Fuzzy c-Means Clustering of Incomplete Data Using Dimension-Wise Fuzzy Variances of Clusters

Ludmila Himmelspach$^{(\boxtimes)}$ and Stefan Conrad

Institute of Computer Science, Heinrich-Heine-Universität Düsseldorf,
40225 Düsseldorf, Germany
{himmelspach,conrad}@cs.uni-duesseldorf.de

Abstract. Clustering is an important technique for identifying groups of similar data objects within a data set. Since problems during the data collection and data preprocessing steps often lead to missing values in the data sets, there is a need for clustering methods that can deal with such imperfect data. Approaches proposed in the literature for adapting the fuzzy c-means algorithm to incomplete data work well on data sets with equally sized and shaped clusters. In this paper we present an approach for adapting the fuzzy c-means algorithm to incomplete data that uses the dimension-wise fuzzy variances of clusters for imputation of missing values. In experiments on incomplete real and synthetic data sets with differently sized and shaped clusters, we demonstrate the benefit over the basic approach in terms of the assignment of data objects to clusters and the cluster prototype computation.

Keywords: Clustering · Fuzzy c-means (FCM) · Incomplete data · Missing values

1 Introduction

Clustering is one of the important and primarily used techniques for the automatic knowledge extraction from large amounts of data. Its aim is to identify groups, so-called clusters, of similar objects within a data set. Data clustering is used in many fields, including database marketing, image processing, bioinformatics, text mining, and many others. The quality of the data plays an important role in the clustering process and might affect the clustering results. However, problems during the data collection and data preprocessing are often inevitable and might lead to uncertain, erroneous, or missing values in the data sets. Since the completion or correction of data is often expensive or even impossible, there is a need for clustering methods that can deal with such imperfect data.

In the literature, several approaches for adapting the fuzzy c-means (FCM) algorithm to incomplete data have been proposed [1–3]. Like the basic FCM algorithm, these methods assume that the clusters are equally sized and shaped. Even if using the Euclidean distance function as a similarity criterion, in practice,

FCM often reliably identifies differently shaped and sized clusters in complete data sets. Therefore, the FCM algorithm is widely used as a stand-alone clustering method as well as in hybrid hierarchical clustering methods [4]. However, the experiments conducted in [5] showed that FCM adapted to incomplete data produces less accurate clustering results on incomplete data with differently sized and scattered clusters than in the case of clusters with equal sizes and volumes. In [6], we proposed an approach for adapting FCM to incomplete data that takes the cluster scatters into account during the imputation of missing values. Since this method computes the cluster dispersions on the basis of completely available features, it cannot be applied on data sets where missing values occur in all features. Moreover, the experiments showed that this approach produces less accurate cluster prototypes than the basic approach. We assume the reason in the calculation and the use of the cluster dispersions. The cluster dispersions are computed on the basis of completely available features, but they are used for the imputation of missing values in the other features. Therefore, if clusters have different extents in different dimensions, the way of using the cluster dispersions for the imputation of missing values as it is used in our previous approach does more harm than good. In this paper, we present an approach for adapting the fuzzy c-means algorithm to incomplete data that uses the dimension-wise fuzzy variances of clusters for imputation of missing values. Our approach involves the cluster shapes and volumes for missing value imputation in a simple and computationally inexpensive way. In experiments on real and synthetic data sets with differently sized and shaped convex clusters, we demonstrate the capabilities of our new approach and show the benefit over the basic approach in terms of the assignment of data objects to clusters and the cluster prototype computation.

The remainder of the paper is organized as follows. In the next section we give a short overview of the basic fuzzy c-means algorithm and the methods for adapting FCM to incomplete data. In Sect. 3 we describe our idea for missing values imputation using the dimension-wise fuzzy variance of clusters and present the modified FCM algorithm. The evaluation results of our method and the comparison with the basic approach are presented in Sect. 4. In Sect. 5 we close the paper with a short summary and the discussion of future research.

2   Approaches for Fuzzy Clustering of Incomplete Data

2.1   Fuzzy c-Means Algorithm (FCM)

The fuzzy c-means algorithm (FCM) is a well known clustering algorithm that partitions a given data set $X = \{x_1, ..., x_n\}$ in a d-dimensional metric data space into c clusters that are represented by their cluster prototypes $V = \{v_1, ..., v_c\}$. Unlike the k-means algorithm [7], which assigns each data object to exactly one cluster, fuzzy c-means algorithm assigns data items to clusters with membership degrees [8]. The membership degree $u_{ik} \in [0, 1]$ expresses the relative degree to which the data point x_k with $1 \leq k \leq n$ belongs to the cluster C_i, $1 \leq i \leq c$.

The objective function of the fuzzy *c*-means algorithm is defined as follows:

$$J_m(X, U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \cdot d^2(v_i, x_k) \ . \tag{1}$$

The similarity between the data items and the cluster prototypes is expressed by the squared distance function. The parameter m, $m > 1$, is the *fuzzification parameter* which determines the vagueness of the resulting partitioning. The objective function of the fuzzy c-means algorithm is minimized using an *alternating optimization (AO)* scheme [8]. The objective function is alternately optimized over the membership degrees and the cluster prototypes in an iterative process.

   The algorithm begins with the initialization of the cluster prototypes v_i which can be either the first c data items of the data set or c randomly chosen data items or c randomly chosen points in the data space. Alternatively, the membership degrees can be initialized. In the first iteration step the membership degrees of each data item to each cluster are updated according to Formula (2).

$$u_{ik} = \begin{cases} \dfrac{\left(d^2(v_i, x_k)\right)^{\frac{1}{1-m}}}{\sum\limits_{l=1}^{c} \left(d^2(v_l, x_k)\right)^{\frac{1}{1-m}}} & \text{if } I_{x_k} = \emptyset, \\ \lambda, \lambda \in [0,1] \text{ with } \sum_{v_i \in I_{x_k}} u_{ik} = 1 & \text{if } I_{x_k} \neq \emptyset, v_i \in I_{x_k}, \\ 0 & \text{if } I_{x_k} \neq \emptyset, v_i \notin I_{x_k}, \end{cases} \tag{2}$$

where $I_{x_k} = \{v_i \mid d^2(v_i, x_k) = 0\}$. In the second iteration step the new cluster prototypes are calculated based on all data items depending on their membership degrees to the cluster (see Formula (3)).

$$v_i = \frac{\sum\limits_{k=1}^{n} (u_{ik})^m x_k}{\sum\limits_{k=1}^{n} (u_{ik})^m}, \qquad 1 \leq i \leq c \ . \tag{3}$$

The iterative process continues as long as the cluster prototypes change up to a value ϵ. Although the fuzzy c-means algorithm is known as a stable and robust clustering algorithm that does not often get stuck in a local optimum [9,10], it is sensible to evaluate the algorithm for different initializations to achieve the optimal partitioning results.

2.2   Different Approaches for Fuzzy Clustering of Incomplete Data

In the literature, several approaches for adapting fuzzy clustering algorithms to incomplete data have been proposed. Some of them such as the *whole-data strategy FCM (WDSFCM)* and the *partial distance strategy FCM (PDSFCM)* [1,2] carry out the analysis only on the basis of available values. Other methods

like the *optimal completion strategy FCM (OCSFCM)* [1,2], the *nearest proto-type strategy FCM (NPSFCM)* [1,2], and the *distance estimation strategy FCM (DESFCM)* [3] impute missing feature values or distances in an additional iteration step of the fuzzy c-means algorithm. In experiments described in [1,5], the lowest missclassification errors have been obtained by PDSFCM, OCSFCM, and NPSFCM. OCSFCM and NPSFCM tend to strengthen the clustering structure of incomplete data which turned out to be beneficial e.g. for determining the optimal number of clusters using cluster validity indexes (CVIs). In the following we focus on the description of OCSFCM because it provides the basis for our approach.

Optimal Completion Strategy FCM (OCSFCM). The idea of the *optimal completion strategy (OCS)* is to iteratively compute the missing values as the additional variables over which the objective function is minimized [1,2]. The fuzzy c-means algorithm is modified by adding an additional iteration step where the missing values are updated according to Formula (4).

$$x_{kj} = \frac{\sum_{i=1}^{c}(u_{ik})^m v_{ij}}{\sum_{i=1}^{c}(u_{ik})^m}, \qquad 1 \le k \le n \text{ and } 1 \le j \le d \ . \tag{4}$$

In this way, the missing values are imputed by the weighted means of all cluster centers in each iteration step.

The advantage of this approach is that missing values are imputed during the clustering process. However, the drawback of the OCSFCM is that the calculation of the cluster prototypes and the imputation of the missing values influence each other because the algorithm does not distinguish between the available and imputed feature values. In [11] the author proposed to diminish the influence of imputed values to the calculation of cluster prototypes by reducing the membership degrees of incomplete data items depending on the number of missing values. The resulting algorithm loses the property of a probabilistic fuzzy clustering algorithm, though. In our approach we tackle this problem by computing the cluster prototypes only on the basis of available feature values.

3   FCM Clustering of Incomplete Data Using Dimension-Wise Fuzzy Variances of Clusters

In [6], we have proposed an approach for adapting FCM to incomplete data which can be regarded as an extension of OCSFCM that takes the dispersions of clusters into account for the imputation of missing values. The experimental results have shown the benefits of this approach over the basic OCSFCM algorithm. However, since the cluster dispersions are computed on the basis of completely available features but used for the imputation of missing values in the other features, this approach is restricted to data sets with equally shaped clusters, i.e. clusters that have the same extents in different dimensions. Therefore, in our new approach

we impute missing values taking the dimension-wise fuzzy variances of clusters into account. The idea of our approach is imputing missing feature values of incomplete data items depending on both the distances between the cluster prototypes and the incomplete data items and the extents of clusters in the corresponding dimensions.

3.1 A New Membership Degree Using Dimension-Wise Fuzzy Variances of Clusters

For each cluster C_i, $1 \leq i \leq c$, we calculate its fuzzy variance w_{ij} in each dimension j, $1 \leq j \leq d$, as an averaged squared distance of data items to their cluster prototypes according to Formula (5).

$$
w_{ij} = \frac{\sum\limits_{k=1}^{n} u_{ik}^m i_{kj} (x_{kj} - v_{ij})^2}{\sum\limits_{k=1}^{n} u_{ik}^m i_{kj}}, \tag{5}
$$

where

$$
i_{kj} = \begin{cases} 1, & \text{if } x_{kj} \text{ is available} \\ 0 & \text{else} \end{cases} \quad \text{for } 1 \leq j \leq d,\ 1 \leq k \leq n. \tag{6}
$$

We calculate the dimension-wise fuzzy variances of clusters using only available feature items. In this way, we avoid the influence of the imputed feature values on the calculation of cluster variances. That makes sense because missing values are imputed by values close to the corresponding feature values of the cluster prototypes. Therefore, taking both the available and the imputed feature values into account for the calculation of cluster variances would reduce the real variances of clusters. Furthermore, using all available feature items for the calculation of the dimension-wise fuzzy variances of clusters makes our approach applicable on incomplete data sets where missing values occur in all features and data items. Calculating the fuzzy variances of clusters for each feature prevents the influence of the distorted estimation of cluster variances in single dimensions to the whole cluster variances in the case of a large amount of missing values in those features.

We integrate the fuzzy variances of clusters in the new membership degree update formula for the imputation of missing values as follows:

$$
u_{(ik)j}^w = \frac{\left(w_{ij}^{-1}(v_{ij} - x_{kj})^2\right)^{\frac{1}{1-m}}}{\sum\limits_{l=1}^{c} \left(w_{lj}^{-1}(v_{lj} - x_{kj})^2\right)^{\frac{1}{1-m}}} \qquad \forall x_{kj} \text{ with } i_{kj} = 0, \tag{7}
$$

$1 \leq k \leq n$, $1 \leq i \leq c$, and $1 \leq j \leq d$. Note that we also compute the membership degrees for the imputation of missing values for each dimension. That implies that a missing value of an incomplete data item in a particular feature

is updated depending on the distances between the corresponding values of the cluster prototypes and the last imputed value and the extents of clusters in that feature. The larger the variance of a cluster and the smaller the distance between the imputed data item and the cluster center in that particular dimension, the higher the new membership degree is. If all clusters are equally sized and shaped, then the membership degree $u_{(ik)j}^{w}$ depends only on the distances between the data items and the cluster prototypes. As in the basic OCS approach the imputation of missing values in our approach depends on the imputed values in the previous iteration.

3.2   FCM for Incomplete Data Using Dimension-Wise Fuzzy Variances of Clusters (FCMDFVC)

We integrate the new membership degree for imputation of missing values in the basic OCSFCM algorithm and refer the resulting algorithm to as *Fuzzy c-Means Algorithm for Incomplete Data using Dimension-wise Fuzzy Variances of Clusters (FCMDFVC)*. The working principle of FCMDFVC is depicted in Algorithm 1 and is basically the same as of OCSFCM.

The algorithm begins with the initialization of cluster prototypes and missing values. The membership degrees are updated in the first iteration step in the same way as in the basic FCM and OCSFCM. The available and the imputed feature values are not distinguished. Since we want to avoid the influence of the imputed values to the calculation of the cluster prototypes, we compute the cluster prototypes only on the basis of available feature values as in PDSFCM according to Formula (8).

$$v_{ij}' = \frac{\sum_{k=1}^{n}(u_{ik})^m i_{kj} x_{kj}}{\sum_{k=1}^{n}(u_{ik})^m i_{kj}} \qquad \text{for } 1 \le i \le c,\ 1 \le j \le d. \tag{8}$$

If the termination condition is not reached, in the third iteration step, missing values are imputed depending on the distances to the cluster prototypes and the fuzzy variances of clusters according to Formula (9).

$$x_{kj} = \frac{\sum_{i=1}^{c}(u_{(ik)j}^{w})^m v_{ij}'}{\sum_{i=1}^{c}(u_{(ik)j}^{w})^m}, \qquad 1 \le k \le n \text{ and } 1 \le j \le d. \tag{9}$$

Basically, the new membership degree for the imputation of missing values can also be integrated in NPSFCM. In this case, the computation of cluster prototypes using only available feature values is essential because the computation of cluster prototypes and the imputation of missing values influence each other even more than in OCSFCM.

Algorithm 1. FCMDFVC(X, c, m, ϵ)

Require: X is a d-dimensional incomplete data set with n data items, $2 \leq c \leq n$ is
a number of clusters, $m > 1$ is a fuzzification parameter, $\epsilon > 0$ is a termination
accuracy

1: Initialize the set of data centers $v' = \{v'_1, ..., v'_c\}$
2: Initialize all missing values x_{kj} in X with random values in the data space
3: $v = \{\}$
4: **repeat**
5:     $v = v'$
6:     Calculate the membership degrees u_{ik} of each data item x_k to each cluster C_i
       according to Formula (2) // *Step 1*
7:     Calculate the set of new cluster prototypes $v' = \{v'_1, ..., v'_c\}$ according to For-
       mula (8) // *Step 2*
8:     **if** $\|v - v'\| > \epsilon$ **then**
9:         Impute the missing values x_{kj} according to Formula (9) // *Step 3*
10:    **end if**
11: **until** $\|v - v'\| < \epsilon$
12: **return** v'

---

4   Data Experiments

In this section we compare our approach with the basic OCSFCM algorithm
in terms of the assignment of data objects to clusters and the cluster pro-
totypes computation. In order to assess the impact of single modifications of
the basic method, in our experiments, we also tested the OCSFCM algorithm
that computes the cluster prototypes only using the available feature values and
the FCMDFVC approach that computes cluster prototypes using available and
imputed feature values.

4.1   Test Data and Experimental Setup

We tested the four above-mentioned approaches on different real and synthetic
data sets. For the sake of brevity, here we only report the results obtained on
the *wine* data set from the UCI Machine Learning Repository [12] and three
synthetic data sets with different properties. The *wine* data set consists of 178
data items, each with 13 features representing the results of a chemical analysis of
three wine types. Corresponding to the wine types, the data items are distributed
in three classes with 59, 71, and 48 instances. The synthetic data sets are depicted
in Fig. 1. Each of the data sets consist of 2000 data items generated by the
compositions of three and five 3-dimensional Gaussian distributions respectively.
The data items are distributed in three clusters with 400, 700, and 900 instances
in the data set *3D-3* and in five clusters with 200, 350, 450, 700, and 300 instances
in the data sets *3D-5* and *3D-5-h*. All clusters have different magnitudes and

partly overlap. The data set *3D-5-h* was generated from the data set *3D-5* by moving the clusters so that two groups of two and three differently sized clusters build a hierarchical structure in the resulting data set. Using the data set *3D-5-h*, we wanted to find out if the ordering of clusters in the data sets affects the performance of the clustering algorithms adapted to incomplete data.
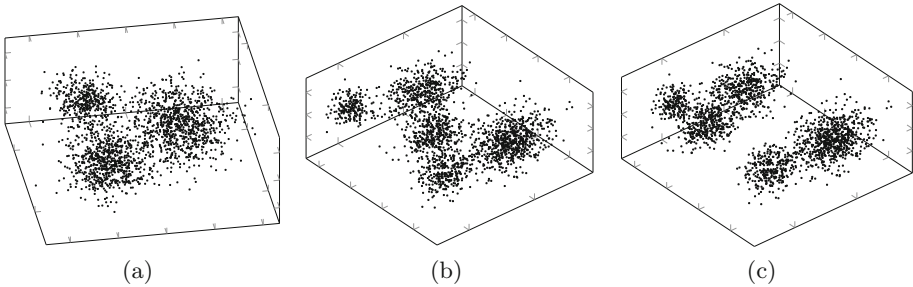

(a)   (b)   (c)

Fig. 1. Test data: (a) 3D-3, (b) 3D-5, (c) 3D-5-h.

In our experiments we first clustered the complete data sets with the basic FCM algorithm and used the resulting clusterings as a baseline for the comparison. Then we generated incomplete data sets by removing values in all features with different probabilities according to the general missing-data pattern [13]. The percentage of missing values was calculated in relation to all feature values in the data sets. In the resulting incomplete data sets, missing values were equally distributed in all features. The general missing-data pattern is common but challenging for clustering methods because missing values occur in many data items. Since we did not adapted the clustering algorithms to incomplete data with conditionally missing values, we deleted the values from the test data sets according to the common missing-data mechanisms MCAR [13]. We clustered the incomplete data sets with the four above-mentioned clustering algorithms. We initialized the cluster prototypes with random values in the data space at the beginning of the algorithms to create the test conditions as real as possible. We used the Frobenius norm distance for the stopping criterion $\|V - V'\| < \epsilon$ defined in Formula (10).

$$\|V - V'\|_F = \sqrt{\sum_{i=1}^{c} \sum_{j=1}^{d} |v_{ij} - v'_{ij}|^2} \qquad \text{for } 1 \leq i \leq c,\ 1 \leq j \leq d. \qquad (10)$$

In all our experiments we set the value ϵ to 0.0001.

4.2   Experimental Results

The performance results of the fuzzy *c*-means clustering algorithms adapted to incomplete data are shown in Figs. 2, 3, 4, and 5. For the sake of uniformity,

we marked the clustering algorithms that only use the available feature items for the computation of cluster prototypes with an asterisk in the legend. That means that our new approach described in Sect. 3 is listed as FCMDFVC* in the legend.

We evaluated the partitioning results produced by the clustering approaches using different crisp and fuzzy indexes. Here, we only report the results for the subset similarity index [14] for comparing the fuzzy partitions produced by the clustering algorithms. Although in previous publications the partitioning results produced by the fuzzy *c*-means algorithms for incomplete data were evaluated using crisp similarity indexes, in our opinion, it makes more sense to use fuzzy indexes to compare the resulting membership degrees. We used the Frobenius norm distance between the terminal prototypes produced by FCM on complete data and the terminal cluster prototypes produced by the four fuzzy clustering approaches on incomplete data. Since there were significant variations in the results from trial to trial, the figures below show the averaged results obtained over 100 trials. For the sake of clarity, we omitted the standard deviations in the diagrams.

Figure 2 presents the performance results for the basic OCSFCM and our approach produced on the incomplete *wine* data set. According to these results the computation of cluster prototypes only using available feature values seems to be the only factor that improved the partitioning results on incomplete data. Our proposal to impute missing values using the dimension-wise fuzzy variance of clusters seems to worsen the partitioning results. Indeed, in almost all experiments this approach produced the least accurate partitioning results and the highest prototype error among all approaches. As we mentioned above, the reason is that the imputation of missing values depends on the imputation in the previous iteration step and influences the computation of the cluster prototypes. Therefore, it is essential to compute the cluster prototypes only on the basis of available feature items in our approach that produced similarly good results as OCSFCM using available feature values for the computation of cluster prototypes.
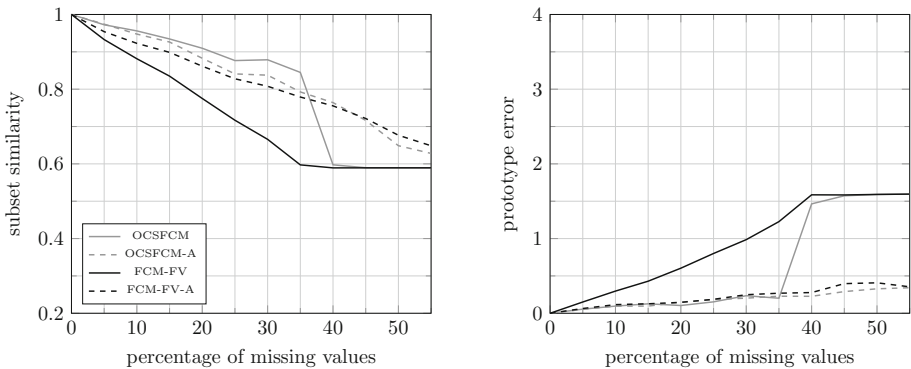

Fig. 2. Averaged results of 100 trials using incomplete *wine* data set.

Since the *wine* data set has 13 attributes, we cannot visualize it and cannot comment on the obtained results. Figures 3 and 4 present the performance results produced by the four approaches on the 3-dimensional data sets with three and five clusters. As the diagrams show, our new approach produced slightly more accurate partitioning results and much more accurate terminal cluster proto-types. While the performance results of OCSFCM and OCSFCM* were similar on the data set *3D-3*, the OCSFCM approach using available feature values for the computation of cluster prototypes produced on average much inaccurate par-titioning results on the data set *3D-5*. This is due to the fact that unlike other methods this approach produced comparably good clustering results only in few of 100 trials and performed poorly on average.

Figure 5 shows the performance results for the four approaches produced on the incomplete *3D-5-h* data set where clusters were moved together building two groups. All clustering approaches produced unstable results on this data set. The basic FCM algorithm produced comparably accurate results on the complete *3D-5-h* data set, we obtained 0.9949 for the crisp subset similarity averaged over
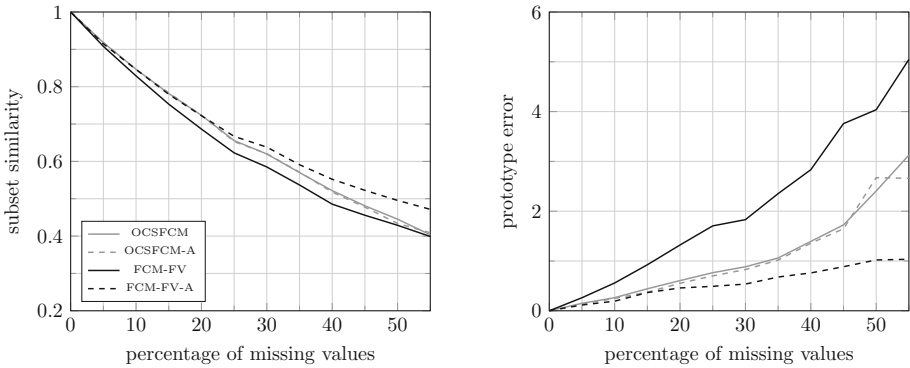

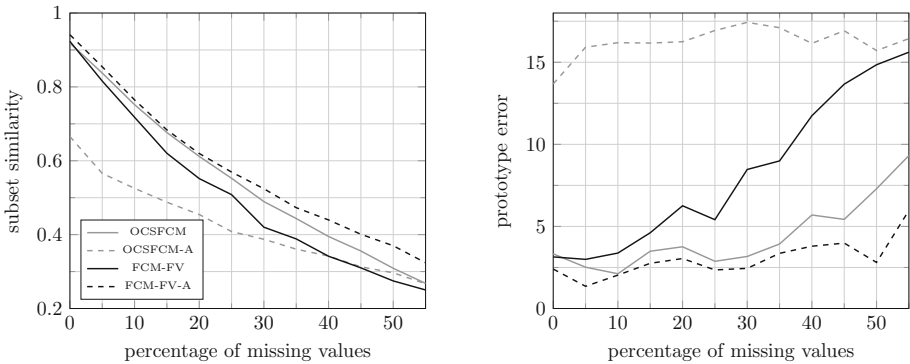**Fig. 3.** Averaged results of 100 trials using incomplete *3D-3* data set.

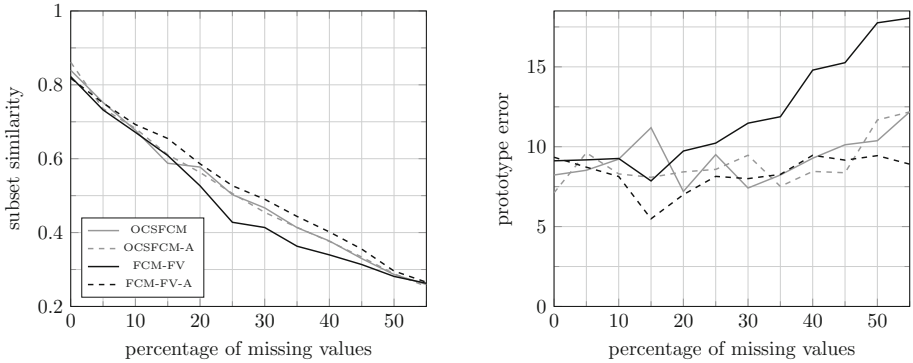**Fig. 4.** Averaged results of 100 trials using incomplete *3D-5* data set.

Fig. 5. Averaged results of 100 trials using incomplete *3D-5-h* data set.

100 trials. That shows that the spatial arrangement of clusters in the data set has a strong influence on the clustering results produced on incomplete data.

5   Conclusions and Future Work

Approaches proposed in the literature for adapting the fuzzy *c*-means (FCM) algorithm to incomplete data assume that the clusters are equally sized and shaped. They produce less accurate clustering results on incomplete data with differently sized and scattered clusters than on the data sets with clusters of equal sizes and volumes. Our previous approach presented in [6] for adapting FCM to incomplete data that takes the cluster scatters into account during the imputation of missing values computes the cluster dispersions on the basis of completely available features, but uses them for the imputation of missing values in the other features. Therefore, if clusters have different extents in different dimensions, it fails to work. In this paper, we presented a new approach for adapting the fuzzy *c*-means algorithm to incomplete data that uses the dimension-wise fuzzy variances of clusters for imputation of missing values. Our approach involves the cluster shapes and volumes for missing value imputation and computes cluster prototypes only using the available feature values. In experiments on real and synthetic data sets with differently sized and shaped clusters, we demonstrated that our new approach produces slightly more accurate partitioning results and much more accurate terminal cluster prototypes than the basic OCSFCM method.

In all our experiments we used incomplete data with missing values MCAR because we did not adapted our approach to incomplete data with a conditional absence of values. In the future we plan to apply the idea of using the class specific probabilities for imputation of missing values as in the approach presented in [15] in order to improve the performance of our approach on data sets with missing values MAR and NMAR. Furthermore, in our experiments we assumed the optimal number of clusters to be known. Determining the optimal number of

clusters on incomplete data is still a challenging problem. Therefore, we want to analyze whether the partitioning results produced by our approach on incomplete data with differently sized and scattered clusters are better than the partitioning produced by the basic OCSFCM for determining the optimal number of clusters using the cluster validity indexes.

References

1. Hathaway, R.J., Bezdek, J.C.: Fuzzy c-means clustering of incomplete data. IEEE Trans. Syst. Man Cybern. Part B **31**(5), 735–744 (2001)
2. Timm, H., Döring, C., Kruse, R.: Fuzzy cluster analysis of partially missing datasets. In: Proceedings of the European Symposium on Intelligent Technologies, Hybid Systems and Their Implementation on Smart Adaptive Systems (EUNITE 2002), pp. 426–431 (2002)
3. Sarkar, M., Leong, T.-Y.: Fuzzy K-means clustering with missing values. In: Proceedings of the American Medical Informatics Association Annual Symposium, pp. 588–592 (2001)
4. van der Laan, M.Y., Pollard, K.S.: A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. J. Stat. Plann. Infer. **117**(2), 275–303 (2003)
5. Himmelspach, L., Conrad, S.: Clustering approaches for data with missing values: comparison and evaluation. In: Proceedings of the Fifth IEEE International Conference on Digital Information Management (ICDIM 2010), pp. 19–28 (2010)
6. Himmelspach, L., Conrad, S.: Fuzzy clustering of incomplete data based on cluster dispersion. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS, vol. 6178, pp. 59–68. Springer, Heidelberg (2010)
7. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281–297 (1967)
8. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
9. Kruse, R., Döring, C., Lesot, M.-J.: Fundamentals of fuzzy clustering. In: Advances in Fuzzy Clustering and Its Applications, pp. 1–30 (2007)
10. Klawonn, F., Kruse, R., Winkler, R.: Fuzzy clustering: more than just fuzzification. Fuzzy Sets Syst. **281**, 272–279 (2015)
11. Timm, H.: Fuzzy-Clusteranalyse: Methoden zur Exploration von Daten mit fehlenden Werten sowie klassifizierten Daten. Ph.D. thesis, Germany (2002)
12. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html
13. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley, New York (2002)
14. Runkler, T.A.: Comparing partitions by subset similarities. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS, vol. 6178, pp. 29–38. Springer, Heidelberg (2010)
15. Timm, H., Döring, C., Kruse, R.: Different approaches to fuzzy clustering of incomplete datasets. Int. J. Approximate Reasoning **35**(3), 239–249 (2004)