Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 114 (2017) 224–233

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA

M&MFCM: Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics

Natacha Gueorguieva[a], Iren Valova[b]*, George Georgiev[c]

[a]*City University of New York, Staten Island, NY 10314, USA*
[b]*University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA*
[c]*University of Wisconsin Oshkosh, Oshkosh, WI 96123, USA*

Abstract

The proposed modification of conventional fuzzy C-means clustering (FCM) algorithm aims to correct some of its shortcomings. We have focused on as missing flexibility in cluster number adaptation; limited cluster type grouping; less than optimal objective function for clusters of unequal size lying very close to each other; considerable computational time particularly in case of high dimensional data. With M&MFCM we propose to replace the usual Euclidean distance with Mahalanobis and Minkowski metrics in order to enhance the cluster detection capacity of FCM by allowing more accurate detection of arbitrary shapes of clusters for high dimensional datasets. Direct replacement of Euclidean distance in the objective function of FCM with Mahalanobis might cause numerical problems as the largest eigenvalues of the fuzzy covariance matrix could produce extremely long clusters thus contradicting the real data distribution. The improvement is achieved by fixing the ratio between the maximal and minimal eigenvalues of the covariance matrix. The parameterized Minkowski distance metric is adapted for implementation with FCM with various settings. We also propose an approach for improving the initial choice of cluster number and for visualization and analysis of cluster results for labeled and unlabeled datasets. Experimental results demonstrate that the proposed M&MFCM and test methodology significantly improve FCM clustering results.

Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.

Keywords: fuzzy clustering; cluster validation; distance metric; Silhouette function; confusion matrix; mapping

* Corresponding author. Tel.: +1-508-999-8502; fax: +1-508-999-9144.
E-mail address: iren.valova@umassd.edu

1877-0509
Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.
10.1016/j.procs.2017.09.064

1. Main text

The purpose of unsupervised learning is to attempt to find natural partitions in the data set. Clustering technique plays an important role in data analysis and interpretation as it allows gaining insight into the nature of the data by discovering hidden structures in it. Data are grouped into clusters so that the objects within a cluster have higher similarity in comparison to one another, but are very dissimilar to data objects in other clusters [1]. Clustering is a classification method for data with unknown distribution; the goal is to find the structure hidden in data, and, as much as possible, to bring together the data with similar nature to the same cluster according to some measure of similarity degree. Most clustering algorithms do not rely on assumptions common to conventional statistical methods, i.e. underlying statistical distribution of data, and therefore they are useful in situations where little to no prior knowledge exists. The potential of clustering algorithms to reveal the underlying structures in data can be employed in a wide variety of applications, including classification, image processing, modeling and identification. Identification of clusters for different types of data and application is based on different types of similarity measures as distance, connectivity, and intensity [1, 2].

The clustering methods and algorithms can be broadly classified into hard and fuzzy clustering based upon membership. Hard clustering methods divide the data into crisp clusters where each pattern (sample) belongs to exactly one cluster. These methods usually lead to local optimum. The development of fuzzy set theory proposed the concept of uncertainty of belonging by transforming the crisp boundary concept into a function of the degree of membership [3, 4]. It ranges between 0 and 1 and therefore the data patterns can belong to more than one cluster. The fuzzy clustering methods lead to global optimum.

The effectiveness of pattern clustering (recognition) is highly dependent on the accurate identification of clusters shapes, which can be influenced by considering prototypes with a geometric structure or by using different distance measures in the objective function [5]. Utilizing patterns as prototypes results in spherical clusters (FCM). It applies the Euclidean distance to measure the dissimilarity between two samples and it is therefore expected to work well when all clusters are spheroids or when all clusters are well separated [6].

Using fuzzy covariance matrices results in ellipsoids (Gustafson-Kessel and Gath-Geva clustering), or, with different kind of prototypes, may result in fuzzy linear varieties (FCV) where the clusters are linear subspaces of the feature space. The latter approaches are driven by optimization of a generalized fuzzy *c*-prototypes functional defined by a measure of similarity between data and cluster center (i.e., prototype) where the fitting prototypes are either straight lines while the measure is the orthogonal distance one, or more generally, prototypes that are convex combinations of points and lines [7, 8]. Authors in [9] propose a modified FCM with entropy regularization and applied it to FCV, which resulted in obtaining linear or ellipsoidal clusters.

Most of the research related to FCM modifications with Mahalanobis and, more recently, with Minkowski distance is directed towards problems of image segmentation [10, 11]. The main goal of this exploration is to detect objects or divide the image into regions which can be considered homogeneous according to a given criteria, such as color, motion, texture etc.

In this research we present an extension of FCM algorithm by using Mahalanobis (MFCM) and parametric Minkowski (&MFCM) dissimilarity distances which are suitable for processing datasets with high dimensions and high number of clusters. The proposed initialization and use of validation measures to determine the initial number of clusters is very effective. We also integrate Sammon mapping and Silhouette function for data visualization and analysis of the clustering results.

This paper is organized as follows: Section 2 introduces objective function fuzzy clustering algorithm, cluster validity measures appropriate for fuzzy clustering as well as an approach for visualization and analysis of experiments for the cases of labeled and unlabeled datasets. Section 3 describes the proposed algorithms MFCM and &MFCM. In Section 4 we present the experimental results and the last Section 5 contains conclusions.

2. Objective function fuzzy based clustering

The objective of function-based fuzzy clustering algorithms is partitioning the data into clusters so that the similarity of data objects within each cluster is maximized and the similarity of data objects among clusters is minimized. In general, the objective function has the following arguments: data matrix, membership matrix and

prototypes of clusters. As this function measures the overall dissimilarity of data objects within each cluster, its minimization corresponds to the best partition of the data set.

Let us assume that a data set has N data objects $\{ \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \}$ each of which has P features and is therefore presented with an $P-$ dimensional feature vector $x_{\boldsymbol{k}} = [x_{k1}, x_{k2}, ..., x_{kP}] \in \Re^{N \text{x} P}$. Then the dataset $\boldsymbol{X}$ is represented as a N x P matrix where each row is an object. Crisp partitioning of $\boldsymbol{X}$ assumes that the number of clusters K is known and it is defined through obtaining of family of subsets $\boldsymbol{Z}$ where $\{\boldsymbol{Z}_i \,|\, 1 \le i \le K\}$ with the following properties [1]:

$$\bigcup_{i=0}^{K} \boldsymbol{Z}_i = \boldsymbol{X}\ ,\ \boldsymbol{Z}_i \bigcap \boldsymbol{Z}_j = \varnothing,\ 1 \le i \ne j \le K\ ,\ \varnothing \subset \boldsymbol{Z}_i \subset \boldsymbol{X},\ 1 \le i \le K \tag{1}$$

Equation (1) determines that the union $\boldsymbol{Z}$ of all subsets $\boldsymbol{Z}_i$ contains all the data $\boldsymbol{X}$. The subsets are disjoint and none of them is empty.

Fuzzy clustering algorithms partition the dataset into K fuzzy clusters where $2 \le K < N$. Each fuzzy partition is represented by a fuzzy KxN membership matrix

$$U = [\mu_{ik}]_{K\text{x}N} \tag{2}$$

μ_{ik} is the membership degree of data object $\boldsymbol{x}_k$ in cluster K_i which satisfies the following constraint:

$$\mu_{ik} \in [0,1],\ 1 \le i \le K,\ 1 \le k \le N;\ \sum_{k=1}^{K} \mu_{ik} = 1, \forall k = 1,2,...,N$$

$$0 \le \sum_{i=1}^{N} \mu_{ik} < K,\ 1 \le k \le N \tag{3}$$

Each K_i cluster is represented by a $\boldsymbol{P}-$ dimensional vector $\boldsymbol{v}_i = [v_{i1}, v_{i2}, ..., v_{iP}] \in \Re^P$ called cluster prototype, which corresponds to the center of this cluster. Therefore, the squared inner product distance norm can be represented with the following equation:

$$\boldsymbol{D}_{ik\,\boldsymbol{Z}}^2 = \left\| \boldsymbol{x}_k - \boldsymbol{v}_i \right\|_{\boldsymbol{Z}}^2 = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T \boldsymbol{Z} (\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{4}$$

2.1. Fuzzy C-means

Fuzzy C-Means (FCM) is a popular distance-based objective function method, which belongs to the set of methods satisfying the sum of squared error criterion [1].

The objective function of FCM is given by

$$J_m^{FCM}(\boldsymbol{U}, \boldsymbol{X}, \{\boldsymbol{A}_i\}) = \sum_{i=1}^{K} \sum_{k=1}^{N} (\mu_{ik})^m\, D_{ik\,\boldsymbol{A}_i}^2 \tag{5}$$

$\boldsymbol{U}$ is the partition matrix, $\boldsymbol{A}_i$ is local norm-inducing matrix which is used as one of the optimization variables in the functionals (1) and (3); and $\boldsymbol{X}$ is a given set of P -dimensional unlabeled data, where $\boldsymbol{U} = [\mu_{ik}]_{K \text{ x } N} \in [0, I]$, $\boldsymbol{X} = \Re^{n \text{ x } N}$, μ_{ik} is the membership degree of data object $\boldsymbol{x}_k$ in cluster K_i which satisfies the following constraint:

$$\sum_{i=1}^{K} \mu_{ik} = 1, \forall\, k = 1,2,...,N \tag{6}$$

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{K} \left(\frac{d_{ij}}{d_{il}} \right)^{\frac{2}{m-1}}},\ i = 1,2,...,K;\ j = 1,2,...,N \tag{7}$$

where $d_{ik}^2 = \sum_{j=1}^{K} d_{ij}^2$ and K represents the number of clusters which needs to be specified prior to running the algorithm.

The performance of the FCM is controlled by the weighting exponent (fuzzifier) $m \in [1, +\infty)$, which impacts the final C-partition of the algorithm. The range of m values recommended in [13] is in the interval $[1.5, 2.5]$. However, further studies indicate that more effective boundaries for the level of fuzziness m is the interval $[1.4, 2.6]$ as it creates significant changes in the FCM membership values. Therefore the latter interval should be considered as more effective [14]. If m approaches a value of one, the partition becomes crisp. The value of m needs to be specified prior running the algorithm. The recommended value of m is two.

The distance norm D_{ikA_i} determined by (1) accounts for clusters of different geometrical shapes in one data set:

$$D^2_{ik,A_i} = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T \boldsymbol{A}_i (\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{8}$$

where the prototype vector $\boldsymbol{v}_i = \|v_{i1}, v_{i2}, ..., v_{in}\| \in \Re^n$ is the center of the cluster K_i. Minimizing the objective function (1) with the constraint (2) and continuing parameters $\boldsymbol{v}_i$ and μ_{ik} requires an alternating optimization scheme as the analytical solution does not exist [3]. One of the drawbacks of FCM algorithm with implementation of Euclidean distance is that it ignores the similarity between some attributes and does not work well in high dimensions. The solution proposed in [10] utilizing Mahalanobis distance works well for image segmentation requiring particular pixel treatment (preprocessing) and also does not work well in higher dimensions [11].

2.2. Cluster validation

One of the most important issues in cluster analysis is the evaluation of clustering results which is the main focus of cluster validity. The latter refers to finding a partitioning that best fits the original data. In general, clustering methods try to discover significant groups present in a data set by searching for clusters with members that are close to each other (have a high degree of similarity) and at the same time are well separated.

The correctness of clustering algorithm results is verified using appropriate criteria and techniques. The two criteria for clustering evaluation and selection of an optimal clustering scheme include compactness and separation [1]. The first one requires the members of each cluster to be as close to each other as possible, which defines the variance and its minimization as a possible measure for compactness. The second criteria - separation - is based on measuring the distance between different clusters, where it can be defined as a distance between the closest and the most distant members of the clusters or the distance between cluster centers.

Here, we use two different categories of fuzzy validation measures: one based only on the matrix $\boldsymbol{U} = \lfloor \mu_{ij} \rfloor$ of membership values of fuzzy data partition (partition coefficient (PC) and classification entropy (CE)) and a second category, which involves both the $\boldsymbol{U}$ matrix and the dataset itself (partition index (SC), separation index (S), and Xie and Beni's index (XB)). The drawback of the first category comes from the lack of direct relation to some properties of the datasets.

PC measures the amount of "overlapping" between clusters. The PC index values range in $[1/K, 1]$, where K is the number of clusters. The optimal number of cluster defines the maximum value of PC. If all membership values of particular fuzzy partition are equal, that is $\mu_{ij} = 1/K$, the PC obtains its lowest value [1].

The formula of CE, which similarly to PC measures the fuzziness of the cluster, is computed for values of c greater than 1 and its values range in the interval $[0, \log_a K]$. The closer the value of CE is to 0, the harder the clustering is. As in the case of PC, the values of CE closer to the upper bound (i.e., $\log_a K$), indicate an absence of any clustering structure in the dataset or inability of the algorithm to extract it.

The next three validation measures belong to the second category, i.e. they involve the membership values as well as the dataset by using fuzzy deviation and cluster variation. The fuzzy deviation of x_j from cluster i, denoted as d_{ij} is defined as the distance between x_j and the center of the cluster weighted by the fuzzy membership function μ_{ij}. Therefore, $d_{ij} = \mu_{ij} \|x_j - v_i\|$.

For cluster i, the summation of the squares of fuzzy deviation of the data point in dataset $\boldsymbol{X}$, denoted as σ_i, is called variation of cluster i. Respectively, the summation of the variations of all clusters σ is called a total variation

of the data set. The compactness of cluster i is determined by $\pi = \sigma_i / N_i$, where N_i is the number of patterns in $i-th$ cluster. Therefore, π is the average variation in cluster i. The separation of fuzzy partition is defined as the minimum distance between the cluster centers given by $d_{min} = min\|v_i - v_j\|$. The next three validation measures satisfying the second criteria are based on the above definitions.

SC is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster. Unlike SC, the separation index S uses a minimum-distance separation for partition validity [4].

Xie and Beni's validation measure quantifies the ratio of the total variation within clusters and the separation of clusters. If clusters are compact and well separated, XB values are small [4].

We have also implemented the Dunn's index (DI) and alternative Dunn index (ADI) although they are suitable for crisp clustering and are not really reliable for the case of overlapping clusters. ADI modifies the original Dunn's index in order to simplify the calculation.

2.3. Fuzzy C-means

Silhouette function (SF) is a very efficient way to analyze the separation distance between the resulting clusters especially in situations of missing dataset class label. SF plot provides a visual measure with a range of [-1, +1] and is a way to assess how close each pattern in one cluster is to patterns in the neighboring one. Therefore, silhouette value measures the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations having values near 1 and poorly clustered observations having values near -1. Silhouette values near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. It is defined as $S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$, where a_i is the average distance between i and all other observation in the same cluster and b_i is the average distance between i and the observations in the nearest neighboring cluster. It is defined by

$$b_i = min_{K_n \in \wp \setminus \boldsymbol{K}(i)} \sum_{j \in K_n} \frac{dist(i,j)}{l(K_k)},$$

where $\boldsymbol{K}(i)$ is the set of all created clusters during the iteration i, $dist(i,j)$ is the distance metric (Euclidean, Mahalanobis etc. used between observations i and j, $l(\boldsymbol{K})$ is the cardinality of set of clusters $\boldsymbol{K}$ [15]. The thickness of the silhouette plot is also very important for the cluster analysis. If very thick for a particular cluster, it most probably includes a group of clusters and therefore further experiments with increased number of clusters are necessary. When all plots are more or less of similar thickness and hence are of similar sizes it can be a sign for correct cluster number choice.

3. Equations

In cluster analysis, the choice of appropriate dissimilarity measure is a requirement. The metric of each cluster is defined by a local norm-inducing matrix $\boldsymbol{A}_i$, which is used as one of the optimization variables in the functionals (1) and (3). Generally, FCM employs Euclidian distance measure, which induces hyperspherical clusters, i.e. $\boldsymbol{A} = \boldsymbol{I}$ and therefore it has poor performance when the clusters have a hyper-ellipsoid shape.

Mahalanobis measures the distance between sample and center for each group in a $P-$ dimensional space defined by N variables and their covariance. If we modify the distance norm presented with (3) in the following way

$$\boldsymbol{D}^2_{ik,\boldsymbol{Z}_i} = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T \boldsymbol{Z}_i (\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{9}$$

where $\boldsymbol{Z}_i = \boldsymbol{Q}_i^{-1}$, $\boldsymbol{Q}_i$ is the inverse covariance matrix of cluster i with additional volume constrain on the determinant of $\boldsymbol{Z}_i$: $|\boldsymbol{Z}_i| = \rho_i,\ \rho_i > 0,\ \forall i$ is the typical way to find a reasonable solution for clusters of different geometrical shapes in one data set. Allowing the matrix $\boldsymbol{Z}_i$ to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume remains constant and at the same time adapting the distance norm to the local topological structure of the data.

The approach we developed when applying Mahalanobis distance in FCM uses Lagrange multiplier method and is similar to the one proposed in [16] combined with the modification proposed in [17]. The latter prevents FCM with Mahalanobis distance from running into numerical problems but at the same time can cause certain degree of overfitting for larger number of clusters. This might appear when clusters are extremely long in the direction of the largest eigenvalues and have little relationship with the real distribution of the data.

The expression used for $\boldsymbol{Z}_i$ calculation is given as

$$\boldsymbol{Z}_i = \left[\rho_i \det(\boldsymbol{F}_i)\right]^{1/n} \boldsymbol{F}_i^{-1} \tag{10}$$

where $\boldsymbol{F}_i$ is the *fuzzy covariance matrix* of the i-th cluster determined by

$$\boldsymbol{F}_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m (\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T}{\sum_{k=1}^{N} (\mu_{ik})^m} \tag{11}$$

Minkowski norm provides a concise distance function that generalizes Euclidean and Manhattan distances and allows various mathematical modifications for a whole class of distance functions suitable for different applications. Minkowski norm is defined by

$$D_m(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt[m]{\sum_{k=1}^{P} \left|\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}\right|^m} \tag{12}$$

where $m \geq 1$ is a parameter which can be adjusted for the specific clustering problems. When the value of $m = 1$, we have the Manhattan distance and the case $m = 2$, corresponds to Euclidean distance.

4. Experiments

We evaluate the performance of the proposed MFCM and &MFCM using 3-D generated datasets [18] as well as two real datasets with high dimensions [19]. We use Sammon mapping, confusion matrix and SF in order to visualize the clustering results for high dimensional datasets and to analyze the number of clusters initially set by ISODATA and validation measures. The choice of Sammon mapping is based on the fact that it tries to preserve interpattern distances, respectively the structure of high dimensional data.

4.1. Experiments with generated data sets

The experiments shown here are performed on two 3-D generated datasets for the purposes of cluster visualization and analysis of the proposed validation measures.

The 3D Tetra dataset consists of 400 data samples and 4 classes, all within a close proximity of each other. This data set examines the algorithm's performance in identifying multiple clusters, even when the patterns are extremely close in space. The four clusters are each almost square in shape, with each cluster filling a quadrant of a larger square (Fig.1).

Since Tetra data is organized into almost spherical patterns, the FCM with Euclidean norm was able to detect the clusters, and the results returned by the confusion matrix prove 100% algorithm accuracy.

While the FCM algorithm with Euclidean metric works well for the Tetra data set, of implementation of proposed modification MFCM to detect two of the four clusters (Fig.2). The calculated confusion matrix proves that 50% of samples are misidentified. The primary reason for this inaccuracy is the use of the diagonal norm, as MFCM

algorithm utilizes a co-variance matrix. The latter provides important information about the orientation and shape of the cluster as the directions of hyper-ellipsoid axes are determined by the eigenvectors.

HeptaM is a 3-D dataset with 399 patterns distributed between 7 classes. Implementation of partition coefficient (PC) validation for FCM with Euclidean metric determined the optimal number of clusters to be 5 (Fig.3) while PC of MFCM accurately calculated it as 7. The visualization of MFCM results are demonstrated with Sammon mapping in Fig.4.

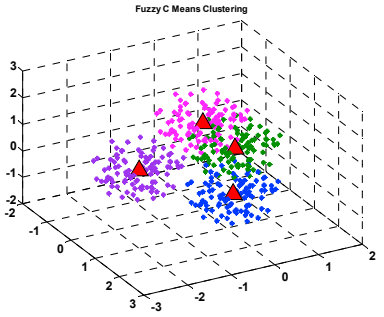

Fig. 1. FCM clustering result for 3-D Tetra

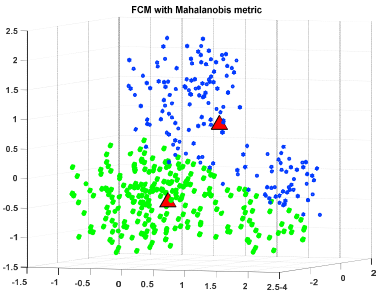

Fig. 2. MFCM clustering result for 3-D Tetra

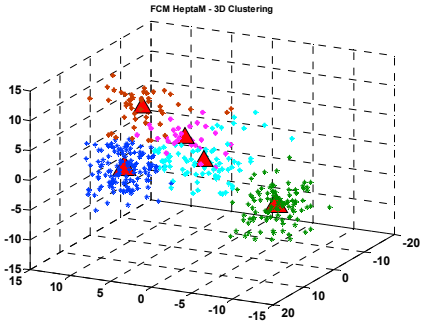

Fig. 3. FCM clustering result for 3-D HeptaM

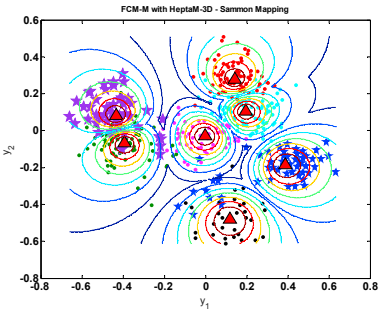

Fig. 4. Sammon mapping of MFCM results for HeptaM

Table I. Validation measures for FCM with Seeds dataset

Validation Measures for FCM -Seeds dataset				
	2 clusters	3 clusters	4 clusters	5 clusters
PC	0.7649611	0.7726316	0.638	0.5553453
CE	0.3829514	0.442937	0.697	0.8663473
SC	0.9254791	0.4363953	0.561	0.4553764
S	0.004407	0.0025337	0.003	0.0026241
XB	2.0621354	3.0293901	2.302	2.0589009
DI	0.0855278	0.215468	0.086	0.069951
ADI	0.0209543	0.0044056	0.001	0.0001

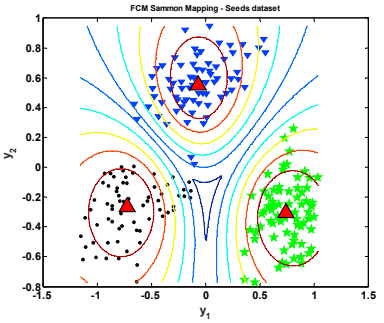

Fig. 5. Sammon mapping of FCM results for Seeds

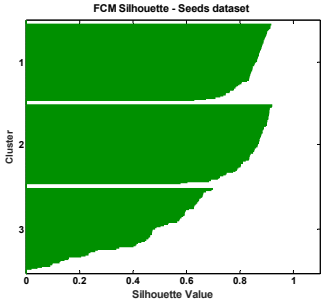

Fig. 6. FCM Silhouette function results for Seeds

Confusion matrix for FCM shows that FCM accurately clusters 64.8% of HeptaM while the number for the proposed MFCM increases to 93.41%.

Table II. Validation measures for MFCM for Housing dataset

FCM Validation Measures - Housing dataset				
	2 clusters	3 clusters	4 clusters	5 clusters
PC	0.87011	0.64714	0.54622	0.47488
CE	0.24133	0.60031	0.83007	1.02137
SC	0.38229	0.27738	0.29396	0.32601
S	7.57E-04	7.42E-04	7.84E-04	8.88E-04
XB	2.60838	2.02322	2.01215	1.52928
DI	0.83238	0.03675	0.05432	0.04885
ADI	0.00752	0.00248	0.00430	8.29E-04

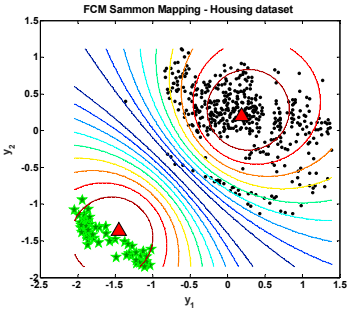

Fig. 7. Sammon Mapping of FCM results for Housing

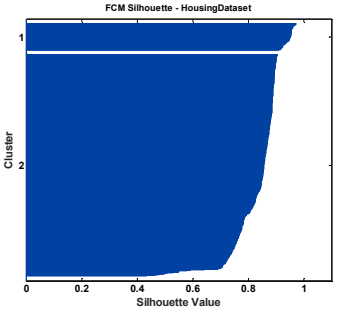

Fig. 8. FCM Silhouette function results for Housing

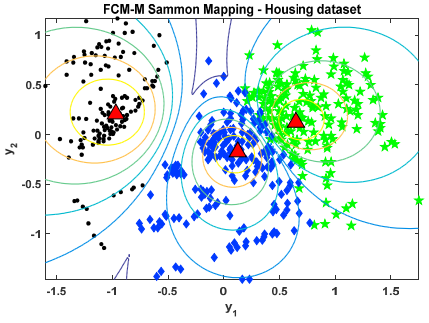

Fig. 9. Sammon mapping of MFCM results

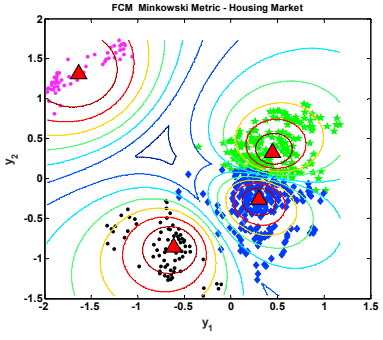

Fig. 10. Sammon Mapping of &MFCM results

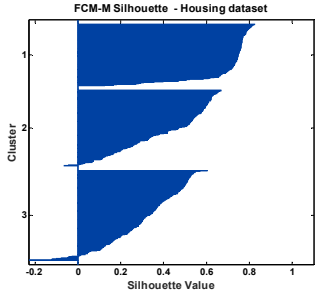

Fig. 11. MFCM Silhouette function results for Housing

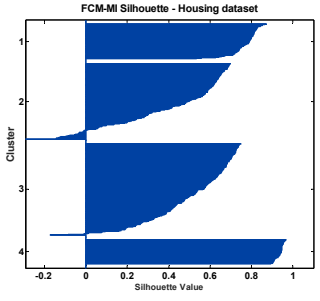

Fig. 12. &MFCM Silhouette function results for Housing

4.2. Experiments with real data sets

Seeds dataset was developed on the base of measurements of geometrical properties of kernels belonging to three different varieties of wheat, respectively Kama, Rosa and Canadian. It includes 210 instances with seven real value

attributes. These are area, parameter and compactness, length of kernel, width of kernel, asymmetry of coefficient and length of kernel groove and they were constructed on the base of soft X-ray technique and GRAINS package. The proposed imaging techniques are considered more sophisticated and less costly than the usually used scanning microscopy or laser technology. The wheat grains are harvested from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences. The respective cluster validation measures (Table I) determined the choice of cluster number FCM algorithm equal to 3.

As the seeds dataset is 7-dimensional, the clustering results are visualized with Sammon mapping presented in Fig.5.

Analysis of confusion matrix and implementation of Silhouette function with cosine distance (Fig.6) proved 96.19% correct classification (respectively 8 misclassified patterns) after implementation of FCM. The results of MFCM and &MFCM on seeds are similar with correct classification rate respectively 91.42% and 94.76% corresponding to 18 and 11 misclassified patterns.

The Sammon mapping of housing dataset results after implementation of FCM with two clusters and the SF are shown in Fig.7 and Fig.8.

Housing dataset with 506 instances and 14 attributes contains information collected by the U.S Census Service from Boston Area. ISODATA determines that the optimal number of clusters is 3. The validation measures presented in Table II show that run with FCM show that 2 is the best value of cluster numbers for this clustering algorithm with Euclidean distance.

Analysis of Fig.8 confirms that FCM needs experiments with various number of clusters as the thickness of the second cluster in Fig.8 assumes the presence of a cluster group.

The proposed MFCM and &MFCM algorithms on the base of validation measures calculated the optimal number of clusters respectively to 3 and 4. The corresponding Sammon mappings are shown in Fig.9 and Fig.10.

We performed an in-depth experimental test on the two proposed algorithms M&MFCM for housing dataset with cluster number varying between 2 and 6. The analysis of the results on the base of silhouette function confirm that the optimal values of clusters respectively for MFCM and &MFCM are the ones determined by the PC calculated for these algorithms. The results of silhouette functions for are demonstrated in Fig.11 and Fig.12.

Experiments performed with &MFCM also included changing the value of parameter $2 \le m \le 6$. Based on the silhouette analysis we concluded that the best value of m for the housing dataset is 2.

5. Conclusions

The modifications proposed in this paper extend the capabilities of FCM for accurate identification of cluster shapes and is supported by the results of high dimensional application. Embedded into the approach are Sammon mapping and silhouette function for better visualization and analysis of clustering results.

Some of the future works include: testing of the modified FCM in other problem domains, establishing a mathematical model for determining their initialization parameters as well as developing of modification which includes more generalized version of Minkowski norm

References

[1] J.C.Bezdek, Pattern Recognition with fuzzy objective function algorithms. Plenum Press, New York, 1981.
[2] J.Kang, L.Min, Q.Luan, Novel modified fuzzy c-means algorithm with applications. Digital Signal Processing, 2009, 19: 309-319.
[3] Dunn, J.C., Well Separated Clusters and Optimal Fuzzy Partitions. J. Cybernetics, 1974, Vol. 4, pp. 95-104.
[4] T.Takagi, M.Sugeno. Fuzzy identification of systems and its application to modeling and control. IEEE Transactions on Systems, Man and Cybernetics, 1985, 15(1):116–132.
[5] Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance Metric Learning, with Application to Clustering with Side-information. In: Advances in NIPS, pp. 505–512. MIT Press, Cambridge (2002)
[6] D.S. Yung and X.Z. Wang, Improving performance of similarity based clustering by feature weight learning, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2002), 556-561.
[7] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson: Detection and characterization of cluster substructure. I. linear structure, fuzzy c-lines, *SIAM J. Appl. Math.*, 40, 2, 339-357 (1981).
[8] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson: Detection and characterization of cluster substructure. II. fuzzy c-varieties and convex combinations thereof, *SIAM J. Appl. Math.*, 40, 2, 358-372 (1981).
[9] H. Ichihashi, K. Honda, N. Tani, Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization, in: 4th Asian Fuzzy System Symposium, 2000, pp. 217–221.

[10] Y. Zhang, Z. Li, J. Cai, J. Wang, Image Segmentation Based on FCM with Mahalanobis distance. Proceedings ICICA'10 of the First International Conference on Information Computing and Applications, pp. 205-212.
[11] A. Ghosh, N.S. Mishra, S. Ghosh, Fuzzy clustering algorithms for unsupervised change detection in remote sensing images, Information Sciences 181(2011) 699–715.
[12] Abder-Rahman Ali, Adeläıde Albouy-Kissi, Antoine Vacavant, Manuel Grand-brochier, Jean-Yves Boire, A Novel Fuzzy C-Means Based Defuzzification Approach with an Adapted Minkowski Distance. 19th Computer Vision Winter Workshop, Krtiny, Czech Republic, February 3–5, 2014.
[13] Nikhil Pal and James Bezdek. *On Cluster Validity for the Fuzzy C-Means Model*. IEEE Trans. Fuzzy Syst 3, pages 370-379, 1995.
[14] Ming Huang, Zhixun Xia, Hongbo Wang, Qinghua Zeng, and Qian Wang. *The range of the value for the fuzzifier of the fuzzy c-means algorithm*. Pattern Recognition Letters, volume 33, issue 16, pages 2280-2284, 2012.
[15] M Maechler, P Rousseeuw, A Struyf, M Hubert, K Hornik, Cluster: cluster analysis basics and extensions. R package version 1 (2), 56, 2012.
[16] Gustafson D, Kessel W, Fuzzy clustering with fuzzy covariance matrix, In: Proceedings of IEEE CDC, San Diego, CA, USA, 1979, pp 761–766.
[17] R. Babuska, P.J. van der Veen, U. Kaymak, Improved covariance estimation for Gustafson-Kessel clustering. Proceedings of the IEEE international conference on Fuzzy Systems. Honolulu, 2002, pp. 1081-1085
[18] Ultsch, A., Clustering with SOM: U*C, In Proc. Workshop on Self-Organizing Maps, Paris, France, (2005) , pp. 75-82.
[19] Lichmsn M., UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2012.