Cluster Tendency Assessment for Fuzzy Clustering of Incomplete Data

Ludmila Himmelspach[1] **Daniel Hommers**[1] **Stefan Conrad**[1]

[1]Institute of Computer Science, Heinrich-Heine-Universität Düsseldorf, Germany

Abstract

The quality of results for partitioning clustering algorithms depends on the assumption made on the number of clusters presented in the data set. Applying clustering methods on real data missing values turn out to be an additional challenging problem for clustering algorithms. Fuzzy clustering approaches adapted to incomplete data perform well for a given number of clusters. In this study, we analyse different cluster validity functions in terms of applicability on incomplete data on the one hand. On the other hand we analyse in experiments on several data sets to what extent the clustering results produced by fuzzy clustering methods for incomplete data reflect the distribution structure of data.

Keywords: fuzzy cluster analysis, incomplete data, cluster tendency, cluster validity

1. Introduction

Clustering is an important unsupervised learning technique for automatic knowledge extraction from data. Its task is exploring the distribution of objects in a data set. In general, clustering is defined as a technique for partitioning data set into groups or *clusters* of similar data objects. Various algorithms for data clustering have been developed and are used in many areas, including database marketing, image processing, information retrieval, and others. One of the well-known and widely used clustering methods is the fuzzy c-means algorithm (FCM) [1], which produces a fuzzy partitioning of a data set into c clusters. An important parameter for clustering quality is the number of clusters c. Applying clustering methods on real data sets missing values turn out to be an additional challenging problem for clustering algorithms. For that reason, several clustering approaches for handling incomplete data have been proposed in literature [2, 3, 4, 5]. Data experiments conducted in [2, 5, 6] have shown that for an optimal number of clusters some of these methods are able to assign data items to clusters quite accurately.

However, in real world applications the optimal number of clusters is generally not known a priori. The performance measures for assessing the cluster tendency were developed for complete data. In this study, we analyse different validity functions from literature in terms of applicability on incomplete data on the one hand. On the other hand we analyse in experiments on several data sets to what extent the clustering results produced by methods for incomplete data reflect the distribution structure of data objects and whether the optimal number of clusters can be determined using original and adapted cluster validity functions.

The remainder of the paper is organised as follows. In Section 2, we give an overview of methods for adapting fuzzy c-means algorithm for incomplete data. We describe different cluster validity functions and propose some ideas for their adaption to incomplete data in Section 3. In Section 4, we present the evaluation results and compare clustering methods and cluster validity functions, respectively. We close the paper with a short summary and discussion of future research in Section 5.

2. Fuzzy Clustering of Incomplete Data

2.1. Fuzzy C-Means Algorithm (FCM)

The fuzzy c-means algorithm (FCM) is a well known clustering algorithm, which produces a partitioning of n data points $X = \{x_1, ..., x_n\}$ in d-dimensional metric data space into c clusters. Instead of "hard" dividing the data set into clusters the fuzzy c-means algorithm assigns data points to clusters with membership degrees [1]. The membership degree $u_{ik} \in [0, 1]$ expresses the relative degree to which data point x_k belongs to the cluster C_i and is calculated as follows:

$$u_{ik} = (D_{ik}^{1/(1-m)})/(\sum_{j=1}^{c} D_{jk}^{1/(1-m)}), \qquad (1)$$

where $m > 1$ is the fuzzification parameter and $D_{ik} = ||x_k - \mu_{C_i}||_A^2 = (x_k - \mu_{C_i})^T A(x_k - \mu_{C_i})$ is the squared A-norm distance between data point x_k and the representative (prototype) of the cluster C_i.

Like the most partitioning clustering algorithms FCM determines an optimal partitioning of a data set in an iterative process. The algorithm begins with initialising cluster prototypes μ_{C_i} $(1 \leq i \leq c)$, which are randomly chosen points in the feature space. In the first iteration step, the membership degrees of each data point x_k to each cluster C_i are calculated according to Formula 1. In the second iteration step, the new cluster prototypes are calculated based on all data points depending on their membership degrees to the cluster:

$$\mu_{ij} = (\sum_{k=1}^{n} (u_{ik})^m x_{kj}) / (\sum_{k=1}^{n} (u_{ik})^m), \qquad (2)$$

for $1 \leq i \leq c$ and $1 \leq j \leq d$.

The iterative process continues as long as the cluster prototypes change up to a value ϵ. In this way the objective function given in Equation 3 is minimised in each iteration step.

$$J_m(U, \mu) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m D_{ik}. \qquad (3)$$

2.2. Different Approaches for Fuzzy Clustering of Incomplete Data

To apply the fuzzy c-means algorithm to incomplete data, in literature several approaches for adapting FCM have been proposed [2, 3, 4, 5]. Some of them perform the clustering analysis using only available feature values. Other methods estimate and replace missing values or distances in each iteration of the algorithm. In experiments described in [2, 6], the lowest missclassification errors have been obtained by PDSFCM, OCSFCM and NPSFCM. In experiments on data with missing values *missing completely at random (MCAR)*, also WDSFCM obtained similarly good results as the other three approaches. In this study, we focus our consideration on these four methods. We do not expand on other approaches because either of their weak performance in terms of accuracy or of their adaption to specific data distribution or specific distribution of missing values in the data set.

2.2.1. Whole-Data Strategy (WDS)

A simple method for adapting FCM for handling incomplete data is the *whole-data strategy (WDS)* [2]. First, incomplete data items are removed from the data set and remaining complete data items are clustered via basic fuzzy c-means algorithm. Afterwards, incomplete data items are assigned to the nearest cluster by calculating the partial distances [7] (cf. Formula 4) between incomplete data items and cluster prototypes.

$$D_{part}(x_k, \mu_i) = \frac{d}{\sum_{j=1}^{d} I_{kj}} \sum_{j=1}^{d} (x_{kj} - \mu_{ij})^2 I_{kj}, \qquad (4)$$

where

$$I_{kj} = \begin{cases} 1 & if\ x_{kj}\ is\ available \\ 0 & else \end{cases}$$

for $1 \leq i \leq c$, $1 \leq k \leq n$ and $1 \leq j \leq d$.

In WDSFCM, the partitioning of data set into clusters is carried out only on the basis of complete data items. That implies that WDSFCM cannot be applied to data sets containing no complete data items. Moreover, in view of cluster tendency assessment, the optimal number of clusters is totally determined by the distribution of complete data items.

2.2.2. Partial Distance Strategy (PDS)

Another approach for adapting the fuzzy c-means algorithm to incomplete data is the *partial distance strategy (PDS)* [2]. For the calculation of membership degrees of data items to clusters in the first iteration step, the squared distance function D is replaced by the partial distance function D_{part} (cf. Equation 4). In this way, the distances between incomplete data items and cluster prototypes are approximated by partial distances and the distances between complete data items and cluster prototypes are calculated in the same way as in FCM. In the second iteration step of the algorithm, the cluster prototypes are calculated only on the basis of all available feature values of data items:

$$\mu_{ij} = (\sum_{k=1}^{n} (u_{ik})^m I_{kj} x_{kj}) / (\sum_{k=1}^{n} (u_{ik})^m I_{kj}) \qquad (5)$$

for $1 \leq i \leq c$ and $1 \leq j \leq d$.

The advantage of this approach, in contrast to the whole-data stategy, is that it can be used even if all data items have missing values. This approach benefits from the fact that all cluster prototypes are calculated as complete vectors. Hence, all available feature values of an incomplete data item can be used for calculation of membership degrees of the data item to all clusters. Furthermore, the entire membership matrix can be used for determining the optimal number of partitions.

2.2.3. Optimal Completion Strategy (OCS)

The idea of the *optimal completion strategy (OCS)* is to estimate missing values depending on the cluster prototypes in an additional iteration step of the fuzzy c-means algorithm [2]. At the beginning of the algorithm, missing values of incomplete data items are replaced by random values in the feature space. The calculation of membership degrees and the cluster prototypes in the first two iteration steps works in the same way as in FCM. The available and estimated values in the data matrix are not distinguished. In a third iteration step missing values of incomplete data items are estimated depending on all cluster prototypes as follows:

$$x_{kj} = (\sum_{i=1}^{c} (u_{ik})^m \mu_{ij}) / (\sum_{i=1}^{c} (u_{ik})^m) \qquad (6)$$

for $1 \leq k \leq n$ and $1 \leq j \leq d$.

In contrast to the aforementioned two approaches, the advantage of OCSFCM in respect of finding the optimal number of clusters is that both membership degrees and feature values (available and estimated) can be used by the cluster validity functions.

2.2.4. Nearest Prototype Strategy (NPS)

The *nearest prototype strategy* (NPS) [2] is a modification of the optimal completion strategy. The missing values of an incomplete data item are completely substituted by the corresponding values of the cluster prototype to which this data item has the highest membership degree or the minimum partial distance, respectively. Thus, in the third iteration step of NPSFCM missing values are estimated as follows:

$$x_{kj} = \mu_{ij} \text{ with } D_{ik} = \min\{D_{part_{1k}}, ..., D_{part_{ck}}\} \tag{7}$$

for $1 \leq k \leq n$ and $1 \leq j \leq d$.

In respect of applying cluster validity functions, NPSFCM has the same advantages as OCSFCM.

3. Cluster Validity Criteria for Fuzzy Clustering Incomplete Data

A common method for determining the optimal number of partitions is to test the clustering algorithm for different numbers of clusters [8]. After each trial, the clustering results are assessed using a cluster validity function. The clustering with the best value for the cluster validity function is rated as the optimal partitioning of a data set. In the literature, various cluster validity functions for fuzzy clustering have been proposed, which consider different aspects of an optimal partitioning and consequently yield different results for different data distributions. Here, we focus on the most known cluster validity functions as partition coefficient (PC) [1], compactness and separation (S) [10], fuzzy hypervolume (FHV) and partition density (PD) [11]. Below, we describe them and give an idea how to apply them to clustering of incomplete data.

3.1. Partition Coefficient (PC)

A simple cluster validity criterion is the *partition coefficient (PC)*, which rates a partitioning of a data set as optimal if data items are clearly assigned into clusters [1]. This means that membership degrees should be near by 1 or near by 0. Since partition coefficient is not normalised to the number of clusters, in our study we used the normalised version of the partition coefficient proposed in [9]. The *normalised partition coefficient (NPC)* is normalised to the range $[0, 1]$. The high value of NPC indicates a good partitioning of the data set. According to [9] NPC is calculated as follows:

$$NPC(U) = \frac{1}{c-1}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\frac{1}{n}\sum_{k=1}^{n}u_{ik}u_{jk}. \tag{8}$$

Since the calculation of PC and NPC is based only on membership matrix, these cluster validity functions can be applied to fuzzy clustering of incomplete data without any changes.

3.2. Compactness and Separation (S)

In the *compactness and separation (S)* cluster validity function [10], the distances between data points and cluster prototypes are related to the distances between clusters:

$$S(U, X, \mu) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^2||(x_k, \mu_i)||^2}{n \min\limits_{1\leq i,j\leq c,\ i\neq j}||(\mu_i, \mu_j)||^2}, \tag{9}$$

where $||.||$ is the Euclidean norm. This cluster validity function is directly based on the definition of cluster analysis that data items in the same cluster are to be as similar and data items of different clusters are to be as dissimilar as possible. Thus, the smallest value for S indicates an optimal c-partitioning of a data set.

With regard to the applicability to clustering of incomplete data, the compactness and separation cluster validity function requires some changes. The calculation of S involves the calculation of distances between data points and cluster prototypes. Since some clustering methods adapted to incomplete data do not estimate missing values, in our study, we approximate the distances $||(x_k, \mu_i)||^2$ by the partial distances $D_{part}(x_k, \mu_i)$ (cf. Equation 4).

3.3. Fuzzy Hypervolume (FHV)

Another cluster validity function is *fuzzy hypervolume (FHV)* [11]. The fuzzy hypervolume rates a fuzzy clustering of a data set as optimal if the clusters are of minimal volume. The determinant of the covariance matrix of cluster Cov_i is used as a measure for the volume (compactness) of clusters. The fuzzy hypervolume is defined as follows:

$$FHV(U, X, \mu) = \sum_{i=1}^{c}\sqrt{det(Cov_i)} \tag{10}$$

with

$$Cov_i = \frac{\sum_{k=1}^{n}(u_{ik})^m(x_k - \mu_i)(x_k - \mu_i)^T}{\sum_{k=1}^{n}(u_{ik})^m}.$$

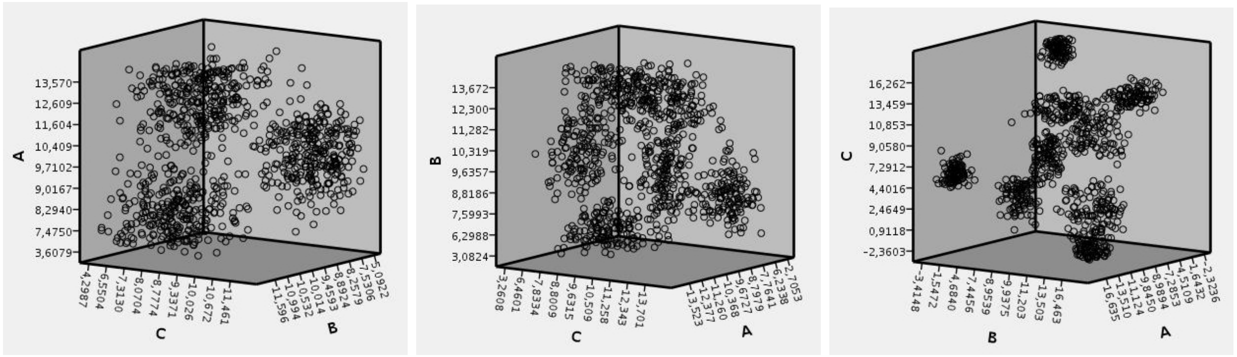
As the compactness and separation validity function the fuzzy hypervolume requires the calculation of distances between data points and cluster prototypes. Pursuing the available case approach, for assessing the cluster tendency on incomplete data we compute the covariance matrices according to [4] as follows:

$$Cov_{i(ml)} = \frac{\sum_{k=1}^{n}(u_{ik})^m I_{k(ml)}(x_{km} - \mu_{im})(x_{kl} - \mu_{il})^T}{\sum_{k=1}^{n}(u_{ik})^m I_{k(ml)}}, \tag{11}$$

where

$$I_{k(ml)} = \begin{cases} 1 & if\ x_{km}\ and\ x_{kl}\ are\ available \\ 0 & else \end{cases}$$

for $1 \leq i \leq c$ and $1 \leq m, l \leq d$.



(a)

(b)

(c)

Figure 1: Test data sets (a) with three, (b) with six and (c) with nine clusters.

3.4. Partition Density (PD)

The fuzzy hypervolume values the quality of a clustering only on the basis of the volume of clusters regardless of their densities. Thus, the large clusters are automatically rated as "bad". To overcome this drawback, the *partition density (PD)* [11] relates the number of data points closely located to cluster prototypes to the volume of clusters:

$$PD(U, X, \mu) = \frac{Z}{FHV} = \frac{Z}{\sum_{i=1}^{c} \sqrt{det(Cov_i)}}, \quad (12)$$

$$\text{where } Z = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}$$

$$\forall\ x_k \in \{x_k \mid (x_k - \mu_i)^T Cov_i^{-1} (x_k - \mu_i) < 1\}. \quad (13)$$

Regarding the applicability of the partition density to incomplete data, this cluster validity measure requires the approximation of the covariance matrix as well as the distances between data points and cluster prototypes. In this study, we approximate the covariance matrix as in FHV (cf. Equation 11) and the distances $dist(x_k, \mu_i)$ by the partial distances $D_{part}(x_k, \mu_i)$ (cf. Equation 4).

4. Data Experiments

In order to test the performance of fuzzy clustering methods adapted to incomplete data and cluster validity functions regarding finding the optimal number of clusters, we have conducted several experiments on three data sets with different numbers of clusters. The data sets were generated by the compositions of three, six and nine 3-dimensional Gaussian distributions, respectively. Each of the data sets consists of 900 data points, which are equally distributed on about same-sized clusters (see Figure 1). Except for a small number of data items, the clusters are well separated in all data sets. Since such a distribution of data is favorable to both the determining the optimal number of clusters and the clear assignment of data items into clusters, the cluster validity functions and the clustering algorithms are expected to work well on the test data. Regarding the comparability of experimental results for different distributions of missing values in data sets, we ensured that the values of different attributes are uncorrelated in all data sets. That is because dependent features do not provide additional information for clustering on the one hand. On the other hand we have observed in further studies that the clustering results on incomplete data with a conditional distribution of missing values are influenced by the correlation between features in the data set.

To generate incomplete data sets the test data have been modified by evenly removing values in all features with probabilities of 10%, 25%, and 40% according to the general missing-data pattern [12]. The percentage of missing values was calculated in relation to all values in the data set. In order to test whether the performance of algorithms depends on a random or conditional reduction of a data set, we removed values from test data according to the common missing-data mechanisms *missing completely at random (MCAR)*, *missing at random (MAR)* and *not missing at random (NMAR)* [12].

In our experiments, we first clustered the complete data sets with basic FCM with different numbers of clusters. We evaluated partitionings of data sets obtained with different cluster validity functions as partition coefficient (PC) [1], partition entropy (PE) [1], normalised partition coefficient (NPC) [9], proportion exponent (PX) [13], compactness and separation index (S) [10], fuzzy hypervolume (FHV) [11], partition density (PD) [11] and average partition density (APD) [11]. Only NPC, S, FHV, PD and APD could correctly determine the optimal number of clusters in all test data sets. Although the clustering structure of data sets was very clear, PC and PE had trouble determining the optimal number of partitions in the data sets. As in other studies, e.g. in [14], these indices underestimated the number of clusters. Due to extreme values out of range we could not compute PX for all numbers of clusters, so we did not determine its optimum.

(a) Averaged optimal number of clusters obtained by WDSFCM on data sets with missing values MCAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***
6C	***6.00***	***6.00***	5.67	***6.00***	***6.00***	5.47	***6.00***	***6.00***	6.93	***6.00***	***6.00***	9.23
9C	8.83	**8.97**	**9.00**	7.30	6.83	7.30	9.33	**9.10**	**9.13**	9.37	**9.13**	11.10

(b) Averaged optimal number of clusters obtained by PDSFCM on data sets with missing values MCAR (asterisks indicate adapted versions of cluster validity measures).

	NPC			S*			FHV*			PD*		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	7.77	7.43	***3.00***	7.40	7.50
6C	**6.10**	**6.07**	5.90	**5.80**	**5.87**	5.27	6.97	9.47	9.10	8.93	9.27	9.17
9C	9.47	9.60	10.10	8.03	7.43	8.17	10.50	10.33	8.80	10.50	9.83	9.33

(c) Averaged optimal number of clusters obtained by OCSFCM on data sets with missing values MCAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	**3.50**	7.03	***3.00***	6.37	7.23
6C	***6.00***	**5.93**	5.23	***6.00***	**6.10**	5.37	***6.00***	**6.30**	8.07	***6.00***	7.30	8.33
9C	**8.90**	8.83	8.67	7.33	6.90	7.40	9.27	10.20	11.00	**9.23**	10.20	11.10

(d) Averaged optimal number of clusters obtained by NPSFCM on data sets with missing values MCAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	**3.03**	***3.00***	***3.00***	**2.97**	***3.00***	**3.33**	7.77	***3.00***	7.60	7.73
6C	***6.00***	**5.97**	5.77	***6.00***	**5.97**	5.30	***6.00***	**6.13**	8.13	***6.00***	9.17	9.13
9C	**9.03**	**8.97**	9.80	8.23	7.97	9.07	**9.13**	**9.10**	10.83	**9.13**	9.70	11.20

Table 1: Over 30 trials averaged optimal number of clusters obtained by different methods on data sets with missing values MCAR.

We clustered the incomplete data with the aforementioned fuzzy c-means algorithms for incomplete data with different numbers of clusters. To create the testing conditions as real as possible, we initialised the cluster prototypes with random values at the beginning. We evaluated obtained clustering results using original and adapted cluster validity functions, respectively. For that reason, we compared the average optimal number of clusters determined by the cluster validity measures over 30 trials of the clustering methods. Note that we did not averaged the values computed by the cluster validity functions because the differences between these values for different numbers of clusters are very small so that an outlier could completely bias the final result. The average number of clusters discovered is only one of the measures for estimating the optimal number of partitions in a data set if the true number of clusters is not known. In this study, we aim to give a deeper insight into the quality of cluster validity indices and clustering methods for incomplete data. Therefore, using the knowledge about the real distribution of test data, we determined the deviation between obtained and true number of clusters. In Tables 1-3 the values are bold-faced if in more than 85% of trials the real number of clusters could be correctly determined so that the cluster tendency was clearly recognisable. The italicised cells indicate the zero deviation from the average number of clusters discovered.

Since in all tests we obtained the same results for APD and PD, below we present the results obtained for NPC, S, FHV and PD organised according to missing-data mechanisms.

4.1. Test Results on Data with Missing Values MCAR

Tables 1 shows the performance results for WDSFCM, PDSFCM, OCSFCM and NPSFCM on data sets with missing values MCAR using different cluster validity functions. For WDSFCM we applied the cluster validity criteria on clusterings produced only on the basis of completely available data items. Since for missing values MCAR the missingness does not depend on the data values (missing or observed) in the data set [12], the structure of the data set is well represented by the complete data items. Due to that the optimal number of clusters could be correctly determined by almost all clustering methods even for a high percentage of missing values in the data sets. Comparing the clustering methods with each other the best performance results were obtained by WDSFCM, OCSFCM and NPSFCM. Since complete data items well represent

(a) Averaged optimal number of clusters obtained by WDSFCM on data sets with missing values MAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	*2.00*	*2.00*	***3.00***	*2.00*	*2.00*	***3.00***	*2.00*	*2.00*	***3.00***	*2.00*	*2.00*
6C	*5.00*	*3.00*	*4.00*	*4.00*	*3.00*	*4.00*	*5.00*	9.23	9.60	*5.00*	8.80	9.20
9C	7.60	6.63	6.23	7.53	6.67	5.67	8.57	11.03	11.63	7.80	11.27	11.47

(b) Averaged optimal number of clusters obtained by PDSFCM on data sets with missing values MAR (asterisks indicate adapted versions of cluster validity measures).

	NPC			S*			FHV*			PD*		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	7.27	7.30	4.17	5.67	6.23
6C	5.77	5.17	4.70	5.17	5.53	4.57	6.37	7.17	8.33	6.90	6.57	8.73
9C	9.00	9.23	5.57	7.37	7.10	7.23	10.87	10.70	8.30	10.43	10.93	8.87

(c) Averaged optimal number of clusters obtained by OCSFCM on data sets with missing values MAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	**3.03**	***3.00***	***3.00***	***3.00***	***3.00***	3.67	5.47	**3.50**	3.67	5.97
6C	5.57	5.37	4.70	5.33	5.47	3.63	**6.07**	6.53	8.07	6.40	6.93	8.07
9C	8.93	7.60	7.00	7.47	6.20	5.77	9.23	9.80	10.77	9.30	10.00	10.53

(d) Averaged optimal number of clusters obtained by NPSFCM on data sets with missing values MAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	3.63	4.60	***3.00***	3.63	5.77
6C	5.37	5.53	4.97	5.37	5.50	4.90	**6.03**	6.27	7.97	**6.03**	6.27	9.13
9C	8.77	8.70	7.93	7.20	6.63	6.40	9.37	9.73	11.07	9.33	10.03	11.13

Table 2: Over 30 trials averaged optimal number of clusters obtained by different methods on data sets with missing values MAR.

the distribution structure of the data set, clusterings produced by WDSFCM are comparable with clusterings produced by the basic FCM on complete data. Estimating missing values by values close to cluster prototypes (by corresponding values of cluster prototypes, respectively), OCSFCM and NPSFCM preserve and strengthen the clustering structure of data. The poorest performance results were obtained by PDSFCM. Experiments conducted in [2, 6] showed that, approximating distances between incomplete data points and cluster prototypes by partial distances, this method assigns data items into clusters as accurately as OCSFCM and NPSFCM for an optimal number of clusters. But in our experiments, the clustering results obtained by PDSFCM were fairly instable regarding the determination of the optimal number of clusters. The number of clusters proposed by this method was underestimated or overestimated from trial to trial. This is due to the fact that the partial distance function does not hold the triangle inequality so that it does not provide reliable estimation of distances for low dimensional data (when the proportion of complete features is relatively small in relation to the number of all features).

Comparing the cluster validity functions the best results were obtained by NPC. Due to the simple distribution structure of data the data items were assigned to the clusters with "precise" membership degrees, what is favorable to the rationale of NPC. Determining the optimal number of clusters the poorest results were obtained for the adapted versions of FHV and PD, which were applied on partitionings produced by PDSFCM. Using only available values for computing the covariance matrices the volumes of clusters were poorly approximated and assessed. There were great differences between covariance matrices computed on complete and incomplete data. Since in most of cases, the optimal number of clusters was correctly determined by other cluster validity indices on the partitionings produced by PDSFCM, the proposed adapted versions of FHV and PD do not seem to be satisfying solutions regarding the cluster tendency assessment in presence of missing values in data.

4.2. Test Results on Data with Missing Values MAR

The performance results in terms of determining the optimal number of clusters obtained by different clustering methods on data sets with missing values MAR are presented in Table 2. Since for missing values MAR the missingness of values depends on values that are observed (and not on components

(a) Averaged optimal number of clusters obtained by WDSFCM on data sets with missing values NMAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	***3.00***	*2.00*	6.63	***3.00***	*2.00*	5.73
6C	*4.00*	*4.00*	3.70	*4.00*	*4.00*	3.70	**6.03**	*4.00*	9.43	7.47	*4.00*	*2.00*
9C	8.47	6.23	4.60	8.43	6.37	3.20	9.60	10.63	10.90	8.50	8.33	10.57

(b) Averaged optimal number of clusters obtained by PDSFCM on data sets with missing values NMAR (asterisks indicate adapted versions of cluster validity measures).

	NPC			S*			FHV*			PD*		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	2.97	***3.00***	***3.00***	3.27	7.27	7.90	7.40	5.77	7.53	7.53
6C	***6.00***	6.00	4.30	*4.00*	6.03	5.63	8.87	8.83	8.30	9.00	9.47	9.03
9C	9.77	7.20	8.40	8.13	7.27	5.10	11.00	9.97	8.63	11.03	9.80	9.13

(c) Averaged optimal number of clusters obtained by OCSFCM on data sets with missing values NMAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	2.90	***3.00***	***3.00***	2.97	***3.00***	6.30	6.40	**3.30**	5.63	6.37
6C	***6.00***	5.20	4.53	*4.00*	5.13	5.50	***6.00***	7.70	8.57	***6.00***	7.40	8.10
9C	8.83	6.73	6.47	8.17	6.93	5.47	9.50	9.73	10.27	9.13	9.97	10.40

(d) Averaged optimal number of clusters obtained by NPSFCM on data sets with missing values NMAR.

	NPC			S			FHV			PD		
	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
3C	***3.00***	***3.00***	3.23	***3.00***	***3.00***	3.10	***3.00***	7.47	7.57	***3.00***	7.47	7.63
6C	***6.00***	5.23	5.57	*4.00*	5.10	4.90	***6.00***	8.57	9.50	***6.00***	8.80	9.73
9C	**8.97**	7.30	6.30	7.93	7.23	5.20	**9.20**	9.60	11.10	**9.07**	10.00	11.20

Table 3: Over 30 trials averaged optimal number of clusters obtained by different methods on data sets with missing values NMAR.

that are missing) [12], we removed values in two attributes depending on values of the third one. In this way, for a certain percentage of missing values there could be more data items afflicted with missing values in a data set with missing values MAR than e.g. in a data set with missing values MCAR. On the other side, there is a completely available attribute in a data set. Since missing values MAR occur in data items with certain properties, the completely available data items do not represent the clustering structure of complete data set as well as in the case of missing values MCAR. This fact provides the explanation for the poor performance of WDSFCM (cf. Table 2 (a)), which carries out the cluster analysis only on the basis of complete data items. On average, the optimal number of clusters determined on clusterings produced by this approach is smaller than the actual one. In contrast, the partitionings of data sets with three clusters produced by OCSFCM and especially by NPSFCM on data with missing values MAR reflects the clustering structure of complete data set almost as well as in the case of missing values MCAR. For a small percentage of missing values, the optimal number of clusters in other data sets could be correctly determined only in at most 50% of trials of OCSFCM and NPSFCM. Whereas, the most stable results were obtained using FHV and PD. In contrast, the most instable results were obtained on clusterings produced by WDSFCM and PDSFCM, especially using the adapted versions of FHV and PD. The “best“ values for c were so varying that no tendency was recognisable regarding the determining the optimal number of clusters.

4.3. Test Results on Data with Missing Values NMAR

Table 3 shows the cluster validity results for clustering algorithms on data with missing values NMAR. Since missing values NMAR induce a conditional reduction of a data set (the missingness of data depends on the missing values themselves [12]), complete data items do not reflect the clustering structure of data set as in the case of missing values MCAR. Therefore, the performance results obtained by WDSFCM slightly differ from the results on data with missing values MAR. Compared to missing values MAR the improvement of the results for this method is based on the fact that values were removed in all attributes and this way more complete data items were contained in data sets. The performance results for OCSFCM and NPSFCM are considerably worse than in case of missing values MCAR. For a small percentage of missing values in data sets the optimal number of clusters could be reliably determined, but for a high percentage of missing values in data sets the clustering structure of data could be determined only for data set with three clusters.

Comparing cluster validity functions with each other a crucial distinction emerges for data with missing values NMAR and MAR. On average, the optimal number of clusters determined by NPC and S is smaller than the actual number of clusters. In contrast, FHV and PD determined a greater number of clusters than the optimum. The reason for this is concerned with the monotonicity properties of these cluster validity functions for increasing number of clusters.

5. Conclusions and Future Work

In this paper, we analysed different cluster validity functions for fuzzy clustering in terms of determining the optimal number of clusters on incomplete data. We proposed some adaptions for cluster validity performance measures involving data matrix for the calculation. Furthermore, in experiments on several data sets we analysed to what extent the clustering results produced by clustering methods for incomplete data reflect the distribution structure of original data. The experimental results have shown that the best performance in terms of cluster tendency assessment was achieved by clustering algorithms OCSFCM and NPSFCM, which estimate missing values by values close to cluster prototypes. In this way, they preserve and strengthen clustering structure of data. In contrast, the clustering results obtained by the partial distance strategy FCM were instable regarding the determining the optimal number of clusters. However, the basic OCSFCM and NPSFCM leave clustering structure (e.g. cluster sizes) out of consideration while estimating missing values. Due to this these methods are instable in assigning data object to clusters on data with differently sized clusters [5]. Therefore, in our future research, we plan to continue working on the improvement of clustering algorithms for incomplete data using cluster dispersion.

Comparing cluster validity functions regarding finding the optimal number of clusters, NPC obtained slightly better results than FHV and PD. However, summarising the results of our study, we do not conclude that NPC is the better cluster validity index for determining the optimal number of clusters. It might produce better results on data with simple distribution structure compared to other indices but, as already mentioned in [14], NPC ignores the geometric structure of clustering. Cluster validity indices like FHV and PD overcome this problem involving the data matrix for calculation. Since not all clustering methods for incomplete data complete the data matrix and the adaptions pursuing the available case approach for FHV and PD do not provide satisfying results, in future we also plan to develop a better adaption of cluster validity functions using volume and density of clusters to incomplete data.

References

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, 1981.

[2] R. J. Hathaway and J. C. Bezdek. Fuzzy c-means Clustering of Incomplete Data, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 31, no. 5, pp. 735–744, 2001.

[3] M. Sarkar and T.-Y. Leong. Fuzzy k-means Clustering with Missing Values. In *Proceedings of American Medical Informatics Association Annual Symposium (AMIA)*, pp. 588–592, 2001.

[4] H. Timm, C. Döring, and R. Kruse. Different Approaches to Fuzzy Clustering of Incomplete Datasets, *International Journal of Approximate Reasoning*, vol. 35, pp. 239–249, 2004.

[5] L. Himmelspach and S. Conrad. Fuzzy Clustering of Incomplete Data Based on Cluster Dispersion. In Proceedings of the 13^{th} *International Conference on Information Processing and Management of Uncertainty* (IPMU 2010), Lecture Notes in Computer Science 6178, pp. 59-68, Springer-Verlag, 2010.

[6] L. Himmelspach and S. Conrad. Clustering Approaches for Data with Missing Values: Comparison and Evaluation. In *Proceedings of the Fifth IEEE International Conference on Digital Information Management* (ICDIM 2010), 2010.

[7] J. K. Dixon. Pattern Recognition with Partly Missing Data, *IEEE Transactions on System, Man and Cybernetics*, vol. 9, pp. 617–621, 1979.

[8] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[9] E. Backer and A. K. Jain. A Clustering Performance Measure based on Fuzzy Set Decomposition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3 (1), pp. 66–74, 1981.

[10] X. L. Xie and G. Beni. A Validity Measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 (8), pp. 841–847, 1991.

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

[12] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*, John Wiley & Sons, 2002.

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

[14] J. C. Bezdek, W. Li, Y. Attikiouzel, and M. P. Windham. A Geometric Approach to Cluster Validity for Normal Mixtures, *Soft Computing*, vol. 1, no. 4, pp. 166–179, 1997.