Fuzzy *K*-means Clustering with Missing Values

Manish Sarkar and Tze-Yun Leong
Department of Computer Science, School of Computing
National University of Singapore
Lower Kent Ridge Road, Singapore: 119260
{manish, leongty}@comp.nus.edu.sg

Fuzzy K-means clustering algorithm is a popular approach for exploring the structure of a set of patterns, especially when the clusters are overlapping or fuzzy. However, the fuzzy K-means clustering algorithm cannot be applied when the data contain missing values. In many cases, the number of patterns with missing values is so large that if these patterns are removed, then the number of patterns to characterize the data set is insufficient. This paper proposes a technique to exploit the information provided by the patterns with the missing values so that the clustering results are enhanced. There are various preprocessing methods to substitute the missing values before clustering the data. However, instead of repairing the data set at the beginning, the repairing can be carried out incrementally in each iteration based on the context. It is thus more likely that less uncertainty is added while incorporating the repair work. Fine-tuning the missing values using the information from other attributes further consolidates this scheme. Applications of the proposed method in medical domain have produced good performance.

Keywords: Fuzzy *K*-means clustering and missing values.

1. Introduction

Motivation: In medicine and biology, we often need exploratory analysis like grouping the patterns such that the patterns within the same cluster have a high degree of similarity, and the patterns from different clusters have a high degree of dissimilarity. Clustering can be formally defined as follows [1]: Given a set of data $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subseteq \mathbf{R}^N$, find an integer K ($2 \le K \le n$) and K number of partitions of X that exhibit categorically homogeneous subsets.

Importance of clustering: Some tasks for which the clustering algorithms can be employed are as follows:

1. Clustering can abstract or compress certain properties of the data set.
2. A classifier can be constructed through clustering. To build a classifier, we group a data set, and subsequently assign a class label (crisp or fuzzy) to each cluster. The class label of a new pattern is determined based on the cluster in which the pattern falls.
3. Clustering can be applied to decide whether the representation of a problem on computers is appropriate for processing. If the representation is not appropriate, then the data set behaves like a set of random numbers without any underlying regularity. In that case, the bad clustering results indicate that the user needs to modify the representation of the problem.

Basics of clustering: Three types of clustering approaches are commonly used [1]. They are (1) hierarchical approach, (2) graph theoretic approach, and (3) objective function-based approach. The objective function-based approach is very popular. One extensively used objective function-type clustering algorithm is *hard K-means clustering* algorithm [1]. It assigns each pattern exactly to one of the clusters assuming well-defined boundaries between the clusters. However, there may be some patterns that belong to more than one cluster. In order to overcome this problem, the idea of *fuzzy K-means* (FKM) algorithm has been introduced. Unlike the hard *K*-means, in the FKM each input pattern belongs to all the clusters with different degrees or membership values. Incorporation of the fuzzy theory in the FKM algorithm makes it a generalized version of the hard *K*-means algorithm. From the psycho-physiological point of view, the problem of pattern clustering is unsuitable for approaches with precise mathematical formulations.

However, the FKM algorithm cannot be applied to the real-life clustering problems when the data contain missing values. The missing values in a pattern imply that the values of some of the attributes of the pattern are unknown. Missing values can occur due to various reasons like (a) patient entries for some attributes are irrelevant or unknown, (b) in the questioning session, the patient did not want to provide the values, (c) errors have led to incomplete attributes, (d) random noises have led to some impossible values, and they have been removed intentionally, (e) patients have died before an experiment was finished.

Problem definition: This paper addresses how to apply the FKM algorithm efficiently in the presence of missing values. It is assumed that the values are missing at random, i.e., the probability of missing a value does not depend on the quantity of the value [6].

Related work: The approaches to deal with missing values can be categorized into the following groups [3] [4][5][6][7]:

Deductive imputation: Missing values are deduced with certainty, or with high probability from the other information of the pattern.

Hot-deck imputation: Missing values are replaced with values from the closest matching patterns.

Mean-value imputation: The mean of the observed values is used to replace the missing values.

Regression-based imputation: Missing values are replaced by the predicted values from a regression analysis.

Imputation using Expectation-Maximization: Missing values are repaired in two steps. In the E-step, the expected value of the loglikelihood is calculated, and in the M-step, the missing values are substituted by the expected values. Then the likelihood function is maximized as if no data were missing.

Overview of the proposed method: Most of the current methods repair or impute the missing values before the clustering starts. This paper attempts to repair the missing data while performing clustering. Exploiting this trick is difficult because while updating a cluster center, the distance between the pattern with missing values and the cluster center cannot be measured. Using the law of large numbers, if we assume that the distances between the cluster center and the patterns form a Gaussian distribution, then the distance between a pattern with missing values and the cluster center can be replaced by the weighted mean of the distances between the cluster center and the complete patterns. The missing values are further fine-tuned by exploiting the information from the other attributes.

2. Background

2.1 Fuzzy Sets

In traditional two-state classifiers, where a class C is defined as a subset of the universal set *X*, any input pattern $\boldsymbol{x} \in X$ can either be a member or not be a member of the given class C. This property of whether or not a pattern $\boldsymbol{x}$ of the universal set belongs to the class C can be defined by a characteristic function $\mu_C : X \to \{0,1\}$ as follows:

$$\mu_C(\boldsymbol{x}) = \begin{cases} 1 \ \textit{iff} \ \boldsymbol{x} \in C \\ 0 \ \text{otherwise} \end{cases} \qquad (1)$$

In real-life situations, boundaries between the classes may be overlapping. Hence, it is uncertain whether an input pattern belongs totally to the class C. To consider such situations, in fuzzy sets [1] the concept of the characteristic function has been modified to the fuzzy *membership function* $\mu_C : X \to [0,1]$. This function is called membership function because larger value of the function denotes more membership of the element to the set under consideration.

2.2. Fuzzy *K*-Means Clustering

Clustering a data set $X \subseteq \mathbf{R}^N$ implies that the data set is partitioned into K clusters such that each cluster is compact and far from other clusters. One way to achieve this goal is through the minimization of the distances between the cluster center and the patterns that belong to the cluster. Using this principle, the hard *K*-means algorithm minimizes the following objective function [8]:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in F_k} d(\boldsymbol{m}_k, \boldsymbol{x}_i) \qquad (2)$$

where $d(\boldsymbol{m}_k, \boldsymbol{x}_i)$ is a distance measure between the center $\boldsymbol{m}_k$ of the cluster F_k and the pattern $\boldsymbol{x}_i \in X$. Eqn. (2) can be rewritten as

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_k(\boldsymbol{x}_i) d(\boldsymbol{m}_k, \boldsymbol{x}_i) \qquad (3)$$

where $\mu_k(\boldsymbol{x}_i) \in \{0,1\}$ is the characteristic function, i.e., $\mu_k(\boldsymbol{x}_i) = 0$ if $\boldsymbol{x}_i \notin F_k$, else $\mu_k(\boldsymbol{x}_i) = 1$. When the clusters are overlapping, each pattern may belong to more that one cluster, i.e., $\mu_k(\boldsymbol{x}_i) \in [0,1]$. Hence, $\mu_k(\boldsymbol{x}_i)$ should be interpreted as a membership function rather than the characteristic function. Therefore, the objective function (3) can be modified to the following:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_k^q(\boldsymbol{x}_i) d(\boldsymbol{m}_k, \boldsymbol{x}_i) \qquad (4)$$

where $\mu_k(\boldsymbol{x}_i) \in [0,1]$ is now a fuzzy membership function, and q is a constant known as the *index of fuzziness* that controls the amount of fuzziness. Since the minimization of the objective function (4) may lead to a trivial solution, the following two constraints are satisfied while minimizing the objective function:

$$\textstyle\sum_{i=1}^{n} \mu_k(\boldsymbol{x}_i) > 0 \quad \forall k \in \{1, 2, ..., K\} \qquad (5)$$

$$\textstyle\sum_{k=1}^{K} \mu_k(\boldsymbol{x}_i) = 1 \quad \forall i \in \{1, 2, ..., n\} \qquad (6)$$

The first constraint guarantees that there is no empty cluster, and the second constraint imposes the condition that each pattern needs to share its membership with all the clusters such that the sum of memberships is equal to one. Differentiating the objective function (4) with the constraints (5) and (6), we obtain

$$\mu_k(\boldsymbol{x}_i) = \frac{1}{\sum_{h=1}^{K} \left(\frac{d(\boldsymbol{m}_k, \boldsymbol{x}_i)}{d(\boldsymbol{m}_h, \boldsymbol{x}_i)} \right)^{2/(q-1)}} \quad \forall i \in \{1, ..., n\}, k \in \{1, ..., K\} \qquad (7)$$

$$\boldsymbol{m}_k = \frac{\sum_{i=1}^{n} \mu_k^q(\boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{i=1}^{n} \mu_k^q(\boldsymbol{x}_i)} \quad k = 1,2,...,K \qquad (8)$$

Eqn. (7) and (8) are used in an iterative fashion to update the memberships and the cluster centers. The updating continues until the changes in the membership values of all the patterns become negligible or the required number of iterations is over (Fig. 1).

The worst-case time complexity of the algorithm is as follows: To find the distance between the cluster center and all the patterns, we need $O(nN)$ computations. For all the clusters, the number of computations needed is $O(nNK)$. If the clustering needs *T* iterations, then the worst-case complexity is $O(nNKT)$.

INPUT:

(1) A set of input data X .

(2) The value of the fuzziness index $q \in (1,\infty)$.

(3) Number of clusters K .

(4) A distance measure $d(\boldsymbol{m}_k, \boldsymbol{x}_i) = (\boldsymbol{m}_k - \boldsymbol{x}_i)' A^{-1}(\boldsymbol{m}_k - \boldsymbol{x}_i)$ between $\boldsymbol{m}_k$ and $\boldsymbol{x}_i$, where *A* is a positive definite matrix.

(5) A small, positive constant ε , and an appropriate matrix norm $\|.\|$.

(6) Maximum number of iterations *T*.

(7) An $n \times K$ matrix U , where the element of the *i*th row and the *k*th column indicates $\mu_k(\boldsymbol{x}_i)$.

ALGORITHM:

```
Assign $t = 0$.
Randomly initiate the fuzzy K-partition of $U^t$ .
DO
    Set $t = t + 1$.
    FOR $k = 1,2,...,K$
        Calculate the cluster center $\boldsymbol{m}_k$ using
            $\boldsymbol{m}_k = \frac{\sum_{i=1}^{n} \mu_k^q(\boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{i=1}^{n} \mu_k^q(\boldsymbol{x}_i)}$
    ENDFOR
    Update $U^{t+1}$ by calculating $U^t$ as follows:
    Determine the content of the following set:
    $I_k = \{k \mid 1 \le k \le K;\ d(\boldsymbol{m}_k, \boldsymbol{x}_i) = 0\}$
    IF $I_k = \varnothing$ ,
        $\mu_k(\boldsymbol{x}_i) = \frac{1}{\sum_{h=1}^{K} \left(\frac{d(\boldsymbol{m}_k, \boldsymbol{x}_i)}{d(\boldsymbol{m}_h, \boldsymbol{x}_i)}\right)^{2/(q-1)}}$
    ELSE $\mu_k(\boldsymbol{x}_j) = 0 \ \ \forall k \in \{1,2,...,K\} - I_k$
        and $\sum_{k \in I_k} \mu_k(\boldsymbol{y}_j) = 1$
    ENDIF
ENDDO UNTIL $\|U_t - U_{t+1}\| > \varepsilon$ OR $t < T$
```

OUTPUT:

(1) $\mu_k(\boldsymbol{x}_i) \ \ \forall i, k$, i.e., the belongingness of the patterns in the clusters.

(2) $u = \underset{k}{\operatorname{argmax}}\, \mu_k(\boldsymbol{x}_i)$. u denotes the cluster in which $\boldsymbol{x}_i$ belongs to *when* the membership is considered crisp.

Fig. 1: *Fuzzy K-means algorithm.*

3. Proposed Method

Algorithm: Let all the missing values in the data set *X* occur in the *d*th attribute. We shall relax this constraint later. Let us call the set of all the patterns with missing values *Z*, and the set of all complete patterns *Y* (i.e., $X = Y \cup Z$). Each pattern $\boldsymbol{z}_j = [z_{j1}, z_{j2}, ..., z_{j(d-1)}, ?, z_{j(d+1)} ..., z_{jN}]' \in Z$ can be made complete by substituting z_{jd} by $\frac{1}{|Y|}\sum_{\boldsymbol{y} \in Y} y_d$, where $|Y|$ indicates the cardinality of the set *Y* and $[u]'$ indicates the transpose of [*u*]. Subsequently, the standard FKM can be applied to the data set since there is no missing value in the data set.

However, we can modify the clustering algorithm so that the substitution operation is more context dependent. In the clustering, we need the substitution operation while finding the distance between a cluster center (say *k*th) and an incomplete pattern. We can fill the pattern at that point of time, and thus, we fill the pattern differently and incrementally for each cluster center. Therefore, instead of filling z_{jd} by $\frac{1}{|Y|}\sum_{\boldsymbol{y} \in Y} y_d$, we fill $(z_{jd} - m_{kd})^2$ by the mean of $\{(y_{id} - m_{kd})^2 \mid i = 1,2,...,|Y|\}$, i.e., $\left[\frac{1}{|Y|}(\sum_{i=1}^{|Y|} y_{id} - m_{kd})\right]^2$. Here the assumption is that the members of $\{(y_{id} - m_{kd})^2 \mid i = 1,2,...,|Z|\}$ are *i.i.d.* (independent and identically distributed), and hence, from the law of large numbers, they form a Gaussian distribution. In the above procedure, we treat each complete pattern $\boldsymbol{y}_i$ equally. However, the complete patterns that are close to m_{kd} should influence the update of the cluster center more. In other words, we can use the concept of weighted mean instead of a simple mean. Hence, we choose the weights as the membership values. Thus, $(z_{jd} - m_{kd})^2$ is substituted by $\left[\frac{\sum_{i=1}^{|Y|}\left(\mu_k(\boldsymbol{y}_i)(y_{id}-m_{kd})^2\right)}{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)}\right]$.

The substituted value $\left[\frac{\sum_{i=1}^{|Y|}\left(\mu_k(\boldsymbol{y}_i)(y_{id}-m_{kd})^2\right)}{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)}\right]$ becomes same for all patterns with missing values although some of the patterns with missing values are very close to the cluster center m_{kd} and some are far away from m_{kd}. If we assume that the weighted distance $(z_{jd} - m_{kd})^2 / \sigma_d^2$ linearly depends on the weighted distance between z_{id}, $\forall i \neq j$, and m_{kd}, then we can estimate $(z_{jd} - m_{kd})^2 / \sigma_d^2$ using the following linear regression or weighted mean:

$$\begin{aligned}(z_{jd} - m_{kd})^2 / \sigma_{kd}^2 &= w_1 (z_{j1} - m_{j1})^2 / \sigma_{k1}^2 + \ldots \\ &+ w_{(d-1)} (z_{j(d-1)} - m_{j(d-1)})^2 / \sigma_{k(d-1)}^2 \\ &+ w_d \frac{\sum_{i=1}^{|Y|}\left(\mu_k(\boldsymbol{y}_i)(y_{id}-m_{kd})^2\right)}{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)} / \sigma_{kd}^2 \qquad (9) \\ &+ w_{(d+1)} (z_{j(d+1)} - m_{j(d+1)})^2 / \sigma_{k(d+1)}^2 \\ &+ \ldots + w_N (z_{jN} - m_{jN})^2 / \sigma_{kN}^2\end{aligned}$$

where w_h indicates the importance of the hth attribute, $\boldsymbol{m}_k = \frac{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)\boldsymbol{y}_i}{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)}$ and $\sigma_{kh}^2 = \frac{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)(y_{ih} - m_{kh})^2}{\sum_{i=1}^{|Y|}\mu_k(\boldsymbol{y}_i)}$. The importance w_h can be determined by using some *a priori* knowledge or by using some feature extraction algorithms (when the data are labeled). In this paper, we are not assuming that we know the importance of the attributes, and hence we are distributing the importance equally among all the attributes by making $w_h = 1/N \ \ \forall h \in \{1, 2, \ldots, N\}$.

Till now we have shown all the derivations when the values are missing only in the dth attribute. Similar procedure can be adopted when we have patterns with missing values in more than one attribute. Thus, the modified FKM needs some extra steps to consider the incomplete patterns.

Particular case: The mean-value imputation, in which the missing value z_{jd} of the pattern $\boldsymbol{z}_j$ is replaced by $\frac{1}{|Y|}\sum_{i=1}^{|Y|} y_{jd}$, can be derived from the proposed method when (a) the cluster centers are assumed to be at the origin, (b) all the patterns receive equal importance, and (c) $w_h = 0,\ \forall h \neq d,\ w_d = 1,$ and (d) the repairing is done only in the first iteration. Moreover, if $w_d = 0$ and all $\sigma_h^2\ \ \forall h \in \{1, 2, \ldots, N\}$ are equal, then the proposed algorithm reduces to that of [8].

Convergence: When the missing value occurs only in the dth attribute, we partition the data set into the two sets Y and Z. If we use the proposed algorithm for this type of data set, we actually minimize the following objective function:

$$\begin{aligned}J = &\sum_{i=1}^{|Y|}\sum_{k=1}^{K}\mu_k^q(\boldsymbol{y}_i)[d(\boldsymbol{m}_i, \boldsymbol{y}_i)]^2 \\ &+ \sum_{i=1}^{|Z|}\sum_{k=1}^{K}\mu_k^q(\boldsymbol{z}_i)[d(\boldsymbol{m}_i, \boldsymbol{z}_i)]^2\end{aligned} \qquad (10)$$

It is straightforward to show that the objective function (10) under the constraints (5) and (6) is monotonically decreasing, and hence, the iterative minimization guarantees the convergence. The same result holds if the values are missing in more than one attribute.

Time complexity: Let us first look at the time complexity when the values are missing only in the dth attribute. For finding the mean and variance of all complete patterns, we need $O(|Y|)$ computations in each iteration. For each iteration and cluster center, we require $O(|Z|N)$ computations to do the regression. Since $n = |X| = |Y \cup Z|$, the time-complexity for all the cluster centers and iterations is bounded by $O(nNTK)$. When the missing values occur in more than one attribute, then the worst-case time complexity becomes $O(nN^2TK)$. Since in practical cases $N << n$, the repair work does not significantly change the order of the time complexity of the original FKM algorithm.

Quality of clustering: The quality of the clustering can be measured in two ways: directly or indirectly. In the direct method, we can apply some cluster validity measures to check whether the quality of the clustering is improving. In the indirect method, we cluster the data using the proposed method, and then the clusters are utilized to build the classifiers. The classifier performance is used as an indirect way to quantify the quality of the clustering. Note that this is possible only when the data are labeled.

4. Results and Discussion

We have conducted the experiments on the Wisconsin-Madison breast Cancer data from UCI machine learning repository [2]. We have compared the result of the proposed algorithm with that of mean substitution, hot deck, regression, EM and C4.5 algorithms. The presence of a breast mass may indicate (but not always) malignant cancer. The University of Wisconsin Hospital has collected 699 samples using the fine needle aspiration test. Each sample consists of the following ten attributes: (1) Patient's i.d., (2) clump thickness, (3) uniformity of cell size, (4) uniformity of cell shape, (5) marginal adhesion, (6) single epithelial cell size, (7) bare nuclei, (8) bland chromatin, (9) normal nucleoli and (10) mitosis. Except the patient's i.d., all other measurements are assigned to an integer value between 1 and 10, with 1 being closest to the benign and 10 the most anaplastic. Each sample is either benign or malignant.

The data set contains 16 samples each with one missing attribute. Since the number of missing values is small, we introduced more missing values with probability 0.25 to all attributes of each pattern. Using the *t*-test, we first ensured that the data are missing at random. We find the quality of the clustering through indirect way, i.e., through classification performance. We partition the data set into training and test sets. The training set consists of some patterns with missing values, but the test set contains only complete patterns. Using the proposed technique, the training set is grouped into *K* clusters, and each cluster is fuzzily labeled. Next, each pattern of the test set is classified based on which fuzzy clusters it falls in. Similar scheme is also used with four other imputation techniques, and the classification performances of these techniques are shown in Table 1. The proposed method performs better than the other methods.

The advantages of the proposed method are: (a) the substitution of a particular missing value is carried out differently for different cluster centers, (b) the substitution is carried out incrementally so that better clusters are formed. The limitations of the proposed method are appearing from the assumptions that it requires: (a) the members of $\{(z_{jd}-m_{kd})^2 \,|\, j=1,2,...,|Z|\}$ to be *i.i.d.*, and (b) the attribute with missing values linearly depends on the other attributes. In future, we would attempt to relax these assumptions. In addition to medical problems, we intend to apply the proposed technique to cluster the microarray genomic data, where missing values are encountered quite often due to the limitations of the experiments.

Table 1: Comparative results of the proposed method with respect to other methods.

Techniques	Classification rates
FKM with hot deck	92.67%
FKM with mean substitution	93.18%
FKM with regression	95.67%
FKM with EM algorithm	96.34%
FKM with the proposed method	98.43%
C4.5 after pruning	94.31%

Acknowledgments: A strategic research grant RP960351 from NSTB and the Ministry of Education, Singapore, has supported this work.

References

[1] Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.

[2] Blake C. L. and C. J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn, 1998.

[3] Ghahramani, Z. and M. I. Jordan. *Learning from incomplete data*. Technical report, AI memo no. 1509, MIT, 1994.

[4] Heitjan D. F. *Annotation: What can be done about missing data? Approaches to imputation.* American Journal of Public Health, vol. 87, no. 4, pp. 548-550, 1997.

[5] Heitjan, D. F. and R. Thomas. *Missing data, types of.* In: Encyclopedia of Statistical Sciences Update, vol. 2, pp. 408-411, Wiley, New York, 1998.

[6] Little R. J. A. and Rubin D. B. *Statistical Analysis with Missing Data,* Wiley, New York, 1987.

[7] Schneider T. *Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values.* Journal of Climate, vol. 14, no. 5, pp. 853-871, 2001.

[8] Timm H. and R. Kruse. *Fuzzy cluster analysis with missing values*. In: Proceedings of 17th International Conference of the North American Fuzzy Information Processing Society (NAFIPS98), pp. 242-246, Pensacola, FL, 1998.