Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method*

Dan Li[1], Jitender Deogun[1], William Spaulding[2], and Bill Shuart[2]

[1] Department of Computer Science & Engineering
University of Nebraska-Lincoln, Lincoln NE 68588-0115
[2] Department of Psychology
University of Nebraska-Lincoln, Lincoln NE 68588-0308

Abstract. In this paper, we present a missing data imputation method based on one of the most popular techniques in Knowledge Discovery in Databases (KDD), i.e. clustering technique. We combine the clustering method with soft computing, which tends to be more tolerant of imprecision and uncertainty, and apply a fuzzy clustering algorithm to deal with incomplete data. Our experiments show that the fuzzy imputation algorithm presents better performance than the basic clustering algorithm.

1 Introduction

The problem of missing (or incomplete) data is relatively common in many fields of research, and it may have different causes such as equipment malfunctions, unavailability of equipment, refusal of respondents to answer certain questions, etc. These types of missing data are unintended and uncontrolled by the researchers, but the overall result is that the observed data cannot be analyzed because of the incompleteness of the data sets. A number of researchers over last several decades have investigated techniques for dealing with missing data [1–4]. Methods for handling missing data can be divided into three categories. The first is *ignoring and discarding data*, and *listwise deletion* and *pairwise deletion* are two widely used methods in this category [2]. The second group is *parameter estimation*, which uses variants of the *Expectation-Maximization* algorithms to estimate parameters in the presence of missing data [1]. The third category is *imputation*, which denotes the process of filling in the missing values in a data set by some plausible values based on information available in the data set [4].

Among all imputation approaches, there are many options varying from simple method such as mean imputation, to some more robust and complicated methods based on the analysis of the relationships among attributes. Principal imputation methods in practice include (a) Mean imputation; (b) Regression imputation; (c) Hot deck imputation; and (d) Multiple imputation [3]. Clustering

* This work was supported, in part, by a grant from NSF (EIA-0091530), a cooperative agreement with USADA FCIC/RMA (2IE08310228), and an NSF EPSCOR Grant (EPS-0091900).

S. Tsumoto et al. (Eds.): RSCTC 2004, LNAI 3066, pp. 573–579, 2004.

algorithms have been widely used in hot deck imputation. One of the most well known clustering algorithms is the K-means method, which takes the number of desirable clusters, K, as input parameter, and outputs a partitioning of K clusters on a set of objects. The conventional clustering algorithms are normally crisp. However, it is sometimes not the case in reality, i.e., an object could be assigned to more than one clusters. Therefore, a fuzzy membership function can be applied to the K-means clustering, which models the degree of an object belonging to a cluster. This brings the basic idea of soft computing, which is tolerant of imprecision, uncertainty and partial truth [5]. In this paper, we present a hot deck missing data imputation method based on soft computing.

2 Missing Data Imputation with K-means Clustering

A fundamental problem in missing data imputation is to fill in missing information about an object based on the knowledge of other information about the object. As one of the most popular techniques in data mining, clustering method facilitates the process of solving this problem. Given a set of objects, the overall objective of clustering is to divide the data set into groups based on similarity of objects, and to minimize the intra-cluster dissimilarity. In K-means clustering, the intra-cluster dissimilarity is measured by the summation of distances between the objects and the centroid of the cluster they are assigned to. A cluster centroid represents the mean value of the objects in the cluster.

Given a set of N objects $X = \{x_1, x_2, ..., x_N\}$ where each object has S attributes, we use x_{ij} ($1 \leq i \leq N$ and $1 \leq j \leq S$) to denote the value of attribute j in object x_i. Object x_i is called a *complete* object, if $\{x_{ij} \neq \phi \mid \forall \ 1 \leq j \leq S\}$, and an *incomplete* object, if $\{x_{ij} = \phi \mid \exists \ 1 \leq j \leq S\}$, and we say object x_i has a missing value on attribute j. For any incomplete object x_i, we use $R = \{j \mid x_{ij} \neq \phi, 1 \leq j \leq S\}$ to denote the set of attributes whose values are available, and these attributes are called *reference* attributes. Our objective is to obtain the values of non-reference attributes for the incomplete objects. By K-means clustering method, we divide data set X into K clusters, and each cluster is represented by the centroid of the set of objects in the cluster. Let $V = \{v_1, v_2, ..., v_K\}$ be the set of K clusters, where v_k ($1 \leq k \leq K$) represents the centroid of cluster k. Note that v_k is also a vector in an S-dimensional space. We use $d(v_k, x_i)$ to denote the distance between centroid v_k and object x_i.

The algorithm for missing data imputation with K-means clustering method can be divided into three processes. First, randomly select K complete data objects as K centroids. Second, iteratively modify the partition to reduce the sum of the distances for each object from the centroid of the cluster to which the object belongs. The process terminates once the summation of distances is less than a user-specified threshold ε. The last process is to fill in all the non-reference attributes for each incomplete object based on the cluster information. Data objects that belong to the same cluster are taken as nearest neighbors of each other, and we apply a nearest neighbor algorithm to replace missing data.

We use generalized L_P norm distance [6] to measure the distance between a centroid and a data object in the cluster, as shown in Equation (1):

$$d(v_k, x_i) = \left(\sum_{j=1}^{S} |x_{i,j} - v_{k,j}|^p \right)^{1/p}. \quad (1)$$

The Euclidean distance is actually L_2 distance and the Manhattan distance is L_1 distance. Another choice is the Cosine based distance which is calculated from Cosine Similarity, as shown in Equation (2):

$$Sim(v_k, x_i) = \frac{\sum_{j=1}^{S} x_{i,j} * v_{k,j}}{\sqrt{\sum_{j=1}^{S} {x_{i,j}}^2 \sum_{j=1}^{S} {v_{k,j}}^2}}, \quad \text{and} \quad d(v_k, x_i) = e^{-Sim(v_k, x_i)}. \quad (2)$$

3 Missing Data Imputation with Fuzzy K-means Clustering

Now we want to extend the original K-means clustering method to a fuzzy version to impute missing data. The reason for applying fuzzy approach is that fuzzy clustering provides a better description tool when the clusters are not well-separated, as is the case in missing data imputation. Moreover, the original K-means clustering may be trapped in a local minimum status if the initial points are not selected properly. However, continuous membership values in fuzzy clustering make the resulting algorithms less susceptible to get stuck in local minimum situation.

In fuzzy clustering, each data object x_i has a membership function which describes the degree that this data object belongs to certain cluster v_k. The membership function is defined in Equation (3):

$$U(v_k, x_i) = \frac{d(v_k, x_i)^{-2/(m-1)}}{\sum_{j=1}^{K} d(v_j, x_i)^{-2/(m-1)}}, \quad (3)$$

where $m > 1$ is the fuzzifier, and $\sum_{j=1}^{K} U(v_j, x_i) = 1$ for any data object x_i $(1 \leq i \leq N)$ [7]. Now we can not simply compute the cluster centroids by the mean values. Instead, we need to consider the membership degree of each data object. Equation (4) provides the formula for cluster centroid computation:

$$v_k = \frac{\sum_{i=1}^{N} U(v_k, x_i) * x_i}{\sum_{i=1}^{N} U(v_k, x_i)}. \quad (4)$$

Since there are unavailable data in incomplete objects, we use only reference attributes to compute the cluster centroids.

The algorithm for missing data imputation with fuzzy K-means clustering method also has three processes. Note that in the initialization process, we pick K centroids which are evenly distributed to avoid local minimum situation. In the second process, we iteratively update membership functions and centroids until the overall distance meets the user-specified distance threshold ε. In this process, we cannot assign the data object to a concrete cluster represented by a cluster centroid (as did in the basic K-mean clustering algorithm), because each data object belongs to all K clusters with different membership degrees. Finally, we impute non-reference attributes for each incomplete object. We replace non-reference attributes for each incomplete data object x_i based on the information about membership degrees and the values of cluster centroids, as shown in Equation (5):

$$x_{i,j} = \sum_{k=1}^{K} U(x_i, v_k) * v_{k,j}, \text{ for any non-refence attribute } j \notin R. \tag{5}$$

4 Experiments and Analysis

We test our algorithms on two types of data sets. One is weather related databases for drought risk management. The other type of data is the Integrated Psychological Therapy (IPT) outcome databases for psychotherapy study. A common property in these two types of data sets is that missing data are present either due to the malfunction (or unavailability) of equipment or caused by the refusal of respondents. We use the Root Mean Squared Error (RMSE) to evaluate the overall performance of the imputation algorithms. For each experiment with user-specified parameters, we randomly remove amount of data from test set, and use them as missing data. We run each experiment ten times and the experimental results are based on the average values of testing. Since each data attribute has different domain, to fairly test our algorithms, we first normalize the data set so that all the data values are between 0 and 100.

4.1 Mean Substitution vs. K-means

We first compare the non-fuzzy K-means imputation algorithm with mean substitution method, as shown in Figure 1. For K-means algorithm, we select Manhattan distance metric to compute the distance between any two data objects, and the numbers of clusters are 4 (left) and 7 (right), respectively. Each experiment is conducted ten times. From Figure 1, it is obvious that imputation with K-means clustering method outperforms widely used mean substitution method. This indicates that it is reasonable to fill in missing (non-reference) attributes based on the information from reference attributes. Given two or more data objects, if they are similar (close) with regard to reference attributes, other non-reference attributes should be similar (close) too. This is the essential assumption based on which our K-means imputation algorithm works.

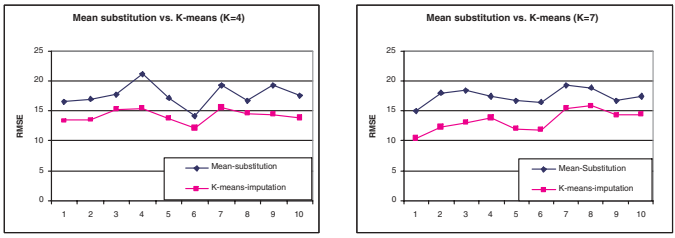

Fig. 1. Mean substitution vs. non-fuzzy K-means imputation.

4.2 K-means vs. Fuzzy K-means

We evaluate and analyze the performance of the basic K-means and the fuzzy K-means imputation algorithms from two aspects. First, we want to see how the percentage of missing data influences the performance of the algorithms. Second, we test on various input parameters (e.g. distance metrics, the value of fuzzifier m, and cluster number K), and conclude with the best values.

Percentage of Missing Data. Table 1 summarizes the results for varying percentages of missing values in the test cases. The experiments are based on two groups of input parameters. First, we select Euclidean distance metric, assume 8 clusters, and set the value of fuzzifier for fuzzy algorithm to 1.5. In the second group of experiments, we use Manhattan distance as the dissimilarity measure, assume 7 clusters, and set the value of fuzzifier to 1.3. We make two observations from Table 1. First, as the percentage of missing values increases, the overall error also increases considering both the basic K-means and the fuzzy K-means imputation algorithms. This is reasonable because we lose more useful information when the amount of missing data increases. The second observation is that the fuzzy K-means algorithm provides better results than the basic K-means method. Especially, when the percentage reaches 20%, the basic K-means algorithm cannot work properly.

Table 1. RMSE for varying percentages of missing values.

	Euclidean Distance, K=8, m=1.5				Manhattan Distance, K=7, m=1.3				
	5%	10%	20%	30%	1%	2%	5%	10%	15%
K-means	14.08	15.93	NA	NA	13.18	13.39	13.94	15.31	15.86
Fuzzy K-means	11.77	12.05	14.41	14.79	10.17	11.51	12.32	13.23	14.57

Effect of Input Parameters. Now, we design experiments to evaluate two missing data imputation algorithms by testing on different input parameters. First, we select three different distance metrics, i.e. Euclidean distance, Manhattan distance, and Cosine-based distance, as shown in Equation (1) and (2).

Table 2 presents the effect of these metrics. We can see that Manhattan distance provides the best result, and the Cosine-based distance is the worst. Again, it can be seen that the fuzzy imputation algorithm outperforms K-means algorithm for all three distance metrics.

Table 2. RMSE for varying distance metrics.

	Manhattan Distance	Euclidean Distance	Cosine-based Distance
K-means	13.37	14.08	17.65
Fuzzy K-means	11.12	11.77	14.99

Next, we want to test on the effect of the value of fuzzifier, which has been used in Equation (3). Since fuzzifier is only a parameter used in fuzzy imputation algorithm, as shown in Table 3, the K-means clustering method does not present much change as the value of m changes. However, for fuzzy algorithm, the change in performance is obvious, and the best value of m is 1.3. When the value of fuzzifier goes to 2, the basic K-means algorithm even outperforms the fuzzy K-means algorithm. This indicates that selecting a proper parameter value is important for system performance.

Table 3. RMSE for varying the values of fuzzifier.

	Euclidean Distance, K=8, 5% missing			
	m=1.1	m=1.3	m=1.5	m=2.0
K-means	13.81	13.71	14.08	13.58
Fuzzy K-means	12.49	10.07	11.77	17.23

5 Conclusion

In this paper, we investigate missing data imputation techniques with the aim of constructing more accurate algorithms. We borrow the idea from fuzzy K-means clustering, and apply it to the problem of missing data imputation. The experimental results demonstrate the strength of this method. We evaluate the performance of the algorithms based on the RMSE error analysis. We discover that the basic K-means algorithm outperforms the mean substitution method, which is a simple and common approach for missing data imputation. Experiments also show that the overall performance of the fuzzy K-means method is better than the basic K-means method, especially when the percentage of missing values is high. We test the performance of our algorithms based on difference input parameters, and find the best value for each parameter.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. of Royal Statistical Society Series **39** (1977) 1–38

2. Gary, K., Honaker, J., Joseph, A., Scheve, K.: Listwise deletion is evil: What to do about missing data in political science (2000) http://GKing.Harvard.edu.
3. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
4. Myrtveit, I., Stensrud, E., Olsson, U.H.: Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering **27** (2001) 999–1013
5. (Zadeh, L.A.) http://www.cs.berkeley.edu/projects/Bisc/bisc.memo.html.
6. Akleman, E., Chen, J.: Generalized distance functions. In: Proceedings of the '99 International Conference on Shape Modeling. (1999) 72–79
7. Krishnapuram, R., Joshi, A., Nasraoui, O., Yi, L.: Low-complexity fuzzy relational clustering algorithms for web mining. IEEE Trans. on Fuzzy Syst. **9** (2001) 595–607