Proceedings of 2013 International Conference on Fuzzy Theory and Its Application
National Taiwan University of Science and Technology, Taipei, Taiwan, Dec. 6-8, 2013

A study on a fuzzy clustering for mixed numerical and categorical 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*—Most clustering methods focus 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), 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 an algorithm that uses the missing categorical data imputation method and distances between numerical data that contain missing values. Furthermore, we apply fuzzy clustering for interpreting results that are vague.**

I. Introduction

Clustering is the most popular method for discovering group and data structures in datasets. It is used for example in data mining and web mining. 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 the 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 fuzzy c-means clustering for mixed data. If we encounter missing numerical data, we use the PDS 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 fuzzy c-means algorithm. Section III presents the clustering algorithm for mixed data. Section IV introduces the missing categorical imputation method, and Section V introduces the notion of distance between data that contain missing values. Section VI proposes a fuzzy clustering algorithm that can treat mixed incomplete data.

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

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

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

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

where θ is the fuzzification parameter ($\theta > 1$). Minimizing the u values of (2) are less fuzzy for values of θ near 1 and

978-1-4799-0386-3/13/$31.00 ©2013 IEEE

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 $G_j = \{g_{ijk}\}, i = 1, \ldots, n, k = 1, \ldots, K_j$

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

Honda *et al.* proposed a method that combined the quantification of categorical data and the fuzzy clustering of numerical data[5]. The variables up to $(m - q)$ are numerical; the rest is categorical. Calculating

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

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

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

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

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

$$\boldsymbol{q}_j^{\mathsf{T}} G_j^{\mathsf{T}} G_j \boldsymbol{q}_j = n \tag{7}$$

Algorithm: Fuzzy c-means algorithm for mixed databases

Step1. 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).

Step2. Update category score $\boldsymbol{q_j}, j = m-q+1, \ldots, m$, using equation (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}.

Step3. Update cluster center b_{cj} using

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

Step4. Update membership u_{ci} using

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

where

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

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

Step5. Let ϵ judgment value for convergence. Compare u_{ci}^{NEW} to u_{ci}^{OLD} using

$$\max_{c,i} \|u_{ci}^{\text{NEW}} - u_{ci}^{\text{OLD}}\| < \epsilon \tag{11}$$

If true, then stop, otherwise return to Step2.

IV. Missing categorical data imputation method

Recently, missing data imputation has been recognized as important task and has developed. However, we are unfamiliar with 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, "novel rough set model based on similarity", as proposed by Sen *et al*[7].

DEFINITION1. (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:

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

DEFINITION2. (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

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

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

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

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

DEFINITION3. (Nearest undifferentiated set (NS) of an object) The nearest undifferentiated set 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

Step1. 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;

Step2. For all the objects x_i that have missing attributes, which means $MAS_i \neq \phi$, do the loop to 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 $a_k(x_{j0})$ :
         $num++;$
Step3.   if $num > 0$, return to Step1, otherwise, go to step4;
Step4.   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 FCM algorithm and FCM algorithm for mixed databases algorithm 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[6]: 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 might 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 paper, we use 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

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

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

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

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

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 (PDSFCM), is obtained by making two modifications of the FCM algorithm. First, we calculate D_{ci} in (10) for incomplete data according to (14) through (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

Step1. If there are missing categorical data, use the imputation algorithm described in Section IV, and get the complete categorical data part $x_{ij} (i = 1, \ldots, n, j = m - q + 1, \ldots, m)$

Step2. Initialize membership u_{ci} and cluster center b_{cj}, then normalize u_{ci} satisfying $\sum_{i=1}^{n} u_{ci} = 1$, $i = 1, \ldots, n$.

Step3. Update the category score

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

according to follow constraint conditions.

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

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

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

Step4. Update cluster center b_{cj} using

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

Step5. Update membership u_{ci} using the following.

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

where D_{ci} is calculated using

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

Step6. Let ϵ judgment value for 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 Step3.

VII. An Experimental Result

In this section, we show the performance of our algorithm for mixed incomplete data. We use 20 artificial samples that have seven attributes(a_1 to a_5 are numerical; the rest is categorical). Each categorical item has two categories (denoted 0 or 1). We employ data sets, in which 25% of the samples have missing attributes, as incomplete data sets.

Table II lists the clustering results to compare complete and incomplete data. Values in bold font correspond to a cluster that the sample belongs to most. This result shows that the clustering algorithm is useful enough for incomplete data.

TABLE I. Complete data

	a_1	a_2	a_3	a_4	a_5	a_6	a_7
x_1	51	6	20	5	53	2	2
x_2	5	49	4	10	50	0	0
x_3	5	4	49	50	10	1	2
x_4	5	51	4	9	50	2	0
x_5	49	5	20	4	49	2	1
x_6	5	4	50	49	9	1	0
x_7	5	50	4	10	50	0	2
x_8	50	5	19	5	49	1	0
x_9	5	51	3	9	50	1	0
x_{10}	49	5	20	5	49	2	2
x_{11}	6	4	51	50	10	2	1
x_{12}	50	4	21	5	50	2	1
x_{13}	4	3	50	50	10	2	2
x_{14}	5	4	49	50	10	0	1
x_{15}	4	50	5	11	49	0	2
x_{16}	50	6	20	5	50	1	0
x_{17}	5	4	50	50	10	0	1
x_{18}	5	49	4	10	51	0	2
x_{19}	5	5	51	51	11	0	0
x_{20}	51	5	20	6	50	2	2

TABLE II. Fuzzy Clustering Result

	complete data			incomplete data		
x_1	0.078	0.374	**0.548**	0.233	0.368	**0.399**
x_2	0.217	**0.401**	0.382	0.244	**0.394**	0.362
x_3	0.217	**0.401**	0.382	0.244	**0.394**	0.362
x_4	0.208	**0.403**	0.389	0.235	**0.398**	0.367
x_5	0.073	0.377	**0.550**	0.073	0.447	**0.479**
x_6	**0.754**	0.140	0.106	**0.748**	0.120	0.132
x_7	0.204	**0.404**	0.392	0.168	**0.431**	0.401
x_8	0.071	0.377	**0.553**	0.055	0.459	**0.485**
x_9	**0.749**	0.142	0.109	**0.743**	0.122	0.134
x_{10}	0.074	0.380	**0.545**	0.081	0.445	**0.474**
x_{11}	**0.754**	0.140	0.107	**0.747**	0.120	0.133
x_{12}	0.076	0.376	**0.548**	0.079	0.444	**0.476**
x_{13}	**0.750**	0.140	0.110	**0.743**	0.122	0.135
x_{14}	0.071	0.378	**0.551**	0.073	0.448	**0.479**
x_{15}	0.208	**0.403**	0.389	0.233	**0.398**	0.368
x_{16}	0.071	0.377	**0.551**	0.071	0.449	**0.481**
x_{17}	**0.758**	0.138	0.104	**0.752**	0.118	0.130
x_{18}	0.208	**0.403**	0.389	0.272	**0.379**	0.349
x_{19}	**0.758**	0.138	0.104	**0.421**	0.281	0.298
x_{20}	**0.404**	0.390	0.206	0.368	0.233	**0.399**

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, and then we used the FCM clustering algorithm. When we encountered numerical missing data, we used the PDS distance for numerical missing data. To obtain better performance during the clustering analysis for mixed data containing missing values, we plan to study other imputation methods for categorical incomplete data and other distance measures, for instance, WDS, OCS, and NPS, for missing numerical values. Furthermore, we will applying this algorithm to real data to check its efficiency.

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

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

[5] Zhexue Huang: Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery 2, pp.283-304, 1998

[6] R. J. Hathaway, 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

[7] Sen Wu, Xiaodong Feng, Yushan Han, Qiang Wang: Missing Categorical Data Imputation Approach Based on Similarity, IEEE International conference on Systems, Man, and Cybernetics, 2012.

[8] Yosr Naija, Kaouther Blibech, Salem Chakhar, Riadh Robbana: Extension of Partitional Clustering Methods for Handling Mixed Data, IEEE International Conference on Data Mining Workshop, 2008.