standard FCM algorithm.

However, FCM algorithms are sensitive to the initial centers, which makes the clustering results unstable. In particular when some data are missing, the selection of the initial cluster centers becomes more important. To address this issue, we consider the AP algorithm, which does not require initial cluster centers and the number of clusters. Three strategies for solving AP clustering of incomplete datasets had been proposed in our previous research [15]. These strategies were simple and easy to implement which directly deal with incomplete dataset using AP algorithm. However, the effect of dataset information on missing attributes had not been studied, by which clustering quality would be affected. In this paper, based on Improved Partial Data Strategy (IPDS), a modified AP algorithm for incomplete data based on K-nearest neighbor intervals (KNNI-AP) is proposed. First, missing attributes are represented by KNNI on the basis of IPDS, which are robust. Second, the clustering problems are transformed into clustering problems with interval-valued data, which may provide more accurate clustering results. Third, AP algorithm simultaneously considers all data points as potential centers, which makes the clustering results more stable and accurate.

The remainder of this paper is organized as follows. Section 2 presents a description of AP algorithm and AP clustering algorithm for interval-valued data (IAP) based on clustering objective function minimization. The KNNI representation of missing attributes and the novel KNNI-AP algorithm are introduced in Section 3. Section 4 presents clustering results of several UCI datasets and a comparative study of our proposed algorithm with KNNI-FCM and other methods for handling missing values using AP. We conclude this work and discuss the future work in Section 5.

2. AP Clustering Algorithm for Interval-Valued Data

2.1. AP Clustering Algorithm. AP algorithm and K-means algorithm have similar objective function, but the AP algorithm simultaneously considers all data points as the potential centers.

Let a complete dataset $X = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in \mathbb{R}$. The goal of AP is to find an optimal exemplar set $X_C = \{x_{c1}, x_{c2}, \ldots, x_{ck}\}$, $(1 < k < N)$, by minimizing the clustering error function:

$$J(C) = \sum_{i=1}^{N} d^2\left(x_i, C\left(x_i\right)\right), \tag{1}$$

where $C(x_i)$ represents the exemplar for given x_i. Each data point only corresponds to a cluster, and each exemplar is an actual data point which is the center of the cluster.

First, AP algorithm takes each data point as the candidate exemplar and calculates the attractiveness information between sample points, that is, the similarity between any two sample points. The similarity can be set according to specific applications; similarity measurement mainly includes similarity coefficient function and distance function. Common distance functions are Euclidean distance, Manhattan distance, and Mahalanobis distance. In the traditional clustering problem, similarity is usually set as the negative of squared Euclidean distance:

$$s(i, j) = -d^2\left(x_i, x_j\right) = -\left\|x_i - x_j\right\|_2^2, \quad i \neq j, \tag{2}$$

where $s(i, j)$ is stored in a similarity matrix, representing the suitability that the sample x_i is the exemplar of the sample x_j. $s(i, i)$ is set for each sample, called "preference." The greater the value is, the more possible the corresponding point is selected as the exemplar. Because all samples are equally suitable as centers, the preferences should be set as a common value P. The number of identified exemplars is influenced by P, which can be changeable for different numbers of clusters. Frey and Dueck [2] suggested preference is the median of the input similarities (resulting in a moderate number of clusters) or their minimum (resulting in a small number of clusters). We also employ [5] to measure the preference values to get more accurate clustering results.

To select appropriate clustering centers, AP algorithm searches for two different pieces of information: responsibility ($r(i, j)$) and availability ($a(i, j)$). $r(i, j)$ sent from sample i to sample j reflects how well-suited sample j is to be served as the cluster center for sample i. $a(i, j)$ sent from sample j to sample i reflects how appropriate for sample i to choose sample j as its exemplar. The message-passing procedure terminates after a fixed number of iterations or the changes in the messages fall below a threshold.

2.2. AP Clustering Algorithm for Interval-Valued Data (IAP). AP algorithm should be adjusted to deal with interval data. Let an M-dimensional interval-valued dataset $\overline{X} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_N}\}$, where the ith sample is expressed as $\overline{X_i} = \{\overline{x_{i1}}, \overline{x_{i2}}, \ldots, \overline{x_{iM}}\}$ and $\overline{x_{il}} = [x_{il}^-, x_{il}^+]$, $(1 \leq l \leq M)$. To find the optimal exemplar set $\overline{X_c} = \{\overline{x_{c1}}, \overline{x_{c2}}, \ldots, \overline{x_{ck}}\}$ $(1 < k < N)$, we minimize the following clustering error function:

$$J(C) = \sum_{i=1}^{N} d^2\left(\overline{x_i}, \overline{C(x_i)}\right), \tag{3}$$

where $\overline{C(x_i)}$ represents the exemplar for given $\overline{x_i}$. The similarity is changed as

$$s(i,j) = -d^2\left(\overline{x_i}, \overline{x_j}\right) = -\left\|\overline{x_i} - \overline{x_j}\right\|_2^2, \quad i \neq j. \tag{4}$$

The Euclidean distance can be defined as

$$\begin{aligned}\left\|\overline{x_i} - \overline{x_j}\right\|_2^2 = &\sum_{l=1}^{M}\left|x_{il}^{+} - x_{jl}^{+}\right|^2 + \sum_{l=1}^{M}\left|x_{il}^{-} - x_{jl}^{-}\right|^2 \\ &+ \sum_{l=1}^{M}\left[\frac{\left|x_{il}^{+} - x_{jl}^{+}\right| + \left|x_{il}^{-} - x_{jl}^{-}\right|}{2}\right]^2.\end{aligned} \tag{5}$$

Similarity matrix of $\overline{X}$ can be calculated accordingly. Then the two pieces of information are updated alternately, which are both zero in the initial stage, and the update process is given as follows:

$$r(i,j) \longleftarrow s(i,j) - \max\left[a\left(i,j'\right) + s\left(i,j'\right)\right],$$

$$a(i,j) \longleftarrow \begin{cases} \min\limits_{i \neq j}\left\{0, r\left(j,j\right) + \sum\limits_{i' \neq i, i' \neq j}\max\left[0, r\left(i',j\right)\right]\right\}, \\ \qquad\qquad i \neq j, \\ \sum\limits_{i' \neq j}\max\left[0, r\left(i',j\right)\right], \quad i = j. \end{cases} \tag{6}$$

To avoid the numerical oscillation, the damping factor λ is introduced as follows:

$$\begin{aligned} R_i &= (1-\lambda)\, R_i + \lambda R_{i-1}, \\ A_i &= (1-\lambda)\, A_i + \lambda A_{i-1}. \end{aligned} \tag{7}$$

The procedure of IAP can be described as follows.
Input is the similarity matrix S and the preference P.
Output is the clustering result.

Step 1. Initialize responsibility ($r(i,j)$) and availability ($a(i,j)$) to zero: $r(i,j) = 0$; $a(i,j) = 0$.

Step 2. Update the responsibilities.

Step 3. Update the availabilities.

Step 4. Terminate the message-passing procedure after a fixed number of iterations or the changes in the messages fall below a threshold. Otherwise go to Step 2.

3. AP Algorithm for Incomplete Data Based on K-Nearest Neighbor Intervals (KNNI-AP)

3.1. K-Nearest Neighbor Intervals of Missing Attributes. As a common method to handle missing data, neighbor imputation has been widely used in many areas [16]. Imputation is the problem of approximating the value of a function for a nongiven point in some space when given the value of that function in points around (neighboring) that point. As a simple imputation, the nearest neighbor algorithm selects the nearest sample and does not consider the neighboring samples at all, which is easy to implement and is commonly used. An improved method is K-nearest neighbor imputation [17], where missing attributes are supplemented by the mean value of the attributes in the K-nearest neighbor values. Subsequently, García-Laencina et al. [18] proposed a K-nearest neighbor interpolation method based on weighted distance characteristics of multiple information. Huang and Zhu [19] introduced a pseudodistance neighbor interpolation method. All the approaches mentioned above developed imputation, which are unsuitable to represent the uncertainty of missing attributes completely.

To produce a robust estimation, K-nearest neighbor intervals (KNNI) of missing attributes are proposed. Let $X = \{x_1, x_2, \ldots, x_N\}$ be an M-dimensional incomplete dataset, which contains at least one incomplete sample with some (but not all) missing attribute values. For an incomplete $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iM}\}$, the K-nearest neighbors should be found first.

On the basis of the Improved Partial Data Strategy (IPDS), the attribute information of both complete sample and incomplete sample (nonmissing attributes) can be fully used. The distance between sample A and sample B can be obtained as follows:

$$\left\|x_a - x_b\right\|_2 = \sqrt{\sum_{j=1}^{M} d_j\left(x_{aj}, x_{bj}\right)^2} \times \frac{M}{\omega}, \tag{8}$$

where

$$\begin{aligned} d_j\left(x_{aj}, x_{bj}\right) &= \begin{cases} 0, & \left(1 - m_{aj}\right)\left(1 - m_{bj}\right) = 0, \\ d_N\left(x_{aj}, x_{bj}\right), & \text{others}, \end{cases} \\ d_N\left(x_{aj}, x_{bj}\right) &= \frac{\left|x_{aj} - x_{bj}\right|}{\max\left(x_j\right) - \min\left(x_j\right)}, \\ m_{ij} &= \begin{cases} 1, & x_{ij} \text{ is missing}, \\ 0, & x_{ij} \text{ is not missing}, \end{cases} \end{aligned} \tag{9}$$

where $1 \leq j \leq M$, $1 \leq i \leq N$. $d_j(x_{aj}, x_{bj})$ represents the distance on the jth attribute between the two samples. ω is the feature dimension in which the two samples are both not missing, and M is the dimensions of all features. $\max(x_j)$ and $\min(x_j)$ are the maximum and minimum of the observation data when the missing attribute exists. m_{ij} is indicator function to explain whether the variable is missing.

According to the principle of the nearest neighbor approach, sample and its nearest neighbor share same or similar attributes. Therefore, for sample A, the ranges of missing attributes are basically between the minimum and maximum values of the corresponding attribute values of its K-nearest neighbors. Then the K-nearest neighbor interval of the sample can be determined, and the dataset can be converted into interval dataset. The missing attribute x_{aj} is represented by its corresponding K-nearest neighbor interval

$\overline{x_{aj}} = [x_{aj}^-, x_{aj}^+]$, and nonmissing attribute x_{cj} can also be rewritten into interval form $\overline{x_{cj}} = [x_{cj}^-, x_{cj}^+]$, where $x_{cj}^- = x_{cj}^+ = x_{cj}$. That is, the original values are unchanged. Then the interval dataset $\overline{X} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_N}\}$ is formed.

The selected K is critical to make the intervals represent the missing attributes effectively. If K is too small, the interval values may not express the missing attribute correctly, which likely leads to a biased estimation. In the extreme situation when K is 1, KNNI is degraded into NNI. However, if K is too large, the interval values also cannot correctly characterize the missing attribute values. In the extreme situation when K is as large as n (the number of samples in the dataset), the missing attribute interval is the range of all samples on the attribute, which is too large to represent the missing attribute properly. This will confuse the attribute characteristics among different clusters and result in unreasonable clustering results. K is not only related with the ratio of the missing attribute, but also related with the distribution of the sample and the relevant clusters. Thus, how to choose an effective K will directly affect the accuracy of clustering.

We randomly selected 3 kinds of three-dimensional data to form a dataset. For example, the number of samples is 900, and the missing rate is 15%. KNNI-AP algorithm is used, respectively, when K is selected from 1 to 50; from the test results we can see that clustering results are basically stabilized when K is more than 10 and there is uncertainty when K is too small. Similar to the above process, for random missing data with different dimensions and different sample numbers, the values of K were tested. It can be found that K selected as the cube root of the sample numbers is more appropriate. Therefore, in this paper, the selected K is the cube root of the sample numbers rounded to the nearest integer.

3.2. AP Algorithm for Incomplete Data Based on KNNI (KNNI-AP). KNNI-AP proposed here deals with clustering problem for incomplete data by transforming the dataset to an interval-valued one. The range of missing attribute interval $\overline{x_{aj}} = [x_{aj}^-, x_{aj}^+]$ will be large if the jth attributes are dispersive in clusters and will be small if the jth attributes are compact in clusters. So the KNNI can represent the uncertainty of missing attributes better. The lower and upper boundaries of missing attributes interval are determined by the distributions of attributes in clusters, that is, by the geometrical structure of clusters which can present to some extent the shape of clusters and sample distribution of the dataset. The proposed KNNI-AP can validate the robustness of clustering pattern.

For an M-dimensional incomplete dataset $X = \{x_1, x_2, \ldots, x_N\}$, the procedure of KNNI-AP can be described as follows.

Step 1. Set K as the cube root of the sample numbers rounded to the nearest integer.

Step 2. The distance between sample A and sample B can be obtained based on the IPDS, and the similarity matrix $S1$ can be constructed.

Step 3. Form the corresponding interval dataset $\overline{X} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_N}\}$. For each missing attribute x_{aj}, find its K-nearest neighbors using $S1$. x_{aj} is represented by $\overline{x_{aj}} = [x_{aj}^-, x_{aj}^+]$, and nonmissing attribute x_{cj} is rewritten into interval form $\overline{x_{cj}} = [x_{cj}^-, x_{cj}^+]$, where $x_{cj}^- = x_{cj}^+ = x_{cj}$.

Step 4. Calculate the similarity matrix S of $\overline{X} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_N}\}$. Choose the parameter of AP: maximum number of iterations performed by AP (default 2000); convergence of the algorithm if the estimated cluster centers stay fixed for convits iterations (default 50); decreasing step of preferences (default 0.01); damping factor (default 0.5).

Step 5. Apply Preference Range algorithm to computing the range of preference. Initialize the preference: $P = P_{\min} - p\text{step}$. Update the preference: $P = P + p\text{step}$.

Step 6. Apply IAP algorithm to generating C clusters. If cluster number is known then judge weather C is equal to the given number of clusters; else a series of Sil values corresponding to the clustering result with different numbers of cluster is calculated.

Step 7. If cluster number is known, algorithm terminates until C is equal to the given number of clusters; else it terminates until Sil is the largest.

4. Simulation Analysis

4.1. Incomplete Datasets. In order to test the proposed clustering algorithm, we use artificially generated incomplete datasets. The scheme for artificially generating an incomplete dataset X is to randomly select a specified percentage of components and designate them as missing. The random selection of missing attribute values should satisfy the following [11]:

(1) each original feature vector x_k retains at least one component;

(2) each attribute has at least one value present in the incomplete dataset X.

At least one-dimensional data exists for each vector data and at least one or more kinds of data exist for each dimension. That is, the data in each row are not empty; each column of data cannot be null. In the following experiments, we test the performance of proposed algorithm on commonly used UCI datasets: Iris, Seeds, Wisconsin Diagnostic Breast Cancer (WDBC), and Wholesale customers, which are taken from the UCI machine repository [20] and are often used as standard databases to test the performance of clustering algorithms.

The Iris dataset contains 150 four-dimensional attribute vectors. The Wine dataset used in this paper contains 178 three-dimensional attribute vectors. The WDBC dataset comprises 569 samples and, for each sample, there are 30 attributes. The Wholesale customers dataset refers to clients of a wholesale distributor containing 440 6-dimensional attribute vectors.

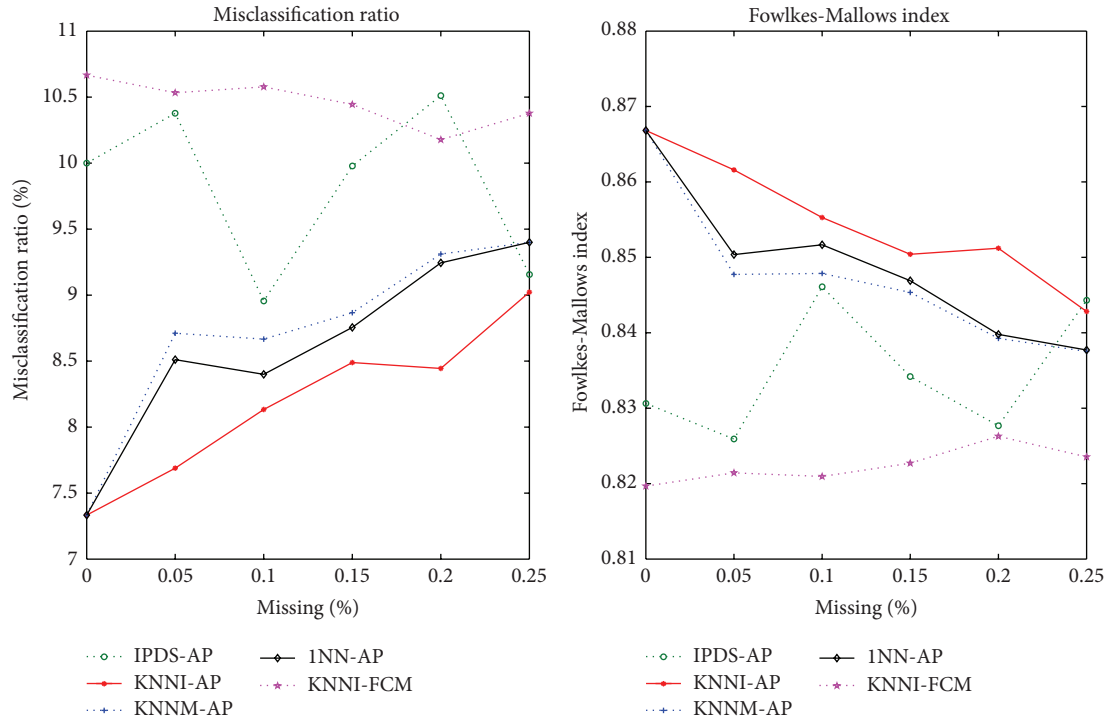

Figure 1: Averaged clustering results of 30 trials using incomplete Iris dataset.

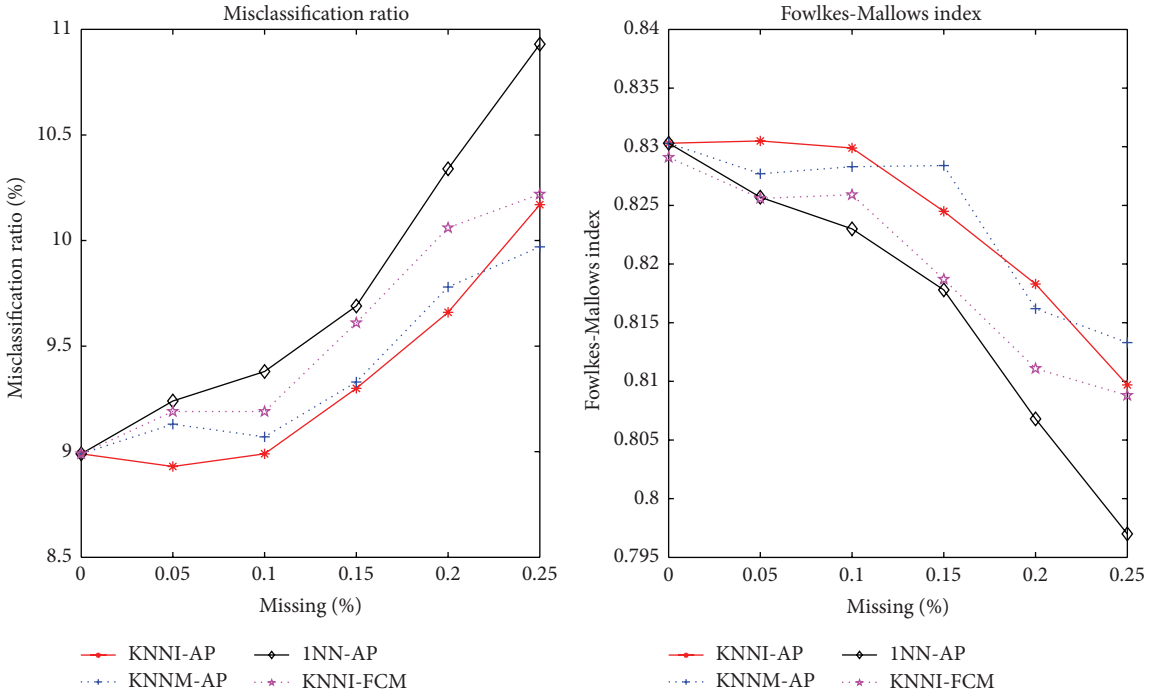

Figure 2: Averaged clustering results of 30 trials using incomplete Wine set.

4.2. Compared Algorithms. To test the clustering performance, we take AP based on the K-nearest neighbor mean (KNNM-AP), AP based on the nearest neighbor (1NN-AP), AP based on IPDS (IPDS-AP), and FCM based on the K-nearest neighbor interval (KNNI-FCM) as compared algorithms. IPDS-AP directly deals with incomplete dataset using AP algorithm, and the others are imputation algorithms using different methods to handle missing values. For KNNM-AP, missing attributes are calculated by the K-nearest neighbor mean; for 1NN-AP, missing attributes are replaced

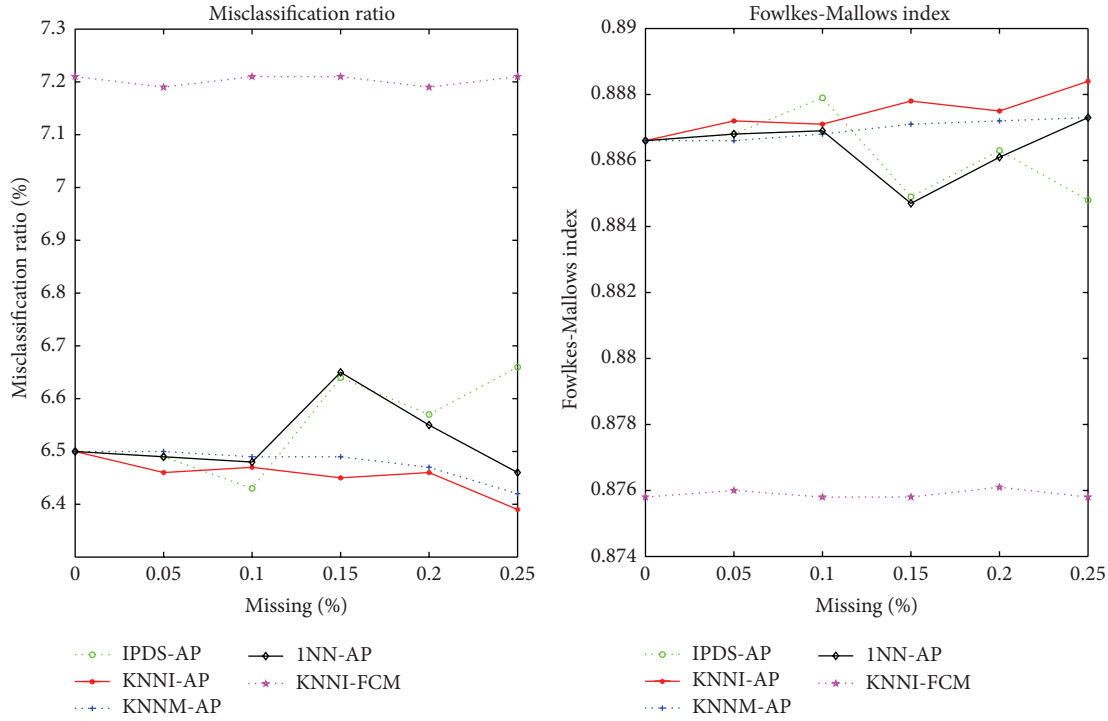

Figure 3: Averaged clustering results of 30 trials using incomplete WDBC dataset.

by the nearest neighbor; for KNNI-FCM, missing attributes are calculated by KNNI similar to KNNM-AP.

4.3. Evaluation Method. To evaluate the quality of clustering results, we use misclassification ratio and Fowlkes-Mallows index [21].

Fowlkes-Mallows (FM) index is used to measure the clustering performance based on external criteria. In general, the larger the FM value is, the better the clustering performance is. The FM index is defined as

$$\mathrm{FM} = \frac{a}{w_1 \cdot w_2} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}, \tag{10}$$

where $w_1 = a + b$ and $w_2 = a + c$.

C_1 is a clustering structure of the dataset and C_2 is a defined partition of the data. We refer to a pair of samples (x_u, x_v) from the dataset using the following terms.

SS: if both samples belong to the same cluster of the clustering structure C_1 and to the same group of partition C_2.

SD: if samples belong to the same cluster of C_1 and to different groups of C_2.

DS: if samples belong to different clusters of C_1 and to the same group of C_2.

DD: if both samples belong to different clusters of C_1 and to different groups of C_2.

a, b, c, and d are the number of SS, SD, DS, and DD pairs, respectively. Then $a + b + c + d = W$, which is the maximum number of all pairs in the dataset (meaning, $W = N(N-1)/2$, where N is the total number of samples in the dataset).

The misclassification rate calculates the proportion of an observation being allocated to the incorrect group. It is calculated as follows: the number of incorrect classifications is divided by the total number of samples.

4.4. Experimental Results and Discussion. For the four datasets, damping factor $\lambda = 0.85$, decreasing step of preferences $p\mathrm{step} = 0.01$, max iteration time $n\mathrm{run} = 2000$, and convergence condition $n\mathrm{conv} = 100$. Because missing data was randomly generated, different tests lead to different results, and we noticed significant variation in the results from trial to trial. To eliminate the variation in the results, Figures 1–4 and Tables 1–4 give the averaged results over 30 trials on incomplete Iris, Wine, WDBC, and Wholesale customers datasets. Figures can intuitively reflect the effects of the algorithms and tables can accurately characterize the clustering results of the algorithms. In particular, 30 trials are generated for each row in the table, and the same incomplete dataset is used in each trial for each algorithm, so that the results can be correctly compared. In the tables, the optimal solutions in each row are highlighted in bold, and the suboptimal solutions are italic.

To test the clustering performance, the clustering results of KNNI-AP, 1NN-AP, KNNM-AP, IPDS-AP, and KNNI-FCM are compared. From figures and tables, it can be seen that KNNI-AP, 1NN-AP, and KNNM-AP reduce to regular AP and KNNI-FCM reduces to regular FCM for 0% missing data. For other cases, different methods for handling missing attributes in AP and FCM lead to different clustering results. The different algorithms result in different misclassification ratio and FM index for the different algorithms. With the growth rate of missing data, the uncertainty of dataset

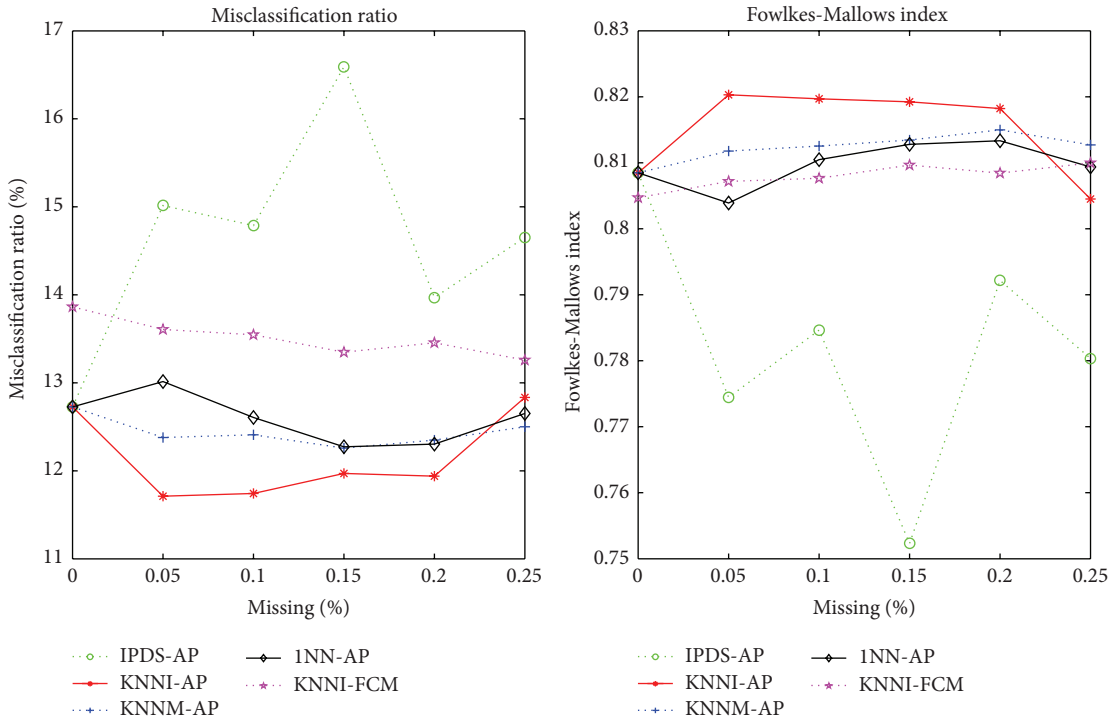

Figure 4: Averaged clustering results of 30 trials using incomplete Wholesale dataset.

increases; therefore, the misclassification ratio increases and FM decreases generally. However, because of handing the missing attributes, the clustering results of algorithms for incomplete data sometimes may be similar to or better than the the results of complete data.

In terms of misclassification ratio, KNNI-AP is always the best performer except for incomplete Wine dataset with 25% missing attributes, incomplete WDBC dataset with 10% missing attributes, and incomplete Wholesale customers dataset with 25% missing attributes. In the three cases, KNNI-AP almost gives suboptimal solutions beside the last case where the result of KNNI-AP is better than the results of IPDS-AP and KNNI-FCM. As for the FM index, KNNI-AP is always the best performer except for incomplete Iris and Wine datasets with 25% missing attributes, incomplete WDBC dataset with 10% missing attributes, and incomplete Wholesale dataset with 25% missing attributes. In the four cases, KNNI-AP almost gives suboptimal solutions beside the last case where the result of KNNI-AP is better than the result of IPDS-AP. From figures and tables, in general, the larger the FM value is, the smaller the misclassification ratio is except for the 25% cases of incomplete Iris and Wholesale datasets. We use misclassification ratio and FM index based on external criteria to accurately evaluate the quality of clustering results.

Comparing KNNI-AP with IPDS-AP, 1NN-AP, and KNNM-AP, the methods are all based on AP algorithm. IPDS-AP ignores missing attributes in incomplete data and scales the partial distances by the reciprocal of the proportion of components used based on the range of feature values, in which the distribution information of missing attributes implicitly embodied in the other data is not taken into account. 1NN-AP substitutes the missing attribute by the corresponding attribute of the nearest neighbor, in which AP algorithm is used to handle the complete dataset. Similarly, missing attributes are supplemented by the mean value of the attributes in the KNNM-AP. Compared with IPDS-AP, KNNI-AP uses the attribute distribution information of datasets sufficiently, including complete data and nonmissing attributes of incomplete data, in which the missing attributes are represented by KNNI on the basis of IPDS. Compared with the other two methods, KNNI-AP achieves interval estimation of missing attributes, taking advantage of the improved IAP, which represents the uncertainty of missing attributes and makes the representation more robust. Furthermore, cluster algorithm with interval data has advantages over cluster with point data, which can present the uncertainty of missing attributes to some degree, thus resulting in more accurate clustering performance.

Comparing KNNI-AP with KNNI-FCM, the methods are both based on KNNI, and the difference between them is the clustering algorithm they use. AP has the advantage that it works for any meaningful measure of similarity between data samples. Unlike most prototype-based clustering algorithms (e.g., K-means), AP does not require a vector space structure and the exemplars are chosen among the observed data samples and are not computed as hypothetical averages of cluster samples. These characteristics make AP clustering particularly suitable for applications in many fields. From our experiments, clustering results of KNNI-AP are far better than those of KNNI-FCM. And in most cases, the methods

Table 1: Averaged results of 30 trials using incomplete Iris dataset.

Missing rate (%)	Misclassification ratio (%)					Fowlkes-Mallows index				
	IPDS	1NN	KNNM	KNNI	KNNI-FCM	IPDS	1NN	KNNM	KNNI	KNNI-FCM
0	10	7.33	7.33	7.33	10.67	0.8306	0.8668	0.8668	0.8668	0.8196
5	10.37	*8.51*	8.71	**7.69**	10.53	0.8259	*0.8504*	0.8477	**0.8616**	0.8214
10	8.96	*8.40*	8.67	**8.13**	10.56	0.8461	*0.8516*	0.8479	**0.8553**	0.8209
15	9.98	*8.76*	8.87	**8.49**	10.56	0.8342	*0.8469*	0.8454	**0.8504**	0.8227
20	10.51	*9.24*	9.31	**8.44**	10.27	0.8277	*0.8397*	0.8392	**0.8512**	0.8262
25	9.16	*9.40*	*9.40*	**9.02**	10.62	**0.8443**	0.8377	0.8375	*0.8428*	0.8235

Table 2: Averaged results of 30 trials using incomplete Wine dataset.

Missing rate (%)	Misclassification ratio (%)					Fowlkes-Mallows index				
	IPDS	1NN	KNNM	KNNI	KNNI-FCM	IPDS	1NN	KNNM	KNNI	KNNI-FCM
0	8.99	8.99	8.99	8.99	8.99	0.8303	0.8303	0.8303	0.8303	0.8291
5	16.94	9.24	*9.13*	**8.93**	9.19	0.7272	0.8257	*0.8277*	**0.8305**	0.8256
10	23.88	9.38	*9.07*	**8.99**	9.19	0.6553	0.8230	*0.8283*	**0.8299**	0.8259
15	24.72	9.69	*9.33*	**9.30**	9.61	0.6417	0.8178	*0.8284*	**0.8245**	0.8187
20	24.80	10.34	*9.78*	**9.66**	10.06	0.6338	0.8068	*0.8162*	**0.8183**	0.8111
25	29.92	10.93	**9.97**	*10.17*	10.22	0.6164	0.7970	**0.8133**	*0.8097*	0.8088

Table 3: Averaged results of 30 trials using incomplete WDBC dataset.

Missing rate (%)	Misclassification ratio (%)					Fowlkes-Mallows index				
	IPDS	1NN	KNNM	KNNI	KNNI-FCM	IPDS	1NN	KNNM	KNNI	KNNI-FCM
0	6.50	6.50	6.50	6.50	7.21	0.8866	0.8866	0.8866	0.8866	0.8758
5	*6.49*	*6.49*	6.50	**6.46**	7.19	*0.8868*	*0.8868*	0.8866	**0.8872**	0.8760
10	**6.43**	6.48	6.49	*6.47*	7.21	**0.8879**	0.8869	0.8868	*0.8871*	0.8758
15	6.64	6.65	*6.49*	**6.45**	7.21	0.8849	0.8847	*0.8871*	**0.8878**	0.8758
20	6.57	6.55	*6.47*	**6.46**	7.19	0.8863	0.8861	*0.8872*	**0.8875**	0.8761
25	6.66	6.46	*6.42*	**6.39**	7.21	0.8848	0.8873	*0.8880*	**0.8884**	0.8758

Table 4: Averaged results of 30 trials using incomplete Wholesale dataset.

Missing rate (%)	Misclassification ratio (%)					Fowlkes-Mallows index				
	IPDS	1NN	KNNM	KNNI	KNNI-FCM	IPDS	1NN	KNNM	KNNI	KNNI-FCM
0	12.73	12.73	12.73	12.73	13.86	0.8084	0.8084	0.8084	0.8084	0.8047
5	15.01	13.02	*12.38*	**11.71**	13.61	0.7745	0.8039	*0.8118*	**0.8203**	0.8071
10	14.79	12.61	*12.41*	**11.74**	13.55	0.7846	0.8105	*0.8125*	**0.8197**	0.8076
15	16.59	12.27	*12.26*	**11.97**	13.35	0.7523	0.8128	*0.8134*	**0.8192**	0.8096
20	13.97	*12.30*	12.35	**11.94**	13.45	0.7922	0.8133	*0.8149*	**0.8182**	0.8084
25	14.65	*12.65*	**12.50**	12.83	13.26	0.7803	0.8094	**0.8127**	0.8045	*0.8099*

based on AP algorithm are also better than KNNI-FCM, which show that AP makes the clustering results more stable and accurate.

5. Conclusion

In this paper, we have studied incomplete data clustering using AP algorithm and presented an AP clustering algorithm based on KNNI for incomplete data. The proposed algorithm is based on the IPDS, estimating the KNNI representation of missing attributes by using K-nearest neighbor interval principle. The proposed algorithm has three main advantages. First, missing attributes are represented by KNNI on the basis of IPDS, which are robust. Second, the interval estimations use the attribute distribution information of datasets sufficiently, which is superior in expressing the uncertainty of missing attributes and enhances the robustness of missing attributes representation. Third, AP algorithm does not require a vector space structure and the exemplars are chosen among the observed data samples and are not computed as hypothetical averages, which makes the clustering results more stable and accurate than other center algorithms.

The results reported in this paper show that our proposed KNNI-AP algorithm is general, simple, and appropriate for the AP clustering with incomplete data. It can be understood that the final clustering results depend on the choice of K and P for KNNI-AP. In the future, our work will focus on the selection of K and P with theoretical basis and the improvement on the similarity measurement of AP when the missing percentage is large, which will be helpful to extend KNNI-AP to solve clustering incomplete data with various missing percentages.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China under Grant nos. 41427806 and 61273233, the Research Fund for the Doctoral Program of Higher Education under Grant nos. 20120002110035 and 20130002130010, the Project of China Ocean Association under Grant no. DY125-25-02, and Tsinghua University Initiative Scientific Research Program under Grant no. 20131089300.

References

[1] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[3] B. J. Frey and D. Dueck, "Response to comment on 'clustering by passing messages between data points'," *Science*, vol. 319, no. 5864, p. 726d, 2008.

[4] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, "Adaptive affinity propagation clustering," *Acta Automatica Sinica*, vol. 33, no. 12, pp. 1242–1246, 2007.

[5] Y. He, Q. Chen, X. Wang, R. Xu, X. Bai, and X. Meng, "An adaptive affinity propagation document clustering," in *Proceedings of the 7th International Conference on Informatics and Systems (INFOS '10)*, pp. 1–7, IEEE, March 2010.

[6] R. J. Little and D. B. Rubin, *Statistical Analysis with Nissing Data*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[7] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 10, pp. 617–621, 1979.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, 1981.

[10] S. Miyamoto, O. Takata, and K. Umayahara, "Handling missing values in fuzzy c-means," in *Proceedings of the 3rd Asian Fuzzy Systems Symposium*, pp. 139–142, 1998.

[11] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 5, pp. 735–744, 2001.

[12] R. J. Hathaway and J. C. Bezdek, "Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm," *Pattern Recognition Letters*, vol. 23, no. 1–3, pp. 151–160, 2002.

[13] D. Li, H. Gu, and L. Zhang, "A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data," *Expert Systems with Applications*, vol. 37, no. 10, pp. 6942–6947, 2010.

[14] D.-Q. Zhang and S.-C. Chen, "Clustering incomplete data using kernel-based fuzzy C-means algorithm," *Neural Processing Letters*, vol. 18, no. 3, pp. 155–162, 2003.

[15] C. Lu, S. Song, and C. Wu, "Afinity propagation clustering with incomplete data," in *Computational Intelligence, Networked Systems and Their Applications*, pp. 239–248, Springer, Berlin, Germany, 2014.

[16] J. L. Ohmann and M. J. Gregory, "Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A.," *Canadian Journal of Forest Research*, vol. 32, no. 4, pp. 725–741, 2002.

[17] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, Clustering, and Data Mining Applications*, pp. 639–647, Springer, Berlin, Germany, 2004.

[18] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[19] X. Huang and Q. Zhu, "A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets," *Pattern Recognition Letters*, vol. 23, no. 13, pp. 1613–1622, 2002.

[20] "UCI machine learning repository," 2014, http://archive.ics.uci.edu/ml/datasets.html.

[21] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: part I," *ACM SIGMOD Record*, vol. 31, no. 2, pp. 40–45, 2002.