Fuzzy Cluster Analysis of Partially Missing Datasets

Heiko Timm, Christian Döring, and Rudolf Kruse
Dept. of Knowledge Processing and Language Engineering
Otto-von-Guericke-University of Magdeburg
Universitätsplatz 2, D-39106 Magdeburg, Germany
{timm,doering,kruse}@iws.cs.uni-magdeburg.de

Abstract

A common problem in data analysis are missing attribute values in datasets. The easiest way to handle such datasets in fuzzy cluster analysis is to discard data with missing values. Since this complete case approach may result in a loss of valuable information and reduced dataset size, we study how missing values can be handled by modified fuzzy clustering methods. These approaches are based on iterated imputation of missing values, available case estimation of the cluster parameters, and the introduction of a class specific probability for missing values. Benchmark datasets with randomly deleted attribute values are used to demonstrate the capability of the presented approaches. Our experiments show that the modified clustering methods are superior to a complete case analysis.

1 Introduction

In many application areas the quality of the data at hand is a serious problem. Often data are noisy or even partially missing. Therefore it is necessary that data analysis methods can handle these problems well [10, 12]. Out of the variety of data analysis methods we focus on fuzzy cluster analysis in this paper [2, 3, 9]. Cluster analysis is a technique for classifying data, i.e., to divide a given dataset into a set of classes or *clusters*. The goal is to divide the dataset in such a way that two cases from the same cluster are as similar as possible and two cases from different clusters are as dissimilar as possible. Thus one tries to model the human ability to group similar objects or cases into classes and categories. In classical cluster analysis each datum must be assigned to exactly one cluster. Fuzzy cluster analysis relaxes this requirement by allowing gradual memberships, thus offering the opportunity to deal with data that belong to more than one cluster at the same time.

Most fuzzy clustering algorithms are objective function based: They determine an optimal classification by minimizing an objective function. In objective function based clustering usually each cluster is represented by a *cluster prototype*. This prototype consists of a *cluster center* (whose name already indicates its meaning) and may be some additional information about the size and the shape of the cluster. The cluster center is an instantiation of the attributes used to describe the domain, just as the data points in the dataset to divide. However, the cluster center is computed by the clustering algorithm and may or may not appear in the dataset. The size and shape parameters determine the extension of the cluster in different directions of the underlying domain.

The degrees of membership to which a given data point belongs to the different clusters are computed from the distances of the data point to the cluster centers w.r.t. the size and the shape of the cluster as stated by the additional prototype information. The closer a data point lies to the center of a cluster (w.r.t. size and shape), the higher is its degree of membership to this cluster. Hence the problem to divide a dataset $X = \{\vec{x}_1, \ldots, \vec{x}_n\} \subseteq \mathbb{R}^p$ into c clusters can be stated as the task to minimize the distances of the data points to the cluster centers, since, of course, we want to maximize the degrees of membership.

Several fuzzy clustering algorithms can be distinguished depending on the additional size and shape information contained in the cluster prototypes, the way in which the distances are determined, and the restrictions that are placed on the membership degrees. There are also some modifications to deal with noisy data which utilize an additional noise cluster or possibilistic membership degrees [4]. Therefore in this paper we focus on the second problem: how to deal with partially missing data?

A dataset has partially missing data if some attribute values of a datum $\vec{x}_j$ are not observed. For example $\vec{x}_j = (x_{j,1}, x_{j,2}, ?, x_{j,4}, ?)$ has missing values in the third and fifth attribute. Only the first, second and third attribute are observed.

The simplest approach to deal with missing values is to remove data points or attributes having missing values from the dataset. Then the data analysis method is executed on the remaining data only. Obviously, this complete case approach is appropriate if missing values are rare. However, if missing values are frequent, the dataset size may be considerably reduced, so that the analysis method can yield unreliable or distorted results. An alternative approach is to impute missing values during data preprocessing. That is, a missing value is replaced with the average attribute value, modal value, or some other estimate. In this case, the data analysis method can be executed on the whole dataset. Although we avoid the problem of considerably reduced dataset size in this way, we still face problems: If we analyze the data, we cannot distinguish between observed and imputed values. Hence, the quality and reliability of the results depends on the quality of the imputation method used. The third approach to handle missing values is to extend our data analysis method so that the method can deal with incomplete data directly.

In this paper we focus on the third approach. We review several approaches to modify fuzzy clustering methods for treating missing values during analysis. In section 2 we discuss several methods which focus on missing values that are missing completely at random. In section 3 we show how to exploit a class specific probability for missing values in fuzzy clustering.

2 Missing Values Missing Completely at Random

Missing values can occur for several reasons. If the probability that a datum is observed only depends on the observed data Y_{obs} but not on the unobserved data Y_{mis}, missing values are called *missing at random* (MAR) [10, 12]. The whole dataset is $Y = (Y_{obs}, Y_{mis})$. If, in addition, the missing values behave like a random sample and do not depend on the observed data Y_{obs} or on the unobserved data Y_{mis}, they are called *missing completely at random* (MCAR) [10, 12]. In this section we focus on the question how to deal with missing values missing completely at random.

2.1 Iterated Imputation

The simplest approach to include data with missing values into fuzzy cluster analysis is to impute missing values in each iteration of the fuzzy clustering algorithm [13]. Compared to an imputation during data preprocessing this offers the advantage to take the computed classification (membership degrees and cluster prototypes) into account. Missing values can be imputed by several methods. For instance, we can impute the corresponding attribute value of the cluster to which the corresponding data point has the highest membership degree. If we are interested in the variances and covariances of clusters we should also modify the formula slightly to obtain the "true", unbiased values. An alternative approach is to use the weighted mean of all cluster centers. That is, a missing value $x_{j,k}$ is computed as $x_{j,k} = \frac{\sum_{i=1}^{c} u_{i,j}^m c_{i,k}}{\sum_{i=1}^{c} u_{i,j}^m}$. $x_{j,k}$ is the k-th attribute value of datum $\vec{x}_j$ and $c_{i,k}$ is the k-th attribute value of cluster center $\vec{c}_i$. The advantage of this approach is that the imputed value changes more smoothly. However, the computation of the "true" variances and covariances becomes more difficult. Therefore, this imputation method is well suited only for the fuzzy c-means algorithm.

The disadvantage of this approach is that the computation of the cluster centers is based on observed *and* imputed attribute values and that the imputation of missing values is based on these cluster centers. Thus, the imputation of a missing value is influenced by its own imputation in the previous iteration. This problem becomes worse if we impute missing values by the center of the cluster with the highest membership degree. By doing so we underestimate the distance between the datum and the cluster and thus overestimate the membership degrees. In other words: Data with missing values tend to have a higher membership degree and therefore they have a stronger influence on clusters than data without missing values.

2.2 Missing Values as an Optimization Problem

An alternative approach is to see the imputation of missing values as an optimization problem [8]. In this case we try to compute the cluster prototypes, the membership degrees between data and clusters, *and* the missing values by fuzzy cluster analysis. Cluster prototypes, membership degrees and missing values are computed iteratively by minimizing the objective function. By computing the derivative of the objective function w.r.t. the missing values, we obtain $x_{j,k} = \frac{\sum_{i=1}^{c} u_{i,j}^m c_{i,k}}{\sum_{i=1}^{c} u_{i,j}^m}$.

If we compare this approach with the iterated imputation by the weighted mean of the centers in the section before, we see that the approaches are identical. Only the ideas they are based on differ. Therefore the problems

described in the previous section also occur here.

2.3 Available Case Approach

The disadvantage of the approaches discussed in the previous sections is that no distinction is made between observed and imputed data during cluster analysis. This results in higher uncertainty of the results. An alternative is to compute the cluster parameters from the observed data only [13]. In this available case approach we compute the center of a cluster by

$$c_{i,k} = \frac{\sum_{j=1}^{n} u_{i,j}^m i_{j,k} x_{j,k}}{\sum_{j=1}^{n} u_{i,j}^m i_{j,k}}, \tag{1}$$

where $c_{i,k}$ is the k-th attribute of the center $\vec{c}_i$, $x_{j,k}$ is the k-th attribute of the datum $\vec{x}_j$ and $i_{j,k}$ is the k-th attribute of the index vector $\vec{i}_j$. $i_{j,k}$ indicates whether an attribute value is observed, i.e. $i_{j,k} = 1$ if the k-th attribute of $\vec{x}_j$ is observed and $i_{j,k} = 0$ otherwise. The covariance matrix $\mathbf{Cov}_i$ can be computed by

$$\mathbf{Cov}_{i(k,l)} = \frac{\sum_{j=1}^{n} (u_{i,j})^m i_{j,k} i_{j,l} (x_{j,k} - c_{i,k})(x_{j,l} - c_{i,l})^\top}{\sum_{j=1}^{n} u_{i,j}^m i_{j,k} i_{j,l}}. \tag{2}$$

We recommend to compute (2) with the centers computed by (1).

Unfortunately, it is not possible to avoid all imputation. The membership degrees of data to clusters are based on the distances between data and clusters. However, if not all attribute values of a datum are observed, we cannot compute these distances. A simple approach to cope with this problem is the following: The distance of a data point to a cluster center is computed as an aggregate of attribute-specific distances. If we assume that all of these attribute-specific distances behave basically in the same way, we may estimate the attribute-specific distance for an unobserved attribute as the mean of the attribute-specific distances for the observed attributes [5]. This leads to the following formula for the fuzzy c-means algorithm:

$$d(\vec{x}_j, \vec{\beta}_i) = \frac{p}{\sum_{k=1}^{p} i_{j,k}} \sum_{k=1}^{p} i_{j,k} (x_{j,k} - c_{i,k})^2. \tag{3}$$

Another idea is to impute the membership degrees directly without estimating the distances between clusters and data before. In probabilistic fuzzy cluster analysis the membership degrees are computed by

$$u_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d^2(\vec{x}_j, \vec{\beta}_i)}{d^2(\vec{x}_j, \vec{\beta}_k)} \right)^{\frac{1}{m-1}}}.$$

The membership degree is based on the *relation* between the distances to different clusters and not on the absolute value. We can estimate this relation if we compute the distances with respect to the observed attribute values only. For the fuzzy c-means algorithm this idea leads to the same membership degrees as (3), because $\frac{p}{\sum_{k=1}^{p} i_{j,k}}$ is a factor in each distance.

Compared with iterated imputation and the optimization based approach the available case method has the advantage that the influence of the imputation on the fuzzy cluster analysis process is considerably reduced.

3 A Class Specific Probability

In some cases the occurrence of missing values can contain additional information. For example if we know that missing values in attribute a occur in cluster A with a probability of 10% and in cluster B with a probability of 50%, we would assign a datum with a missing values in attribute a rather to cluster B than to cluster A. So the question is: How can we use a class dependent probability for missing values to assign incomplete data points to clusters?

The problem is to combine the class dependent probability with the distance measure of the fuzzy clustering method. For this the probability based distance measure of the fuzzy maximum likelihood estimation algorithm (FMLE) [6] is a good choice. In the FMLE the data points are interpreted as realizations of c p-dimensional normal distributions, where c is the number of clusters. The distance computed in the FMLE is inversely proportional to the posterior probability that a datum was created by the probability distribution of the corresponding cluster.

The case of a class dependent probability for missing values $\vec{mv}_i$ can be integrated into this model [14]. The data points are now seen as realizations of a p-dimensional normal distribution N_i, from which the k-th parameter is missing with a probability mv_{ik} and which is chosen with a probability P_i. However, the posterior probability that a datum $\vec{x}_j$ belongs to class i is difficult to compute if we assume that the decision, which attributes are missing, is made after the creation of a datum. Therefore the model is changed in such a way that first class i is chosen with a probability P_i. Then the decision is made which attributes are missing based on the probabilities $\vec{mv}_i$. And finally the datum is created by the normal distribution N_{ik}. k is an index that indicates which attributes are missing. Because the data belonging to class i shall be created by the same normal distribution N_i, the normal distribution N_{ik} are the marginal distributions of N_i. Because of this fact, the posterior probabilities are the same for both models.

This assumption leads to the following posterior probability (likelihood) that a datum x_j with a missing value in the k-th attribute was created by the normal distribution N_i.

$$\frac{P_i \cdot (1-mv_{i1}) \cdot \ldots \cdot (1-mv_{i(k-1)}) \cdot mv_{ik} \cdot (1-mv_{i(k+1)}) \cdot \ldots \cdot (1-mv_{ip})}{(2\pi)^{p/2}\sqrt{\det(\mathbf{A}_i)}} \exp(-\frac{1}{2}(\vec{x}_j-\vec{c}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j-\vec{c}_i)) \quad (4)$$

Based on the discussion above, attributes in which $\vec{x}_j$ has missing values are neglected for the computation of

$$\frac{\exp(-\frac{1}{2}(\vec{x}_j-\vec{c}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j-\vec{c}_i))}{(2\pi)^{p/2}\sqrt{\det(\mathbf{A}_i)}}. \quad (5)$$

The distance between a datum $\vec{x}_j$ and a cluster $\vec{c}_i$ of the FMLE is inversely proportional to the posterior probability that a datum $\vec{x}_j$ with a missing value in the k-th attribute was created by the normal distribution N_i. Following the model of Gath and Geva the distance is computed by

$$\begin{aligned} d^2(\vec{x}_j,(\vec{c}_i,\mathbf{A}_i,P_i,\vec{mv}_i)) &= \frac{1}{P_i \cdot (1-mv_{i1}) \cdot \ldots \cdot (1-mv_{i(k-1)}) \cdot mv_{ik} \cdot (1-mv_{i(k+1)}) \cdot \ldots \cdot (1-mv_{ip})} \\ &\quad \cdot\sqrt{\det(\mathbf{A}_i)}\exp\left(\frac{1}{2}(\vec{x}_j-\vec{c}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j-\vec{c}_i)\right). \end{aligned}$$

Just as in the definition of the posterior probability, attributes in which $\vec{x}_j$ has missing values are not taken into account for the computation of

$$\sqrt{\det(\mathbf{A}_i)}\exp\left(\frac{1}{2}(\vec{x}_j-\vec{c}_i)^\top \mathbf{A}_i^{-1}(\vec{x}_j-\vec{c}_i)\right) \quad (6)$$

Because the data belonging to the same class are created based on the same normal distribution whether they contain missing values or not, the center of the cluster is computed by neglecting missing values as presented in the available case approach.

4 Experiments

We tested the methods presented in the previous sections with the breast cancer dataset [11] and the wine data set [1]. We classified the breast cancer dataset with modified versions of the fuzzy c-means algorithm into three clusters. The breast cancer dataset consists of 699 data points with 9 attributes. 458 data points belong to the class *benign* and 241 data points to the class *malign*. Only 16 attribute values are missing. Therefore we randomly added missing values by deleting attribute values. Fig. 1 shows the number of classification errors in relation to the probability with which we added new missing values. The results show convincingly that we should only remove data points with missing values from the dataset if missing values are rare. Because of the high number of attributes in the dataset all approaches managed well to deal with missing values. We used the wine dataset to distinguish between these approaches. The 13 attributes of the dataset are the result of a chemical analysis of wine. The dataset is divided into three classes with 59, 71, and 48 data points. We analyzed the dataset with the modified version of the Gustafson–Kessel algorithm [7]. For the analysis we used the attributes 7, 10, and 13 and added missing values with a probability of up to 40%. The results are shown in fig. 2. As in the previous example, merely omitting data with missing values is a bad solution if more than only a few data points have missing values. If we compare the approach based on iterated imputation (this includes the objective function based approach) with the available case approach, we see that the available case approach has a lower number of misclassifications. It may be possible to reduce the number of misclassifications

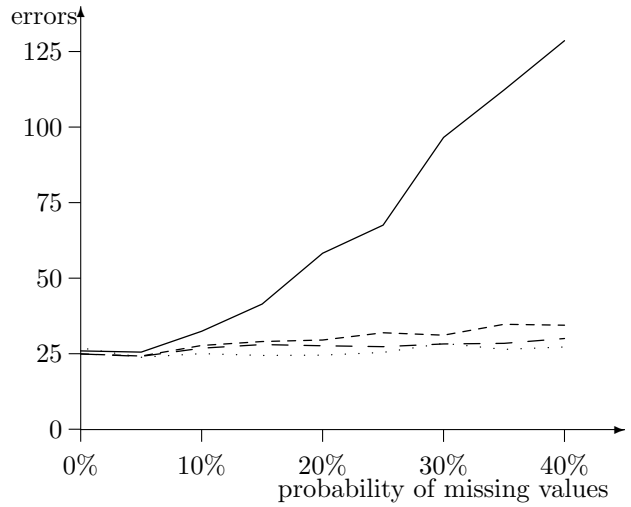

Figure 1: Number of misclassified data points, **fuzzy c-means algorithm**, 3 clusters, Wisconsin breast cancer dataset, **missing values MCAR**. Straight line: complete case analysis, dotted line: imputation with cluster center, dashed line (short dashes) imputation with weighted cluster centers, dashed line (long dashes): available case approach.

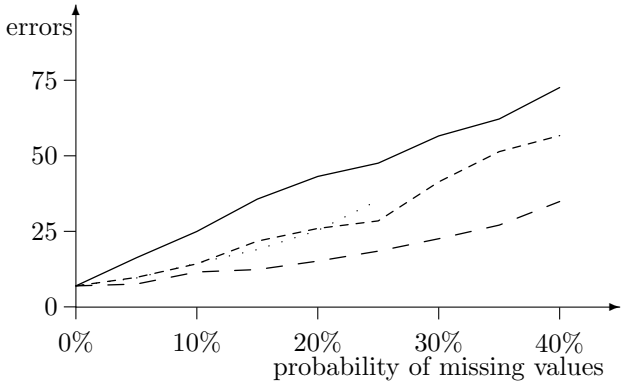

Figure 2: Number of misclassified data points, **Gustafson–Kessel algorithm**, 3 clusters, wine dataset, **missing values MCAR**. Straight line: complete case analysis, dotted line: imputation with cluster center, dashed line (short dashes) imputation with weighted cluster centers, dashed line (long dashes): available case approach.

if we use a more sophisticated method for imputing missing values. However, this would also lead to increased computational costs. Therefore we recommend the available case approach that combines low computational costs with good results.

To test whether it is useful to include a class dependent probability, we used the wine dataset with the attributes 7, 10, and 13 and deleted one attribute in each class with a class dependent probability between 20% and 60%. Furthermore we deleted attributes in the dataset completely at random with a probability of 5%. We classified this dataset with the approach presented in section 3 and the available case modification of the FMLE. Thus the results of both algorithms are comparable. The results in fig. 3 show that the number of misclassifications is lower if we include a class dependent probability into the analysis.

5 Conclusion

In this paper we reviewed several approaches to handle missing values. If missing values occur completely at random (MCAR) we can use iterated imputation or the available case approach. Both approaches are superior to a complete case analysis. If we include data with missing values we can use the inherent redundancy of the dataset to find a classification.

When using the fuzzy c-means algorithm both imputation methods can lead to suitable results. However, if we use more flexible fuzzy clustering methods like the Gustafson-Kessel algorithm, we recommend to use the available case approach. In our experiments this approach showed a high robustness against missing values. The computational costs to extend our fuzzy clustering model to deal with missing values are very small. Thus, it scales up well to larger datasets.

If we recognize that missing values occur with a class specific probability, we can improve the classification if we use this additional information. This can be done with the approach discussed in section 3. Compared with the standard-FMLE the additional computational costs remain low.

In the future we want to study the approaches to deal with missing values in real world problems.

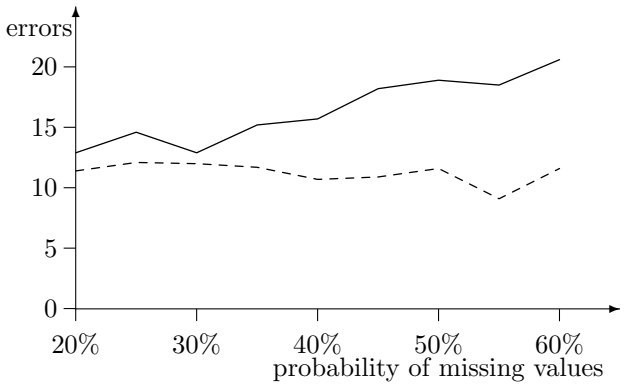

Figure 3: Number of misclassified data points, FMLE, 3 clusters, wine dataset, **missing values with a class specific probability**. Straight line: available case approach, dotted line: exploiting class specific probabilities.

References

[1] Aeberhard, S., Coomans, D., and de Vel, O.: Comparison of Classifiers in High Dimensional Settings. Tech Rep. 92—02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland, 1992.

[2] Bezdek, J.C. and Pal, S.K. (eds.): Fuzzy Models for Pattern Recognition: methods that search for structures in data. IEEE Press, Piscataway, 1992.

[3] Bezdek, J.C., Keller, J., Krishnapuram, R., and Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer, Boston, London, 1999.

[4] Davé, R.N. and Krishnapuram, R.: Robust Clustering Methods: A Unified View, IEEE Transactions on Fuzzy Systems, 5(2), 270—293, 1997.

[5] Dixon, J.K.: Pattern Recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics, 9(6), 617—621, 1979.

[6] Gath, I. and Geva, A.B.: Unsupervised Optimal Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11, 773—781, 1989.

[7] Gustafson, E.E. and Kessel, W.C.: Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE CDC, San Diego, Californien, 761—766, 1979.

[8] Hathaway, R.J. and Bezdek J.C.: Fuzzy c-Means Clustering of Incomplete Data. IEEE Trans. on Systems, Man, and Cybernetics - Part B, 31(5), 735—744, 2001.

[9] Höppner, F., Klawonn, F., Kruse, R. and Runkler, T.: Fuzzy Cluster Analysis, Wiley, Chichester, New York, 1999.

[10] Little, R.J.A. und Rubin, D.A.: Statistical analysis with missing data. John Wiley and Sons, New York, 1987.

[11] Mangasarian, O.L. und Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News, Vol. 23, Nr. 5, 1—18, 1990.

[12] Schafer, J.L.: Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.

[13] Timm, H. and Klawonn, F.: Classification of Data with Missing Values. Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT '98), 1304—1308, Aachen, Deutschland, 1998.

[14] Timm, H. and Klawonn, F.: Different Approaches for Fuzzy Cluster Analysis with Missing Values, Proceedings of 7th European Congress on Intelligent Techniques & Soft Computing, Aachen, Germany, 1999.