Missing Categorical Data Imputation for FCM Clusterings of Mixed Incomplete Data

Takashi Furukawa
Graduate school of Engineering
Hokkai-Gakuen University
Sapporo, Japan
Email: 6512103t@hgu.jp

Shin-ichi Ohnishi
Faculty of Engineering
Hokkai-Gakuen University
Sapporo, Japan
Email: ohnishi@hgu.jp

Takahiro Yamanoi
Faculty of Engineering
Hokkai-Gakuen University
Sapporo, Japan
Email: yamanoi@hgu.jp

***Abstract*—The Data mining is related to human congnitive ability, and one of popular method is fuzzy clustering. The focus of fuzzy c-means (FCM) clustering method is normally used on numerical data. However, most data existing in databases are both categorical and numerical. To date, clustering methods have been developed to analyze only complete data. Although we, sometimes, encounter data sets that contain one or more missing feature values (incomplete data) in data intensive classification systems, traditional clustering methods cannot be used for such data. Thus, we study this theme and discuss clustering methods that can handle mixed numerical and categorical incomplete data. In this paper, we propose some algorithms that use the missing categorical data imputation method and distances between numerical data that contain missing values. Finally, we show through a real data experiment that our proposed method is more effective than without imputation, when missing ratio becomes higher.**

Keywords-clustering; incomplete data; mixed data; FCM.

I. INTRODUCTION

Clustering is the most popular method for discovering group and data structures in datasets in data intensive classification systems. Fuzzy clustering allows each datum to belong to some clusters. Thus data are classified into an optimal cluster accurately [1]. The k-means algorithm is the most popular algorithm used in scientific and industrial applications because of its simplicity and efficiency. Whereas k-means gives satisfactory results for numeric attributes, it is not appropriate for data sets containing categorical attributes because it is not possible to find a mean of categorical value. Although, traditional clustering methods handle only numerical data, real world data sets contain mixed (numerical and categorical) data. Therefore, traditional clustering methods cannot be applied to mixed data sets. Recently, clustering methods that deal with mixed data sets have been developed [4][5].

Moreover, when we analyze real world data sets, we encounter incomplete data. Incomplete data are found for example through data input errors, inaccurate measures, and noise. Traditional clustering methods cannot be directly applied to data sets that contain incomplete data, so we need to treat such data. A common approach to analyzing data with missing values is to remove attributes and/or instances with large fractions of missing values. However, this approach excludes partial data from analytical consideration and hence compromises the reliability of results. Therefore, we need analytical tools that handle incomplete categorical data, a process that is called imputation. To date, many imputation methods have been proposed, but most apply only to numerical variables. Thus, when analyzing categorical data or mixed data containing missing values, one has to eliminate from consideration data with missing values. Moreover, an imputation method applicable to fuzzy clustering is rare.

Fuzzy c-means (FCM) clustering is a very popular fuzzy extension of k-means. However, FCM for mixed data cannot be applied to data that contains missing data. Therefore, we use the imputation method for missing categorical data, and then we apply FCM clustering for mixed data. If we encounter missing numerical data, we use the partial distance [7] instead of the Euclidean distance.

In this paper, we describe the development of a fuzzy clustering algorithm for mixed data with missing numerical and categorical data. The next section introduces the FCM algorithm. Section III presents the clustering algorithm for mixed data. Sections IV and V introduce the missing categorical imputation method, and the notion of distance between data that contain missing values. Section VI proposes a fuzzy clustering algorithm that can treat mixed incomplete data. Finally, Section VI shows through a real data experiment that our proposed method is more effective than without imputation, when missing ratio becomes higher.

II. FUZZY c-MEANS CLUSTERING

The FCM algorithm proposed by Dunn [1] and extended by Bezdek [2] is one of the most well-known algorithms in fuzzy clustering analysis. This algorithm uses the squared-norm to measure similarities between cluster centers and data points. It can only be effective in clustering spherical clusters. To cluster more general datasets, a number of algorithms have been proposed by replacing the squared-norm with other similarity measures [3]. The notation that we use throughout is as follows. Let $\boldsymbol{x}_i = (x_{ij}), i = 1, \ldots, n, j = 1, \ldots, m$ is a feature value of the i^{th} data vector, c is the number of clusters. $\boldsymbol{b}_c = (b_{c1}, \ldots, b_{cm})^{\mathsf{T}}$ is the cluster center of the c^{th} cluster, u_{ci} is the degree to which x_i belongs to the c^{th} cluster. Then, u_{ci} satisfies the following constraint

$$\sum_{c=1}^{C} u_{ci} = 1, \ i = 1, \ldots, n \tag{1}$$

 ISBN: 978-1-61208-340-7

The FCM algorithm for solving (2) alternates the optimizations of L_{fcm} over the variables u and b

$$L_{fcm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\left(\sum_{j=1}^{m}(x_{ij}-b_{cj})^2\right) \quad (2)$$

where θ is the fuzzification parameter ($\theta > 1$). Minimizing the u values of (2) are less fuzzy for values of θ near 1 and fuzzier for large values of θ. The choice $\theta = 2$ is widely accepted as a good choice of fuzzification parameter.

III. Fuzzy c-means clustering for mixed databases

The FCM algorithm has been widely used and adapted. However, only numerical data can be treated; categorical data cannot. When we analyze categorical data, we have to implement a quantification of such data. For example, suppose we obtained n sample data that have m categorical data consisting of K_j categories.

Then, the j^{th} item data can be expressed as an $(n \times K_j)$ dummy variable matrix $G_j = \{g_{ijk}\}, i = 1,\ldots,n, k = 1,\ldots,K_j$

$$g_{ijk} = \begin{cases} 1, & \text{data } i \text{ contains category } k \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Honda et al. proposed a method that combined the quantification of categorical data and the fuzzy clustering of numerical data [6]. The variables up to $(m-q)$ are numerical; the rest is categorical. Calculating

$$L = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\left(\sum_{j=1}^{m-q}(x_{ij}-b_{cj})^2 + \sum_{j=m-q+1}^{m}(\boldsymbol{g}_{ij}^{\mathsf{T}}\boldsymbol{q}_j - b_{cj})^2\right) \quad (4)$$

where $\boldsymbol{q}_j$ is a categorical score, which can be computed as follows

$$\boldsymbol{q}_j = \left(G_j^{\mathsf{T}}\left(\sum_{c=1}^{C} U_c^{\theta}\right)G_j\right)^{-1}\left(\sum_{c=1}^{C}\boldsymbol{b}_{cj}G_j^{\mathsf{T}}U_c^{\theta}\mathbf{1}_n\right) \quad (5)$$

To obtain a unique solution, we impose the following constraint.

$$\mathbf{1}_n^{\mathsf{T}}G_j\boldsymbol{q}_j = 0 \quad (6)$$

$$\boldsymbol{q}_j^{\mathsf{T}}G_j^{\mathsf{T}}G_j\boldsymbol{q}_j = n \quad (7)$$

Algorithm: Fuzzy c-means algorithm for mixed databases

1. Initialize membership $u_{ci}, c = 1,\ldots,C, i = 1,\ldots,n$ and cluster center $b_{cj}, c = 1,\ldots,C$, then normalize u_{ci} satisfying (1).
2. Update category score $\boldsymbol{q_j}, j = m-q+1,\ldots,m$, using (5) according to constraint conditions (6) and (7). We then interpret $\boldsymbol{g}_{ij}^{\mathsf{T}}\boldsymbol{q}_j$ as the j^{th} numerical score x_{ij}.
3. Update cluster center b_{cj} using

$$b_{cj} = \frac{\sum_{i=1}^{n} u_{ci}^{\theta}x_{ij}}{\sum_{i=1}^{n} u_{ci}^{\theta}} \quad (8)$$

4. Update membership u_{ci} using

$$u_{ci} = \left(\sum_{l=1}^{C}\left(\frac{D_{ci}}{D_{li}}\right)^{\frac{1}{\theta-1}}\right)^{-1} \quad (9)$$

where

$$D_{ci} = \|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2 \quad (10)$$

If $\boldsymbol{x}_i = \boldsymbol{b}_c$, $u_{ci} = 1/C_i$

5. Let ϵ judgment value for convergence. Compare u_{ci}^{NEW} to u_{ci}^{OLD} using

$$\max_{c,i}\|u_{ci}^{\mathrm{NEW}} - u_{ci}^{\mathrm{OLD}}\| < \epsilon \quad (11)$$

If true then stop, otherwise return to Step 2.

IV. Missing categorical data imputation method

Recently, missing data imputation has been recognized and developed as an important task. However, we are not accustomed to combining the clustering algorithm and the imputation method. Most missing data imputations are restricted to only numerical data. There are a few methods that permit missing categorical data or mixed data imputation [8][9]. If attributes and/or instances are missing, we do not apply the clustering algorithm. Instead, we apply the imputation method to fill the missing values, and then we can apply the clustering algorithm. In this paper, we use the missing categorical data imputation method, a"novel rough set model based on similarity", as proposed by Sen et al. [7].

DEFINITION 1. (Missing Attribute Set) An incomplete information system is denoted $S =< U, A, V, f >$; with attribute set $A = \{a_k | k = 1,2,\ldots,m\}$; V is the domain of the attribute. $V = V_k$, V_k is the domain of the attribute a_k, which is the category value. $a_k(x_i)$ is the value of attribute a_k of object x_i, and "$*$" means missing value. The missing attribute set (MAS) of object x_i is defined as follows:

$$MAS_i = \{k \mid a_k(x_i) = *, k = 1,2,\ldots,m\}$$

DEFINITION 2. (Similarity between objects) For two objects $x_i \in U$ and $x_j \in U$, their similarity $P_k(x_i, x_j)$ of attribute a_k is defined as

$$P_k(x_i,x_j) = \begin{cases} 1, a_k(x_i) = a_k(x_j) \wedge a_k(x_i) \neq * \wedge a_k(x_j) \neq * \\ 0, a_k(x_i) \neq a_k(x_j) \vee a_k(x_i) = * \vee a_k(x_j) = * \end{cases} \quad (12)$$

Then the similarity of the two objects of all attributes is defined as:

$$P(x_i,x_j) = \begin{cases} 0, \exists a_k \in A(a_k(x_i) \neq a_k(x_j) \wedge a_k(x_i) \neq * \\ \wedge a_k(x_j) \neq * \\ \sum_{k=1}^{m} P_k(x_i,x_j), others \end{cases} \quad (13)$$

The similarity matrix is $M(i,j) = P(x_i, x_j)$.

DEFINITION 3. (Nearest undifferentiated set (NS) of an object) The NS of object $x_i \in U$ is defined as a set NS_i of objects that have a maximum similarity:

$$NS_i = \{j \mid (M(i,j) = \max_{x_k \wedge k \neq i}(M(i,k))) \wedge M(i,j) > 0\}$$

Algorithm: Missing Categorical Data Imputation

1. Set parameter $num = 0$ to record the quantity of imputation data in the current iteration; for all the $x_i \in U$, if x_i has missing attribute, compute its missing attribute set MAS_i and nearest undifferentiated set NS_i;

2. For all the objects x_i that have missing attributes, which means $MAS_i \neq \phi$, do the perform loop for all the $k \in MAS_i$ in order:

2.1 if $|NS_i| = 0$,
break(to deal with the next missing attribute object);

2.2 if $|NS_i| = 1$, assume $j \in NS_i$ and $a_k(x_j) \neq *$, then:

$$a_k(x_i) = a_k(x_j);$$

$num++$;

2.3 if $|NS_i| \geq 2$,

2.3.1 If there exists $m, n \in NS_i$ satisfied $(a_k(x_m) \neq *) \wedge (a_k(x_n) \neq *) \wedge (a_k(x_m) \neq a_k(x_n))$, set:

$$a_k(x_i) = *;$$

2.3.2 Otherwise, if there exists $j_0 \in N$ and $a_k(x_{j0})$: $num++$;

3. if $num > 0$, return to Step 1, otherwise, go to step 4;

4. End. Other methods can be used.

V. Distances between data that contain missing values

In some situations, the feature vectors in $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ can have missing feature values. Any data with some missing feature values are called incomplete data. The original FCM algorithm and the FCM algorithm for mixed databases is a useful tool, but it is not directly applicable to data that contain missing values. Hathaway et al. proposed four approaches to incomplete data[7]: the whole data strategy (WDS), the partial distance strategy (PDS), the optimal completion strategy (OCS), and the nearest prototype strategy (NPS). In WDS, if the proportion of incomplete data is small, then it may be useful to simply delete all incomplete data and apply FCM to the remaining complete data. WDS should be used only if $\frac{n_p}{n_x} \leq 0.75$, where $n_p = |X_P|$ and $n_s = |X| \cdot m$. The cases when missing values $\|X_M\|$ are sufficiently large that the use of the WDS cannot be justified entails calculating partial (squared Euclidean) distances using all available (non-missing) feature values, and then scaling this quantity by the reciprocal of the proportion of components used. For this study, we used the PDS approach for mixed databases containing incomplete data.

In the PDS approach, the general formula for the partial distance calculation of D_{ci} is

$$D_{ci} = \frac{m}{I_i} \sum_{j=1}^{m} (x_{ij} - b_{cj})^2 I_{ij} \tag{14}$$

where

$$I_{ij} = \begin{cases} 0 & (x_{ij} \in X_M) \\ 1 & (x_{ij} \in X_P) \end{cases} \text{ for } 1 \leq i \leq n, 1 \leq j \leq m \tag{15}$$

$$I_i = \sum_{j=1}^{m} I_{ij} \tag{16}$$

$$X_P = \{x_{ij} | \text{ the value for } x_{ij} \text{ is present in } X\}$$

$$X_M = \{x_{ij} | \text{ the value for } x_{ij} \text{ is missing from } X\}$$

For example, let $m = 3$ and $n = 4$. Denoting missing values by *,

$$X = \begin{bmatrix} 1 \\ * \\ * \\ 4 \\ * \end{bmatrix}$$

Then, $X_P = \{x_1 = 1, x_4 = 4\}$, $X_M = \{x_2, x_3, x_5\}$, and

$$\begin{aligned} D_{ci} &= \|\boldsymbol{x}_i - \boldsymbol{b}_c\|_2^2 \\ &= \|(1 \ * \ * \ 4 \ *)^{\mathsf{T}} - (5 \ \ 6 \ \ 7 \ \ 8 \ \ 9)^{\mathsf{T}}\|_2^2 \\ &= \frac{5}{(5-3)}((1-5)^2 + (4-8)^2) \end{aligned} \tag{17}$$

The PDS version of the FCM algorithm, is obtained by making two modifications of the FCM algorithm. First, we calculate D_{ci} in (10) for incomplete data according to (14) – (16). Second, we replace the calculation of $\boldsymbol{b}$ in (8) with

$$b_{cj} = \frac{\sum_{i=1}^{n} u_{ci}^{\theta} I_{ij} x_{cj}}{\sum_{i=1}^{n} u_{ci}^{\theta} I_{ij}} \tag{18}$$

VI. FCM for mixed databases with incomplete data

For clustering analysis, treating missing data becomes especially important when the fraction of missing values is large and the data are of mixed type. We combine the FCM algorithm for mixed databases with the imputation method and the PDS approach to construct a FCM algorithm for mixed databases containing missing values. Here, we assume incomplete mixed data x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, the values up to $m - q$ correspond to numerical data and the rest is categorical. The dummy valuable matrix $G_j = \{g_{ijk}\}$, $k = 1, \ldots, K_j$, is described in equation (3). Applying the FCM algorithm to mixed databases that contain incomplete data is considered as follows:

Algorithm: FCM for mixed databases containing incomplete data

1. If there are missing categorical data, use the imputation algorithm described in Section IV, and separate the complete categorical data $x_{ij} (i = 1, \ldots, n, j = m - q + 1, \ldots, m)$

2. Initialize membership u_{ci} and cluster center b_{cj}, then normalize u_{ci} satisfying $\sum_{i=1}^{n} u_{ci} = 1$, $i = 1, \ldots, n$.

3. Update the category score

$$\boldsymbol{q}_j = \left(G_j^{\mathsf{T}} \left(\sum_{c=1}^{C} U_c^{\theta} \right) G_j \right)^{-1} \left(\sum_{c=1}^{C} \boldsymbol{b}_{cj} G_j^{\mathsf{T}} U_c^{\theta} \mathbf{1}_n \right) \tag{19}$$

according to the following constraint conditions:

$$\mathbf{1}_n^{\mathsf{T}} G_j \boldsymbol{q}_j = 0 \tag{20}$$

 ISBN: 978-1-61208-340-7

$$\boldsymbol{q}_j^{\mathsf{T}} G_j^{\mathsf{T}} G_j \boldsymbol{q}_j = n. \tag{21}$$

We interpret $\boldsymbol{g}_{ij}^{\mathsf{T}} \boldsymbol{q}_j$ to be the j^{th} numerical score x_{ij}.

4. Update cluster center b_{cj} using

$$b_{cj} = \frac{\sum_{i=1}^{n} u_{ci}^{\theta} I_{ij} x_{cj}}{\sum_{i=1}^{n} u_{ci}^{\theta} I_{ij}} \tag{22}$$

5. Update membership u_{ci} using

$$u_{ci} = \left(\sum_{l=1}^{C} \left(\frac{D_{ci}}{D_{li}} \right)^{\frac{1}{\theta-1}} \right)^{-1} \tag{23}$$

where D_{ci} is calculated form

$$D_{ci} = \frac{m}{I_i} \sum_{j=1}^{m} (x_{ij} - b_{cj})^2 I_{ij} \tag{24}$$

6. Let ϵ be a set value to judge convergence. Then compare u_{ci}^{NEW} to u_{ci}^{OLD} using

$$\max_{c,i} \| u_{ci}^{\mathrm{NEW}} - u_{ci}^{\mathrm{OLD}} \| < \epsilon \tag{25}$$

If true, then stop, otherwise return to Step 3.

VII. Experimental Results

In this section, we show the performance of our algorithm for mixed incomplete data.

We use credit approval datasets from UCI Machine Learning Repository which have 683 samples, 15 attributes(6 is numerical and the rest categorical), and 53 missing values. Table I lists the type of attribute("N" is numerical and "C" is categorical) and the number of missing values. This database has its own real classification result, i.e., each sample has been classified into 2 groups "+" or "−".

Fig. 1 the result for this incomplete mixed data using the proposed fuzzy clustering method; The number of samples with membership value over 10 intervals between 0 and 1 are found. 77% of the "+" group samples have high membership in cluster1 and almost all of "−" group samples are strongly classified in cluster 2. The fuzzification parameter θ is 1.2. The result shows that our proposed method is applicable for these real data.

Next, we compared following four methods; (I) Using Imputation method for categorical missing values, PDS for non-imputation missing values and numerical missing values (Imp + PDS). (II) Using imputation method for categorical missing values, WDS for non-imputation missing values and PDS for numerical missing values (Imp + PDS + WDS). (III) Using PDS for all missing values (PDS). (IV) Using WDS for all missing values (WDS).

Fig. **??** shows the results of these 4 methods. This graph's positive area indicate numbers of "+" group samples that have membership value to cluster1, and area of negative indicate "-" group samples data to cluster2. Results of (II) and (IV) have enough high membership samples (from 0.9 to 1.0).

Finally, we change missing ratio (from 0.7% to 0.9% and 1.2%) and apply four methods. Results are shown in Fig. 3 to 6. (II)Imp+PDS+WDS and (III)PDS cannot classify enough when missing ratio is high (1.2%) in Fig. 4 and 5. Further in Fig. 6, (IV)WDS have a lot of samples that cannot be used in any missing ratio (especially in high missing ratio), because this method except all sample data which contain missing values(describe value of 0 in each graph). From these point of view, (I) Imp+PDS can be better method for these dataset as shown in Fig. 3.

TABLE I. ATTRIBUTES AND MISSING VALUES

Attribute	type	Category	Missing
A_1	C	2	12
A_2	N	-	12
A_3	N	-	0
A_4	C	4	6
A_5	C	3	6
A_6	C	14	6
A_7	C	9	6
A_8	N	-	0
A_9	C	2	0
A_{10}	C	2	0
A_{11}	N	-	0
A_{12}	C	2	0
A_{13}	C	3	0
A_{14}	N	-	13
A_{15}	N	-	0

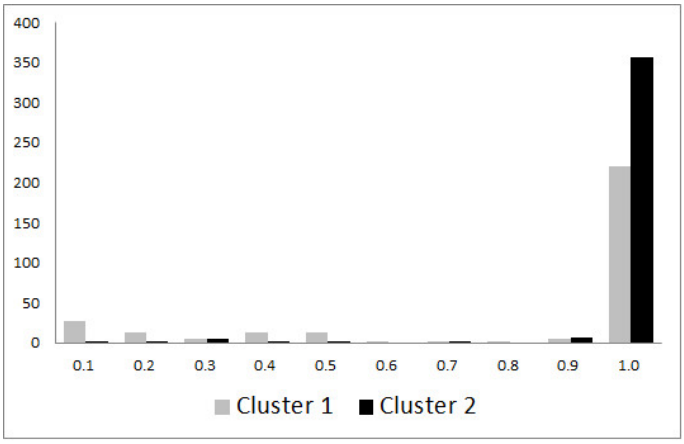

Fig. 1. Fuzzy Clustering Result(Real data)

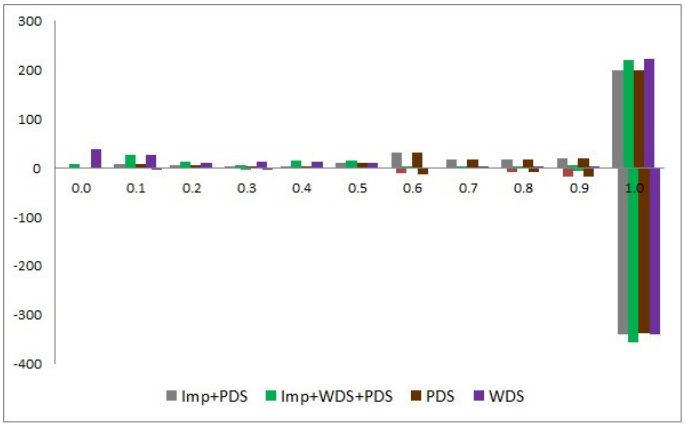

Fig. 2. Comparing four methods

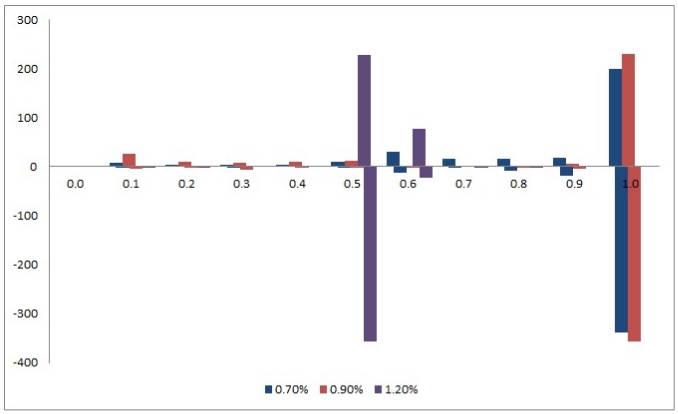

Fig. 3. Missing Result of PDS (I)

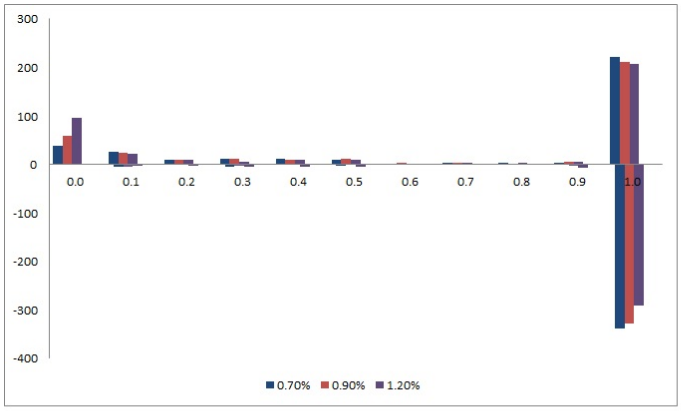

Fig. 4. Missing result of WDS (II)

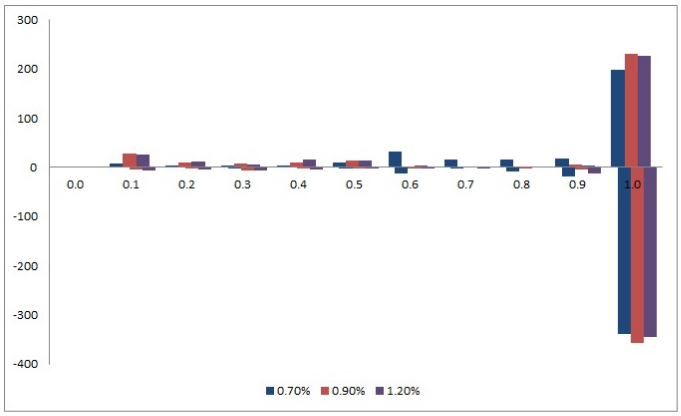

Fig. 5. Missing result of Imp + PDS (III)

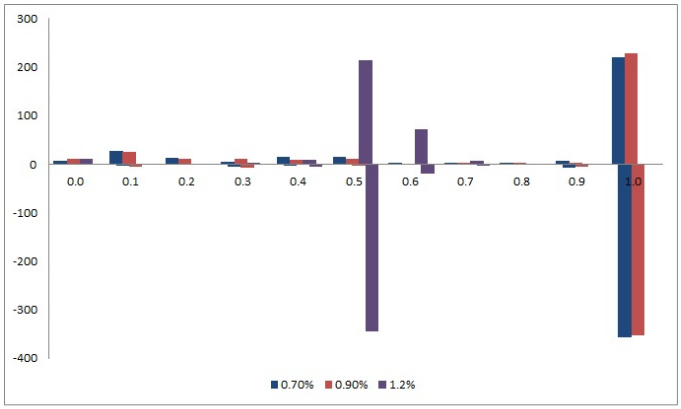

Fig. 6. Missing result of Imp + WDS + PDS (IV)

VIII. CONCLUSION

In this paper, we discussed a FCM clustering algorithm that handles mixed data containing missing values. In our study, we applied the imputation method to missing categorical data before clustering, followed by the FCM clustering algorithm. When we encountered numerical missing data, we used the PDS (and WDS) distance for numerical missing data. A real data experiment shows that our proposed method is more effective than without imputation, when missing ratio becomes higher. To obtain better performance during the clustering analysis for mixed data containing missing values, we plan to apply our algorithm to another datasets, too.

REFERENCES

[1] J. C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3: 32-57, 1973.

[2] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981

[3] W. Sen, F. Xiaodong, H. Yushan, and W. Qiang, Missing Categorical Data Imputation Approach Based on Similarity, IEEE International conference on Systems, Man, and Cybernetics, 2012.

[4] Y. Naija, K. Blibech, S. Chakhar, and R. Robbana, Extension of Partitional Clustering Methods for Handling Mixed Data, IEEE International Conference on Data Mining Workshop, 2008.

[5] K. L. Wu and M. S. Yang, Alternative c-means clustering algorithms, Pattern Recognition vol. 35, pp. 2267-2278, 2002.

[6] K. Honda and H. Ichihashi, Fuzzy c-means clustering of mixed databases including numerical and nominal variables, Cybernetics and Intelligent Systems, 2004 IEEE Conference on, vol. 1, 2004

[7] R. J. Hathaway and J. C. Bezdek, Fuzzy c-means Clustering of Incomplete Data, IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol.31, No. 5, pp.735-744, 2001

[8] K. Bache and M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science [retrieved: Feb. 2014]

 ISBN: 978-1-61208-340-7