Different Approaches for Fuzzy Cluster Analysis with Missing Values

H. Timm
Department of Computer Science
University of Magdeburg
Universitätsplatz 2
D-39106 Magdeburg, Germany
E-Mail: heiko.timm@cs.uni-magdeburg.de

F. Klawonn
Department of Electrical Engineering
and Computer Science
Ostfriesland Univ. of Appl. Sciences
Constantiaplatz 4
D-26723 Emden, Germany
E-Mail: klawonn@et-inf.fho-emden.de

Abstract: Fuzzy cluster analysis is a method for unsupervised data analysis. Because in many data analysis applications the occurrence of missing values is often inevitable, there is a need to deal with this problem. In this paper several approaches are discussed how fuzzy clustering algorithms can deal with missing values. This also includes the presentation of a new approach.

1 Introduction

In real world applications missing values are quite common. They occur for many reasons. For example, if data are gathered with questionnaires, they may be incomplete due to procedural factors. Among these are errors in data entry that created invalid codes, inapplicability of attributes, e.g. age of marriage if a person has never married, or refusal to respond, e.g. if a question concerns income. Other reasons may be that attributes are left blank because for some cases their values are dependent on other attributes or the information is simply missing, e.g. due to faults in sensors.

In the following the expression "missing value" means that for a datum the values of some of its attributes are unknown. If we denote an unknown value by "?", we can write a datum $\mathbf{x_j}$ with the value of the third attribute missing as $\mathbf{x_j} = (x_{j1}, x_{j2}, ?, x_{j4}, \ldots, x_{js})$.

It is obvious that the way in which we deal with missing values depends on their causes. For example we have to handle missing values that are due to the fact that an attribute is inapplicable in a different way than missing values caused by a refusal to respond. Therefore, if a given dataset contains missing values, the first step in data analysis is to examine the reasons for the presence of missing values. Then, in a second step, we look for correlations in the dataset. Depending on the results of these two steps we deal with missing values in different ways. The approaches to deal with missing values can be classified into the following categories [7, 11].

- We delete data with missing values from our dataset or, if missing values occur in some attributes very often, we delete those attributes from our dataset.
- We impute missing values with various methods.
- We choose a method in our data analysis process that tolerates missing values.

The deletion of data or attributes from a given data set is appropriate only, if missing values are rare. If the percentage of missing values is high, the resulting loss of information is not acceptable. Very common is the imputation of missing values by statistical methods. However the quality of this approach depends on the quality of the imputations. Therefore there is also an interest in methods for data analysis that tolerate missing values. In this paper we focus on how fuzzy clustering algorithms can deal with missing values.

Fuzzy cluster analysis is a method for unsupervised data analysis. It is used to classify data, i.e. to divide a given dataset into a set of classes or clusters. The criteria for the quality of the classification are that data belonging to the same cluster are homogeneous and data belonging to different clusters are inhomogeneous. There are several fuzzy clustering algorithms, each of which has its own strengths and weaknesses [8]. These algorithms divide a given dataset into clusters of different size and shape, e.g. spheres, ellipsoids, lines, quadrics or polygons. There are versions which determine the number of clusters automatically as well as robust versions, i.e. fuzzy clustering algorithms which can deal with noise and outliers [2, 4, 10, 12]. Most fuzzy clustering

algorithms are objective function based algorithms. They divide the given data set by minimizing the distances between the clusters and the data assigned to them.

Currently the usual way to deal with missing values in fuzzy cluster analysis is data preprocessing. That is, data with missing values are deleted, the corresponding attributes are removed from the dataset for all data, values are imputed before executing a fuzzy clustering algorithm [7, 9]. This paper presents and discusses some approaches to extend fuzzy clustering algorithms to deal with missing values.

2 Missing values containing no information

Very often missing values are caused by coincidence, error, fault, etc. In these cases a missing values means only that some piece of information is missing. Therefore the question is how to use as much information as possible for classification without reducing the reliability of the results. In the following two different approaches to deal with missing values caused by these reasons are presented and discussed.

A very simple approach to deal with missing values is to impute them during the execution of the fuzzy clustering algorithm [14]. This offers the opportunity to take into account additional knowledge about the computed cluster centers.

A fuzzy clustering algorithm based on this approach classifies a dataset by imputing missing values, computing the distances between the given data and the clusters, and determining the membership degrees and the cluster prototypes. These steps are repeated until the algorithm converges. The algorithm can be initialized e.g. by computing the clusters by neglecting missing values as proposed in the following section.

If there is no further knowledge, the missing values can be imputed using the corresponding attribute values of the cluster to which the datum was assigned with the highest membership degree. However, we have to be very careful if we impute values, because e.g. if we impute a missing value by the corresponding attribute values of the cluster with the highest membership degree, we tend to underestimate the distance. The consequence is that we overestimate the membership degree and therefore the influence of that datum on the cluster. That means that the algorithm extends the influence of data with missing values instead to reduce it as a human would.

This approach is very close to data preprocessing. The only difference is that the computed classification can be used as additional information for the imputation. Therefore this approach has the same disadvantages as data preprocessing methods: if there is a high percentage of missing values, the validity of the computed classification depends on the quality of the imputations. Besides, the use of the computed classification for the imputation of values might lead to the problem of a self fullfilling prophecy.

Another approach is to to adapt the computation formulae in such a way that they can handle data vectors with missing values [14]. We do so by trying to omit the missing values but taking into account the known ones.

Just as a standard fuzzy clustering method an algorithm based on this approach iteratively computes the clusters and the membership degrees of the data to the clusters until the algorithm converges. The differences to the corresponding "normal" fuzzy clustering algorithm are the modified computation formulae. In the following the fuzzy-c-means algorithm [1] will be modified in this way. However the ideas can be used as well for other fuzzy clustering algorithms as e.g. the Gustafson-Kessel algorithm [5] or the algorithm presented by Gath and Geva [6].

The fuzzy-c-means algorithm computes the center of a cluster $\mathbf{c_i}$ by

$$\mathbf{c_i} = \frac{\sum_{j=1}^{n} (u_{ij})^m \mathbf{x_j}}{\sum_{j=1}^{n} (u_{ij})^m}, \tag{1}$$

where u_{ij} is the membership degree of data vector $\mathbf{x_j}$ to cluster $\mathbf{c_i}$, m is the fuzzifier, and n is the total number of data vectors. The center of a cluster is the weighted mean of the data vectors assigned to it and therefore the attributes of the center are the weighted mean of the corresponding attributes of the data vectors assigned to it. That offers the possibility to neglect some attributes of a datum and to take into account the other attributes as shown below. If we extend the membership degree u_{ij} to a membership vector $u_{ij}\mathbf{v_j}$, where $\mathbf{v_j} = (v_{j1}, v_{j2}, \ldots, v_{js})$, $v_{jl} \in \{0, 1\}$, $l \in \{1, 2, \ldots, s\}$, and s is the dimension of the feature space, we can neglect selected parameters of the given datum $\mathbf{x_j}$ by setting v_{jl} to 0.

If we set the corresponding parameter v_{jl} to 0 for all missing values x_{jl} before executing a fuzzy clustering algorithm, and if we compute the centers of the clusters by

$$c_{il} = \frac{\sum_{j=1}^{n} (u_{ij})^m v_{jl} x_{jl}}{\sum_{j=1}^{n} (u_{ij})^m v_{jl}} \text{ for } l \in \{1, \ldots, s\}, \tag{2}$$

we can compute the centers of the clusters regardless whether there are missing values or not.

The computation of the membership is based on the distances of the data to the corresponding clusters. It is obvious that the distance has to be estimated if a datum has a missing value. Therefore this approach is based on estimations similar to the first one in this section. In the following we present two approaches to estimate the distance and thus the membership degree of a datum with missing values to a cluster.

A common approach for this estimation is to neglect those attributes in which a datum has a missing value [13, 14]. That estimation can be interpreted as if the missing attribute does not at all influence the membership degrees of that datum. It means that the membership degrees are estimated directly, because the computed distance is nearly always too small.

Another approach to estimate the distance is based on the assumption that the distance of a datum with missing values to a cluster regarding only the dimension in which that datum has a missing value is similar to the distances of the known attribute to the corresponding attributes of the cluster [14]. That means missing values are imputed in every iteration of the clustering algorithm. However it should be noted that this estimation of missing values is only used to compute the distances. Based on these distances the membership degrees are computed as usual.

If we take a look at these approaches, we see that they have different properties due to the point in the algorithm at which we estimate. The estimation of the distance between a datum $\mathbf{x_j}$ and a cluster $\mathbf{c_i}$ leads to a higher membership degree to the clusters for which $\mathbf{x_j}$ is typical and to a lower membership degrees to clusters for which $\mathbf{x_j}$ is untypical. The reason for this property is that one approach changes the distances to each class and the other does not. Because we estimate the distance for the attributes in which $\mathbf{x_j}$ has a missing value based on the distances of the given attribute values, the relation of the distances changes. For example, if we have a datum with missing values, $\mathbf{x_1} = (2, ?)$, and two cluster centers $\mathbf{c_1} = (1, 3)$ and $\mathbf{c_2} = (5, 5)$, we estimate the distances $d^2(\mathbf{x_1}, \mathbf{c_1}) = 2$ and $d^2(\mathbf{x_1}, \mathbf{c_2}) = 18$. This leads to membership degrees $u_{11} = 0.9$ and $u_{21} = 0.1$, if we estimate the distances by the second approach, and to $u_{11} = 0.75$ and $u_{21} = 0.25$ otherwise. This effect increases if a datum contain more than one missing value.

As the first approach presented in this section the second one is still based on an estimation. But due to the way in which the prototypes are computed, the influnce of the estimations is lower than in that first one. Therefore the obtained classification is more reliable.

Both approaches have in common that the weight of a datum with missing values is the same as that of a datum with no missing values. However both approaches are based on estimations. Therefore the classification of a datum with missing values is not as reliable as that of a datum with no missing values. This can be taken into account by reducing the membership degrees of data with missing values. E.g. a datum with missing values could be assigned to a cluster of untrustworthiness analogous to the approach of Davé to deal with noise [3, 4].

3 Missing values with class dependent probability

The approaches presented in the previous section are based on the assumption that a missing value contains no information. However, sometimes a missing value hints to the class the specific datum might belong to. E.g. in a questionnaire or a medical report some attributes might be left blank because they are inappropriate for a specific class or the same information is contained in other attributes. In these cases the probability for missing values is often class specific. The idea is to use this additional information for fuzzy cluster analysis. We apply it here to the algorithm presented by Gath and Geva [6].

The fuzzy clustering algorithm presented by Gath and Geva is an extension of the Gustafson Kessel algorithm that takes the size and the density of the clusters into account. Gath and Geva interpret the data set as a realization of c p-dimensional normal distributions, where c is the number of clusters. That is a datum $\mathbf{x_j}$ is created with a prior probability P_i by the normal distribution N_i with the expected value $\mathbf{c_i}$ and the covariance matrix $\mathbf{A_i}$. The algorithm computes the classification as a maximum likelihood classifier. I.e. it computes P_i and N_i based on the current membership degrees by maximizing the likelihood that the data assigned to a cluster belong to that cluster.

The case of a class dependent probability for missing values $\mathbf{mv_i}$ can be integrated into this model. The data are now seen as a realization 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 $\mathbf{x_j}$ 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 that first with a probability P_i class i is chosen. Then with the probabilities $\mathbf{mv_i}$ the decision is made, which attributes are missing. 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.

$$\frac{P_i \cdot (1 - mv_{i1}) \cdot \ldots \cdot (1 - mv_{i(k-1)}) \cdot mv_{ik} \cdot (1 - mv_{i(k+1)}) \cdot \ldots \cdot (1 - mv_{ip})}{(2\pi)^{p/2}\sqrt{\det(\mathbf{A_i})}} \exp(-\frac{1}{2}(\mathbf{x_j} - \mathbf{c_i})^\top \mathbf{A^{-1}}(\mathbf{x_j} - \mathbf{c_i})) \quad (3)$$

Based on the discussion above attributes in which $\mathbf{x_j}$ has missing values are neglected for the computation of

$$\frac{\exp(-\frac{1}{2}(\mathbf{x_j} - \mathbf{c_i})^\top \mathbf{A^{-1}}(\mathbf{x_j} - \mathbf{c_i}))}{(2\pi)^{p/2}\sqrt{\det(\mathbf{A_i})}}. \quad (4)$$

The distance function between a datum $\mathbf{x_j}$ and a cluster c_i of the algorithm presented by Gath and Geva is reverse proportional to the posterior probability that a datum $\mathbf{x_j}$ 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 posterior probability (3) leads to:

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

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

$$\sqrt{\det(\mathbf{A})} \exp(\frac{1}{2}(\mathbf{x} - \mathbf{c})^\top \mathbf{A^{-1}}(\mathbf{x} - \mathbf{c})) \quad (5)$$

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, i.e. by:

$$c_{il} = \frac{\sum_{j=1}^{n} (u_{ij})^m v_{jl} x_{jl}}{\sum_{j=1}^{n} (u_{ij})^m v_{jl}} \text{ for } l \in \{1, \ldots, s\}, \quad (6)$$

Analogous to the centers the covariance matrices $\mathbf{A_i}$ are computed by:

$$A_{i_{kl}} = \frac{\sum_{j=1} v_{jl} v_{jk} u_{ij}^m (x_{jk} - c_{ik})(x_{jl} - c_{il})^\top}{\sum_{j=1}^{n} v_{jl} v_{jk} u_{ij}^m} \quad (7)$$

The probabilities P_i and $\mathbf{mv_i}$ are computed by:

$$P_i = \frac{\sum_{j=1}^{n} u_{ij}^m}{\sum_{j=1}^{n} \sum_{l=1}^{c} u_{ij}^m}, \quad (8)$$

$$mv_{ik} = \frac{\sum_{j=1}^{n} u_{ij}(1 - v_{jk})}{\sum_{j=1}^{n} u_{ij}}, \quad (9)$$

where $v_{jk} = 0$ if the k-th attribute of datum x_j is missing and $v_{jk} = 1$ if it is given.

In contrast to the approaches presented in the previous section, this approach is well defined. The computation formulae are only based on the assumption that the probability for a missing value is class dependent. If this assumption is not valid, this approach should not be used.

4 Conclusion

This paper presents some approaches for fuzzy cluster analysis to deal with missing values. If missing values occur for random, there is no well defined solution. The distance between the data and the corresponding clusters always has to be estimated. Therefore the influence of the estimations on the classification should be as

low as possible to obtain a reliable classification. This can be done by neglecting attributes in which a datum has missing values. If some information about the cause of the missing values is available, it can be used. For that this paper presents an approach how to deal with missing values if there is a class dependent probability for their occurance. This new approach has the advantage against the first ones that it is well defined. The classification is not based on estimations. However the reliability of this approach is based on the validity of its underlying assumption. Therfore the reliability of this new approach depends on the reliability of the assumption that there is a class dependent probability for missing values.

References

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

[2] K.K. Chintalapudi, M. Kam. A Noise-Resistant Fuzzy C Means Algorithm for Clustering. Proceedings of FUZZ-IEEE 1998, pp. 1458-1463.

[3] R.N. Davé. Characterization and Detection of Noise in Clustering. Pattern Rcognition Letters, 12(11), pp. 657-664, 1991.

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

[5] E.E. Gustafson, W.C. Kessel. Fuzzy Clustering with a fuzzy Covariance Matrix. Proc. IEEE CDC, San Diego, Calif., pp. 761-766, 1979.

[6] I. Gath and A.B. Geva. Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intelligence 11, 773-781, 1989.

[7] J.E. Hair, R.E. Anderson, R.L. Tatham and W.C. Black. Multivariate Data Analysis with Readings. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[8] F. Höppner, F. Klawonn, R. Kruse and T. Runkler. Fuzzy Cluster Analysis. Wiley, Chichester, 1999.

[9] P.R. Krishnaiah and L.N. Kanal. Handbook of Statistics 2. Classification, Pattern Recognition and Reduction of Dimensionality. North Holland, Amsterdam, 1990.

[10] R. Krishnapuram and J. Keller. A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems, 1, pp. 98-110, 1993.

[11] W.Z. Liu, A.P. White, S.G. Thompson and M.A. Bramer. Techniques for Dealing with Missing Values in Classification. in: Advances in Intelligent Data Analysis. Eds. X. Liu, P. Cohen, M. Berthold. Springer, Berlin, 1997.

[12] N.R. Pal, K. Pal and J.C. Bezdek. A Mixed c-Means Clustering Model. Proceedings of the sixth IEEE International Conference on Fuzzy Systems, pp. 11-21, 1997.

[13] D. Steinhausen and K. Langer. Clusteranalyse - Einführung in Methoden und Verfahren der automatischen Klassifikation, Walter de Gruyter, Berlin, 1977.

[14] H. Timm, F. Klawonn. Classification of Data with missing values. Proceedings of EUFIT '98, pp. 639-644, 1998.