Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **19** (2019) **011304** (92-102) AKU J. Sci. Eng. **19** (2019) **011304** (92-102)

Doi: 10.35414/akufemubid.429540

Araştırma Makalesi / Research Article

Comparison of FCM, PCM, FPCM and PFCM Algorithms in Clustering Methods

Özer Özdemir[1*], Aslı Kaya[2]

[1,2]Eskişehir Technical University, Faculty of Science, Department of Statistics, Eskişehir, TURKEY.

e-posta: [1]ozerozdemir@eskisehir.edu.tr, ORCID ID: http://orcid.org/0000-0003-2446-5139

e-posta: [2]asli.k@eskisehir.edu.tr, ORCID ID: http://orcid.org/0000-0003-2155-9391

[*]Sorumlu Yazar/Corresponding Author

Geliş Tarihi: 01.06.2018 ; Kabul Tarihi: 12.03.2019

Keywords

Fuzzy c-means; Possibilistic c-means; Fuzzy possibilistic c-means; Possibilistic fuzzy c-means

Abstract

Clustering is a process of dividing the objects into subgroups so that the same set of data is similar, but the data of different clusters is different. The basis of the fuzzy clustering algorithms is the C- Means families and the strongest algorithm is the Fuzzy C-means (FCM) algorithm. In this study; FCM, Possibilistic Fuzzy C-means (PFCM), Fuzzy Possibilistic C-means (FPCM) and Possibilistic C- means (PCM) algorithms are used to classify the several real data sets which are E.coli, wine and seed data sets into different clusters by MATLAB program. Also, the results of PFCM, FPCM, PCM and FCM algorithms are compared according to the classification accuracy, root mean squared error (RMSE) and mean absolute error (MAE). The results show that the PFCM and FPCM algorithms have better performance than FCM and PCM according to criteria for comparing the performances.

Kümeleme Yöntemlerinde BCO, OCO, BOCO ve OBCO Algoritmalarının Karşılaştırılması

Anahtar kelimeler

Bulanık c- ortalamalar; Olabilirlikli c-ortalamalar; Bulanık olabilirlikli c-ortalamalar; Olabilirlikli bulanık c-ortalamalar

Öz

Kümeleme, nesneleri özelliklerine göre kümelere bölme işlemidir, böylece aynı veri kümesi benzerdir, farklı kümelerin verileri farklıdır. Bulanık kümeleme algoritmalarının temeli C- ortalamalar aileleridir ve en güçlü algoritma Bulanık C- ortalamalar (BCO) algoritmasıdır. Bu çalışmada; BCO, Olabilirlikli Bulanık C-ortalamalar (OBCO), Bulanık Olabilirlikli C-ortalamalar (BOCO) ve Olabilirlikli C- ortalamalar (OCO) algoritmaları, E.koli, şarap ve tohum veri setleri olarak ifade edilen birkaç gerçek veri setini farklı kümeler halinde sınıflandırmak için MATLAB programı vasıtasıyla kullanılmıştır. Ayrıca, OBCO, BOCO ve OCO ve BCO algoritmaları sonuçları sınıflandırma doğruluğuna, hata kareler ortalamasının karekökü (HKOK) ve ortalama mutlak hata (OMH) değerlerine göre karşılaştırılmıştır. Deney sonuçları, performans karşılaştırmada kullanılan kriterlere göre OBCO ve BOCO algoritmalarının BCO ve OCO algoritmalarından daha iyi performansa sahip olduğunu göstermektedir.

1. Introductıon

Data analysis is necessary when we want to get some knowledge about the system. If data are unlabeled, we need clustering methods in order to associate a label to a subset of data that are slightly close together. Clustering of any data set is a process of partitioning the data set into subgroups. Let the number of data points in a data set X be n, then the number of subgroups c is such that $1<c<n$. The output of the models is generally a set of (cn) values $\{u_{ik}\}$ that can be conveniently arrayed as $(c*n)$ matrix $U=[u_{ik}]$.The clustering algorithms use a distance norm to calculate the membership values. Generally, the Euclidean Distance norm is used. Clustering analysis can be grouped under two headings as hard and fuzzy (Bora

and Gupta 2014). The hard (conventional) clustering methods restrict that each point of the data set belongs to exactly one cluster. That is, in hard clustering analysis membership values only take 0 and 1 values (Cebeci and Yildiz 2015). However, in practice clusters may overlap and data points may belong to more than one cluster. In this case the membership degrees of a data point to clusters should be a value between zero and one (Cebeci et al. 2017).

It has been considered an important point in the evaluation of the concept of uncertainty in modern sense (Zadeh 1965). It has been revealed the fuzzy set theory of objects with imprecise boundaries (Berry 2003). Zadeh introduced the idea of partial memberships described by membership functions. The fuzzy clustering analysis allows the data to belong to more than one cluster by using multiple membership values. This membership values takes values between 0 and 1 (Zadeh 1965).

One of the most widely used fuzzy clustering algorithms is FCM algorithm (Bezdek 1981). Object function of FCM algorithm:

$$J_{FCM}(V,U,X)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\left\|x_k-v_i\right\|,\quad 1<m<\infty \quad (1)$$

where n is the total number of patterns in a given data set and c is the number of clusters; m is a factor which defines the fuzziness degree of the partition; $X=\{x_1,\ldots,x_n\}\subset R$ and $V=\{v_1,\ldots,v_c\}\subset R$ are the feature data and cluster centroids; $U=[u_{ik}]$ and $c*n$ is a fuzzy partition matrix composed of the membership degree of pattern x_k to each cluster i. The weighting exponent m is called the being effective on the clustering performance of FCM algorithm (Şahinli 1999).

FCM algorithm is a partition algorithm (Singhal and Deepika 2016). Regardless of how many "clusters" actually have in the data set, it finds a fuzzy c-partition in a particular data set. Also, the main constraint of FCM algorithm is that the sum of each column in membership matrix U must be equal to 1. FCM algorithm produces the memberships of the data points that are related to the distance of that data point from the centers of the clusters. If a data point is equidistant from the clusters, then it will have the same membership value in each cluster. FCM algorithm has problems dealing with noise and outliers. The main problem with FCM algorithm is that the noise points or the outliers are also accounted in the membership values. The second problem is that FCM algorithm detects spherical clusters. It is not effective in finding other cluster shapes (Ozdemir and Kaya 2018).

In the case of outliers or noise in a data set, FCM algorithm is not appropriate because a single outlier can completely effect the partitioning result in FCM algorithm (Şanlı and Apaydın 2006). Since FCM memberships do not always explain the degrees of belonging for the data well, a possibilistic approach which is called as PCM has been proposed to clustering to correct this weakness of FCM algorithm (Krishnapuram and Keller 1993).

Pal, Pal and Bezdek proposed FPCM model and algorithm that generates both the membership and typicality values (Pal, Pal and Bezdek 1997). Memberships and typicalities are significant for the accurate characteristic of data substructure in clustering problem (Jafar and Sivakumar 2012).

PFCM algorithm is a hybrid algorithm of PCM and FCM algorithms (Pal et al. 2005). PFCM algorithm solves the noise sensitivity defect of FCM algorithm and overcomes the coincident clusters problem of PCM algorithm. But the noise data have an influence on the estimation of centroids. PFCM algorithm creates memberships and possibilities concurrently for each cluster along with the usual prototypes or cluster centers (Timm et al. 2004).

Deciding the number of clusters for all the algorithms mentioned above is an important step. “Validation indices” were proposed to determine the optimal number of clusters. There are several validity indices that are valid in the fuzzy environment.

Saad and Alimi used the Fukuyama-Sugeno index to determine of optimal number of cluster in several data sets in their study (Saad and Alimi 2016).

Ozdemir and Kaya conducted comparisons of the Xie- Beni, Partition Coefficient, Modified Partition Coefficient, Classification Entropy, Kwon, Separation indices in conjunction with the FCM algorithm on widely used data sets. They found some of the mentioned indices incorrectly recognize optimal cluster numbers c for data sets (Ozdemir and Kaya 2018).

Correa et al. performed a comparison of PCM algorithm, FPCM algorithm, Robust fuzzy possibilistic c-means (RFPCM) algorithm and FCM with Gustafson-Kessel algorithm applied to feature extraction on vineyard images. They found that in relation to the runtime, the best performance was obtained for RFPCM algorithm (Correa et al. 2011).

Anderson et al. introduced a new method for comparing soft (fuzzy, probabilistic and possibilistic) partitions based on the earth mover's distance and the ordered weighted average (Anderson et al. 2013).

Ganbold and Chasia were aimed to compare the output of an artificial neural network algorithm and PCM algorithm, an improvement of the FCM algorithm, on both moderate resolutions Landsat 8 and a high resolution Formosat 2 images. PCM algorithm produced a more realistic and reliable result in their study because it considered others' factors like the degree of belongingness, compatibility and typicality to give a possibility of a pixel belonging to a given class (Ganbold and Chasia 2017).

In this paper, we aimed to compare FCM, PCM, PFCM and FPCM algorithms with a validation index called Performance Index (PI) (or updated from Fukuyama-Sugeno index) by using E.coli, wine and seed data sets by writing MATLAB program codes without using any ready-packages as the first in the literature. So, the rest of the paper is organized as follow: Second section explains PCM, FPCM, PFCM algorithms, validation index and criteria for comparing performances of FCM, PCM, FPCM and PFCM algorithms respectively. Third section presents experimental analysis and results using by FCM, PCM, FPCM and PFCM algorithms. Final section is the conclusion and discussion part.

2. Material and Method

2.1 Possibilistic c –means (PCM) algorithm

In order to prevent the outliers from being accounted in, another clustering technique was introduced by Krishnapuram and Keller (1993), named PCM. In contrast to FCM algorithm, membership value generated by PCM algorithm can be interpreted as "degree of belongingness or compatibility or typicality" (Krishnapuram and Keller 1993). Typicality degrees are defined to build prototypes that characterize data subcategories, taking into account both the common points of the category members and their distinctive features as compared to other categories. Typicality values with respect to one cluster do not depend on any of the prototypes of other clusters. Degree of typicality helps the distinction between the highly atypical member of the cluster and the partly atypical member of the group.

PCM algorithm relaxes the row sum constraint of FCM algorithm. The main constraint of PCM algorithm is that each membership value in U can be anything between 0 and 1 or equal to any one of them, i.e. $0 \leq u_{ik} \leq 1$. So these values are called the typicalities of the data points in each cluster. The objective function of PCM algorithm can be formulated as follows:

$$J_{PCM}(V,U,X) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} + \sum_{i=1}^{c} \eta_i \sum_{k=1}^{n} (1-u_{ik})^{m} \quad (2)$$

where n is the total number of patterns in a given data set and c is the number of clusters; m is a factor which defines the fuzziness degree of the partition; d_{ik}^2 is the distances; $U = [u_{ik}]$ is a fuzzy partition of the matrix X.

η_i, is called, "scale" or "typicality" parameter and it estimated from the data. It is calculated as:

$$\eta_i = \frac{\sum_{k=1}^{n} u_{ik}^m \left\| x_k - v_i \right\|^2}{\sum_{k=1}^{n} u_{ik}^m} \quad (3)$$

where n is the total number of patterns in a given data set; $m \in [1, \infty)$ is a parameter which defines the fuzziness degree of the partition; $X = \{x_1, \ldots, x_n\}$ and $V = \{v_1, \ldots, v_c\}$ are the feature data and cluster centroids; $U = [u_{ik}]$ and *c*n* is a fuzzy partition matrix composed of the membership degree of pattern x_k to each cluster i.

The membership value u_{ik}, in case PCM algorithm, will be calculated from the following equation:

$$u_{ik} = \left[1 + \frac{d_{ik}^{2\ \frac{1}{m-1}}}{\eta_i} \right]^{-1} \quad (4)$$

where d_{ik}^2 is the distances; η_i is the scale parameter. PCM algorithm consists of the following steps:

S1: Fix the number of clusters c; fix m, $1 < m < \infty$;
S2: Set iteration counter l=1;
S3: Initialize the possibilistic c-partition $U^{(0)}$;
S4: Estimate η_i
S5: Repeat
S5.1: Update the prototypes using $U^{(l)}$, as indicated below;
S5.2: Compute $U^{(l+1)}$;
S5.3: Increment l;
Until $(\left\| U^{(l-1)} - U^{(l)} \right\| < \varepsilon)$;

Because each data point in PCM techniques is classified as only one set at a time, the clusters do not have too much mobility.

The problem with PCM algorithm is that sometimes, it produces coincident clusters. Since PCM algorithm often produces coincident clusters, FPCM and PFCM algorithms were introduced (Nefti and Oussalah 2004).

2.2 Fuzzy possibilistic c-means (FPCM) algorithm

The objective function of FPCM algorithm includes both memberships and typicalities as shown in Eq. (5)

$$J_{m,\eta}(U,T,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik}^m + t_{ik}^\eta) \left\| x_k - v_i \right\|^2 \quad (5)$$

Subject;

$$m > 1, \eta > 1, 0 \le u_{ik}, t_{ik} \le 1 \quad (6)$$

$$\sum_{i=1}^{c} u_{ik} = 1, \forall k \quad (7)$$

$$\sum_{k=1}^{n} t_{ik} = 1, \forall i. \quad (8)$$

where m and η are the exponents for fuzziness and typicality respectively. Under the 6, 7, 8 constraints and c-means optimization conditions $\sum_{i=1}^{c} u_{ik} = 1$, we will make the following initial conditions or extreme of $J_{m,\eta}(U,T,V)$ in terms of Lagrange multiplier theorem as follows:

$$u_{ik} = \left[\sum_{j=1}^{c} \frac{d_{ik}^{2\ \frac{2}{m-1}}}{d_{jk}^2} \right]^{-1}, 1 \le i \le c; 1 \le k \le n \quad (9)$$

$$t_{ik} = \left(\sum_{j=1}^{n} \left(\frac{d_{ik}^2}{d_{ij}^2} \right)^{2/(\eta-1)} \right)^{-1}, \forall i, k \quad (10)$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik}^m + t_{ik}^\eta) x_k}{\sum_{k=1}^{n} (u_{ik}^m + t_{ik}^\eta)}, \forall i \quad (11)$$

The FPCM algorithm consists of the following steps:

S1: Given a preselected number of clusters *c* and a chosen value for m, initialize the fuzzy partition matrix and typically the partition matrix with constraint in (7) and (8), respectively.

S2: Calculate the center of the fuzzy cluster, v_i for $i = 1,2, \ldots, c$ using Eq. (11).
S3: Use Eq. (9) to update the fuzzy membership u_{ik}.
S4: Use Eq. (10) to update the typically membership t_{ik}.
S5: If the improvement in $J_{m,\eta}(U,T,V)$ is less than a certain threshold (ϵ), then stop; otherwise, go to S2.

The main problem of FPCM algorithm is the constraint, which corresponds to the sum of all typical values of all data in the cluster - especially for a large data set (Pal et al. 2005).

2.3 Possibilistic fuzzy c-means (PFCM) algorithm

To obtain a stronger candidate for fuzzy clustering, Pal, Pal, Keller and Bezdek proposed PFCM algorithm in 2005. PFCM algorithm can avoid overlapping clusters and at the same time is less sensitive to outliers (Pal et al. 2005). PFCM algorithm uses a combination of PCM algorithm and FCM algorithm's objective functions. Object function of PFCM algorithm:

$$J_{m,\eta}(U,T,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta}) \times \left\|x_k - v_i\right\|^2 + \sum_{i=1}^{c}\delta_i \sum_{k=1}^{n}(1-t_{ik})^{\eta}, \quad (12)$$

Subject to

$$\sum_{i=1}^{c} u_{ik} = 1, \forall k \quad (13)$$

$$a > 0,\ b > 0, m > 1, \eta > 1, 0 \le u_{ik}, t_{ik} \le 1 \quad (14)$$

The relative significance between membership values and typicality values is determined by parameters a and b (Timm et al. 2004).

The objective function $J_{m,\eta}$ can be minimized $d_{ik} = \left\|x_k - v_i\right\|^2 > 0,$ for every i, k, m *and* $\eta > 1$ as well as X contains a minimum of c different data. The membership degree is updated with Eq. (15), the typicality values with Eq. (16) and the prototypes with Eq. (17).

$$u_{ik} = \left[\sum_{j=1}^{c}\frac{d_{ik}^{2}}{d_{jk}^{2}}^{\frac{1}{m-1}}\right]^{-1}, 1 \le i \le c; 1 \le k \le n \quad (15)$$

$$t_{ik} = \frac{1}{1 + \left(\frac{b\left\|x_k - v_i\right\|^2}{\delta_i}\right)^{1/(\eta-1)}}, \quad 1 \le i \le c; 1 \le k \le n \quad (16)$$

$$v_i = \sum_{k=1}^{n}\left(au_{ik}^{m} + bt_{ik}^{\eta}\right)x_k \Big/ \sum_{k=1}^{n}\left(au_{ik}^{m} + bt_{ik}^{\eta}\right), 1 \le i \le c \quad (17)$$

The basic steps of PFCM algorithm are given below:

S1: Initialize the number of clusters c, the partition matrix, such that $U^{(0)}$, the typicality matrix $T^{(0)}$, the termination tolerance $\varepsilon > 0$ and the user defined constants.
S2: Calculate the cluster prototypes using Eq. (17)
S3: Update the partition matrix by using Eq. (15)
S4: Update the typicality matrix using Eq. (16)
S5: Repeat from S2 until the improvement of objective function between two consecutive iterations is less than the termination tolerance ϵ.

2.4. Validation Index

The correctness of clustering algorithm results is verified using appropriate criteria and techniques. Since the scores obtained using the c-means family algorithms depend on the choice of c (the number of clusters), it is necessary to validate each result of the partitions once they are found. This validation is performed by a specific algorithm that allows assuming the appropriate value of the number c. We call this algorithm "validity index of the classification". It evaluates each class and determines the optimal or valid partition.

The main idea of the validity functions based on fuzzy partitioning: less fuzziness partitioning is more the performance is better (Saad and Alimi 2012).

During the last years, it has been proposed many validity indices. Fukuyama and Sugeno tried to model the cluster validation by exploiting the compactness and the separation. (Saad and Alimi 2012). This index is called Fukuyama-Sugeno index. Because of difficulties to write codes in Matlab for PCM, FCM, PFCM and FPCM algorithms together and insufficient obtained results for other validation indices for these algorithms to data sets in experimental analysis, we only use one validation index which is the new one after updating from Saad and Alimi. This index is called PI (Saad and Alimi 2016).

Performance Index (PI): PI updated from Fukuyama-Sugeno index is based on compactness and the separation (Saad and Alimi 2016). PI is criterion to choose a good clustering. Optimal clusters should minimize distance within clusters (intra cluster or cluster compactness) and maximize distance between clusters (inter cluster or cluster separation). The minimal value of index designates a "good" clustering in relation to others.

$$PI = J_m - K_m \tag{18}$$

where J_m is a compactness measure and K_m is a degree of separation between each cluster (v_i) and the mean ($\bar{v}$) of cluster centroids. For example for FCM algorithm, $J_m = J_{FCM}$ and

$$K_m = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\, d^2(v_i - \bar{v}) \tag{19}$$

It is clear that for compact and well-separated clusters one expects small values for PI. The first term in brackets measures the compactness of the clusters while the second one measures the distances of the clusters representative (Saad and Alimi 2016).

2.5 Criteria for comparing performances of FCM, PCM, FPCM and PFCM algorithms

FCM, PCM, FPCM and PFCM algorithms are compared in previous experiences using the following criteria:

Root Mean Squared Error (RMSE): The evaluation metric used by all algorithms of clustering is RMSE. RMSE is calculated by the root of the averaging all squared errors between the original data (X) and the corresponding predicted values data ($\bar{X}$).

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(x_{ik} - \bar{x}_{ik})}{n}} \tag{20}$$

where n is the total number of patterns in a given data set and c is the number of clusters; $x_{ik}\ and\ \bar{x}_{ik}$ the actual and predicted rating values data respectively.

Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE: \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_i| \tag{21}$$

where x_i and $\bar{x}_i$ the actual and predicted rating values data respectively.

Accuracy: Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions the model got right. Formally, accuracy has the following definition:

$$Accuracy = \frac{number\ of\ correct\ samples}{total\ number\ of\ samples} * 100 \tag{22}$$

3. Experimental Analysis

To show the feasibility of the methodology mentioned in this study, one performs some experiments to compare the performances of all algorithms with some numerical data sets. For application, E.coli, wine and seed data sets were taken (https://archive.ics.uci.edu/ml/index.php).

Wine data set: The wine data set contains the results of a chemical analysis of wines grown in a

specific area of Italy. Three types of wine are represented in the 178 samples, with the results of 13 chemical analyses recorded for each sample. The Type variable has been transformed into a categorical variable (Asuncion and Newman 2007).

Seed data set: Measurements of geometrical properties of kernels belongs to three different varieties of wheat. A soft X-ray technique and GRAINS package were used to construct all seven, real-valued attributes (Charytanowicz et al. 2010).

E.coli data set: The E.coli data set contains 336 numbers of instances and 7 attributes. This data set has been originally used to predict the cellular localization sites of E.coli proteins (Nakai and Kanehisa 1991).

The code used in this work was developed in MATLAB version R2015a. We wrote all the codes in MATLAB for FCM, PCM, FPCM and PFCM algorithms. FCM, PCM, FPCM and PFCM algorithms are implemented under the same initial values and stopping conditions. Fuzzifier parameter (m) and typicality parameter (η) is chosen 2 because it is used for the mathematical simplifications or by practice. Stopping criteria $\varepsilon = 1e-6$, number of iteration is equal to 100. The initialization typicality matrix T is performed in a random manner. PCM, FPCM and PFCM algorithms generally need an initial U matrix from a previous FCM run. So we run FCM algorithm firstly, then PCM, PFCM and FPCM algorithms run based on result of FCM' membership matrix. On the other hand, validity indices were calculated for the selection of the optimal set number of FCM, PCM, PFCM and FPCM algorithms. For the calculations, the codes were written in MATLAB. Other indices, except for the Performance index, failed to achieve results that were appropriate for PCM, PFCM and FPCM algorithms. Therefore, we continued with PI.

3.1 Obtained Results for Seed Data Set

In the first experiment, PI was calculated according to the number of clusters and shown in Table 1.

Table 1. PI generated by FCM, PCM, FPCM and PFCM algorithms for seed data set

Algorithm / c	PI FCM	PI PCM	PI FPCM	PI PFCM
2	-1.23e+03	-2.34e+04	-3.26e+04	-2.31e+04
3	*-1.28e+03*	-3.88e+04	-3.73e+04	-2.07e+04
4	-1.20e+03	-3.85e+04	-3.32e+04	-2.06e+04
5	-1.15e+03	-2.13e+04	5.95e+03	-2.05+04
6	-1.09e+03	NaN	5.84e+03	NaN
7	-1.13e+03	NaN	-5.05e+03	NaN
8	-1.12e+03	NaN	4.25e+03	NaN
9	-1.15e+03	NaN	4.89e+03	NaN
10	-1.12e+03	NaN	4.90e+03	NaN

As we seen in Table 1, optimal cluster number is equal to 3. FCM, PCM, FPCM and PFCM algorithms applied separately on the data set. The clusters are plotted and shown in Fig. 1, Fig. 2, Fig. 3 and Fig. 4 respectively. Because of showing similar results and being unnecessary for the other data sets called Wine and E.coli, these kind of figures are only shown for Seed data set for experimental analysis of this study.

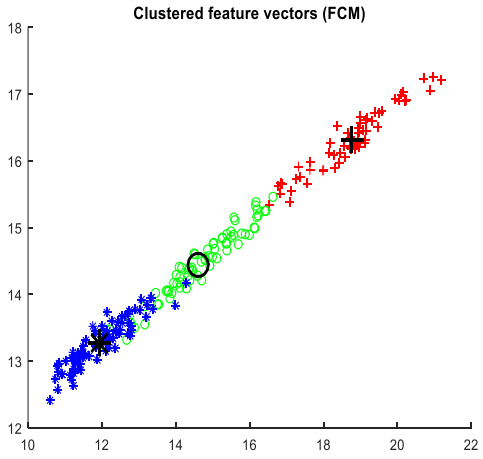

Fig. 1. FCM clusters for seed data set

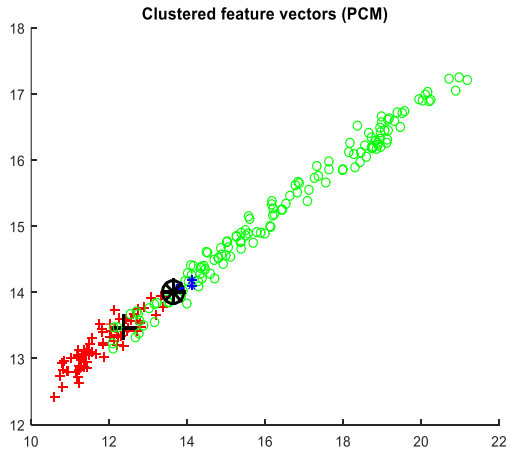

Fig. 2. PCM clusters for seed data set

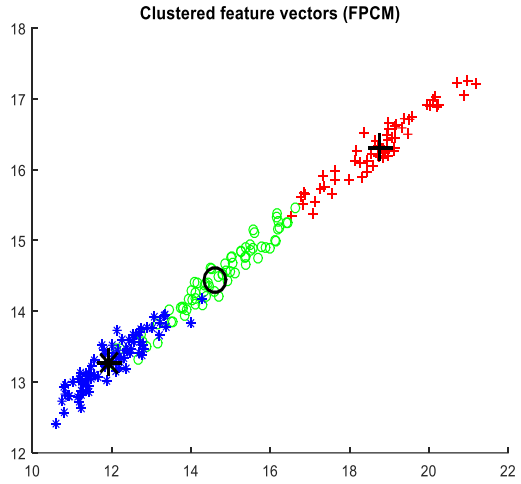

Fig. 3. FPCM clusters for seed data set

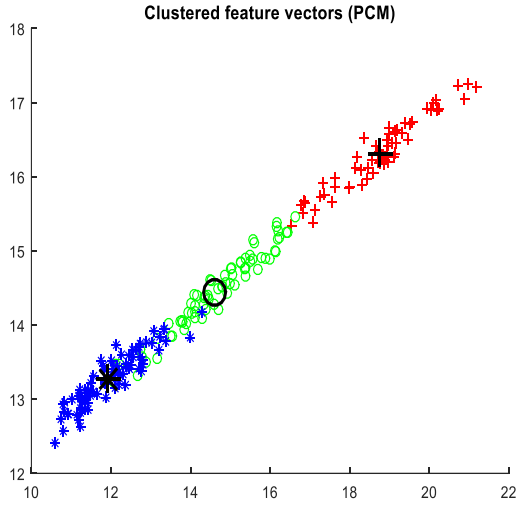

Fig. 4. PFCM clusters for seed data set

After FCM, PCM, FPCM and PFCM algorithms have been applied, obtained clusters and calculated cluster centroids, which are weighted mean (prototypes) of the data, are shown in Fig. 1, Fig. 2, Fig. 3 and Fig. 4. Similar clusters have been obtained and close centers (prototype) have been calculated from FCM, FPCM and PFCM algorithms, on the other hand, two cluster centers have been overlapped in PCM algorithm for all figures. It is clear that the quality of clustering performance from PCM is weak in Fig. 2.

The comparison of classification process is given in Table 2.

Table 2. Comparison of classification process for seed data set

Algorithm	Iteration	Accuracy	RMSE	MAE
FCM	43	% 89.52	1.672	18.037
PCM	334	% 45.71	4.368	43.613
FPCM	45	% 89.5	2.607	30.45
PFCM	46	% 89.59	0.167	2.204

As we seen in Table 2, PCM algorithm gives the worse result than FCM, PCM and FPCM algorithms according to RMSE, MAE, iteration and accuracy. FCM, FPCM and PFCM algorithms are produced approximately similar results.

3.2 Obtained Results for Wine Data Set

In the second experiment, PI was calculated according to the number of clusters and shown in Table 3.

Table 3. PI generated by FCM, PCM, FPCM and PFCM algorithms for wine data set

Algorithm / c	PI FCM	PI PCM	PI FPCM	PI PFCM
2	-1.03e+08	-1.05e+08	-1.03e+08	-1.01e+08
3	*-1.06e+08*	-1.07e+08	-1.06e+08	-1.07e+08
4	-1.06e+08	1.06e+08	-1.06e+08	-1.06e+08
5	-1.06e+08	-1.06e+08	-1.06e+08	-1.06+08
6	-1.05e+08	-1.05e+08	-1.05e+08	-1.05e+08
7	-1.05e+08	-1.04e+08	-1.06e+08	-1.04e+08
8	-1.04e+08	-1.04e+08	2.34e+06	1.23e+06
9	-1.04e+08	-1.04e+08	1.27e+06	1.25e+06
10	-1.04e+08	-1.05e+08	3.01e+05	1.26e+06

As we seen in Table 3, optimal cluster number is equal to 3; because when PI is minimal value, the classification is good. FCM, PCM, FPCM and PFCM algorithms applied separately on the data set.

The comparison of classification process is given in Table 4.

Table 4. Comparison of classification process for wine data set

Algorithm	Iteration	Accuracy	RMSE	MAE
FCM	99	% 67.98	154.9	1896.5
PCM	248	% 9.55	393.94	4436.4
FPCM	98	% 71.91	150.72	1627.9
PFCM	89	% 73.03	146.12	478.6

As we seen in Table 4, PFCM algorithm gives the better result than FCM, PCM and FPCM algorithms according to RMSE, MAE, iteration and accuracy.

3.3 Obtained Results for E.coli Data Set

In the third experiment, PI was calculated according to the number of clusters and shown in Table 5.

Table 5. PI generated by FCM, PCM, FPCM and PFCM algorithms for e.coli data set

Algorithm / c	PI FCM	PI PCM	PI FPCM	PI PFCM
2	-6.31e+03	6.88e+03	-6.3e+03	-6.01e+08
3	*-5.98e+03*	6.48e+03	-5.96e+03	-5.97e+08
4	-5.66e+03	-7.24e+03	-5.55e+03	-5.56e+08
5	-5.38e+03	-7.24e+03	-7.26e+03	-5.56+08
6	-7.27e+03	-7.24e+03	-7.27e+03	-7.25e+08
7	-7.26e+03	-7.24e+03	-7.27e+03	-7.04e+08
8	-7.26e+03	-7.24e+03	-7.27e+03	-7.23e+06
9	-7.26e+03	-7.24e+03	-7.26e+03	NaN
10	-7.26e+03	-7.24e+03	-6.10e+03	Nan

As we seen in Table 5, optimal cluster number is equal to 6 for FCM, FPCM and PFCM algorithms. In PCM algorithm, optimal cluster number is equal to 5. FCM, PCM, FPCM and PFCM algorithms applied separately on the data set.

The comparison of classification process is given in Table 6.

Table 6. Comparison of classification process for e.coli data set

Algorithm	Iteration	Accuracy	RMSE	MAE
FCM	102	% 67.86	12.71	54.93
PCM	49	% 40.26	21.62	84.78
FPCM	30	% 75. 6	0.29	2.58
PFCM	159	% 58.93	0.15	1.26

As we seen in Table 6, PCM algorithm gives the worse result than FCM, PCM and FPCM algorithms according to RMSE, MAE, iteration and accuracy. FCM, FPCM and PFCM algorithms are produced approximately similar results.

4. **Conclusion**

In this study, we applied FCM, PCM, FPCM and PFCM algorithms to cluster the data of wine, seed, and E.coli. For each algorithm, we prepared different m-files codes in MATLAB. All experiments were performed under the same circumstances. The first step of all algorithms is the step of selecting the parameters. In this step, initial parameters are chosen randomly. All algorithms have run several times in order to obtain good results for all applications.

Table 7. Comparison of all algorithms with PI according to all data sets

	n	di	c	nc	PI FCM	PI PCM	PI FPCM	PI PFCM
Seed	210	8	3	3	- *1.28e+03*	3.88e+04	3.73e+04	2.066e+04
Wine	178	13	3	3	- *1.061e+08*	1.067e+08	1.061e+08	1.066e+08
E.coli	336	8	7	5-6	- 7.269e+03	7.241e+03	7.27e+03	7.25e+08

n: Number of data
c: Number of original clusters
di: Number of data items
nc: Optimal number of clusters
PI FCM: Performance Index of FCM
PI PCM: Performance Index of PCM
PI FPCM: Performance Index of FPCM
PI PFCM: Performance Index of PFCM

PI was calculated to measure the quality of the clusters and to determine the optimal number of clusters. According to Table 7, wine data set was divided into 3 clusters for all algorithms; seed data set was separated to 3 clusters for all algorithms and E.coli data set was separated to 6 clusters for FCM, PFCM and FPCM algorithms. But, E.coli data set was divided into 5 clusters for PCM algorithm.

Also, the results of PFCM, FPCM, PCM and FCM algorithms were compared according to the classification accuracy, RMSE and MAE.

Table 8. Comparison of all algorithms with classification criteria according to all data sets

	Seed	Wine	E.coli
FCM			
Accuracy	% 89.52	% 67.98	% 67.86
RMSE	1.672	154.9	12.71
MAE	18.037	1896.5	54.93
PCM			
Accuracy	% 45.71	% 9.55	% 40.26
RMSE	4.368	393.94	21.62
MAE	43.613	4436.4	84.78
FPCM			
Accuracy	% 89.5	% 71.91	% 75. 6
RMSE	2.607	150.72	0.29
MAE	30.45	1627.9	2.58
PFCM			
Accuracy	% 89.59	% 73.03	% 58.93
RMSE	0.167	146.12	0.15
MAE	2.204	478.6	1.26

According to Table 8, results of PCM algorithm have been the worst performance in all algorithms. However; in wine, seed and E.coli data sets; FPCM and PFCM algorithms' performances have been similar. FPCM and PFCM algorithms gave the better results than FCM and PCM for all data sets. Because PFCM and FPCM algorithms have produced good clustering results by the influence of the typicality matrix.

5. References

Anderson, D. T., Zare, A. and Price, S., 2013. Comparing fuzzy, probabilistic and possibilistic partitions using the earth mover's distance. *IEEE Transactions on Fuzzy Systems*, **21**, 766-775.

Asuncion, A. and Newman, D. J., 2007. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

Berry, M. W., 2004. Survey of Text Mining. Springer-Verlag, New York, USA.

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum.

Bora, D. J. and Gupta, A. K., 2014. A comparative study between fuzzy clustering algorithm and hard clustering algorithm. *International Journal of Computer Trends and Technology*, **10**, 108-113.

Cebeci, Z. and Yildiz, F., 2015. Comparison of k-means and fuzzy c-means algorithms on different cluster structures. *Journal of Agricultural Informatics*, **6**, 13-23.

Cebeci, Z., Kavlak, A.T. and Yıldız, F., 2017. Validation of fuzzy and possibilistic clustering results. International Artificial Intelligence and Data Processing Symposium, IEEE.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Lukasik, S. and Zak, S., 2010. A complete gradient clustering algorithm for features analysis of x-ray images. Information Technologies in Biomedicine, Springer-Verlag, Berlin-Heidelberg.

Correa, C., Valero, C., Barreiro, P., Diago, M. P. and Tardaguila, J., 2011. A comparison of fuzzy clustering algorithms applied to feature extraction on vineyard. International Conference in Advances in Artificial Intelligence, 234-239.

Ganbold, G. and Chasia, S., 2017. Comparison between possibilistic c-means and artificial neural network classification algorithms in land use/ land cover classification. *International Journal of Knowledge Content Development and Technology*, **7**, 57-78.

Jafar, M. O. A. and Sivakumar, R., 2012. A study on possibilistic and fuzzy possibilistic c- means clustering algorithms for data clustering. International Conference on Emerging Trends in Science, Engineering and technology, 90-95.

Krishnapuram, R. and Keller, J., 1993. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, **1**, 98-110.

Nakai, K. and Kanehisa, M., 1991. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Genetics*, **11**, 95-110.

Nefti, S. and Oussalah, M., 2004. Probabilistic-fuzzy clustering algorithm. IEEE international Conference on Systems, Man and Cybernetics, 4786-4791.

Ozdemir, O. and Kaya, A., 2018. Effect of parameter selection on fuzzy clustering. *Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi*, **2**, 22-33.

Ozdemir, O. and Kaya, A., 2018. K-medoids and fuzzy c-means algorithms for clustering CO2 emissions of

Turkey and other OECD countries. *Applied Ecology and Environmental Research*, **16**, 2513-2526.

Pal, N. R., Pal, K. and Bezdek, J. C., 1997. A mixed c-means clustering model. IEEE International Conference Fuzzy Systems, 11 -21.

Pal, N. R., Pal, K., Keller, J. M. and Bezdek, J. C., 2005. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, **13**, 517-530.

Saad, M. F. and Alimi, A.M., 2012. Validity index and number of cluster. *International Journal of Computer Science Issues*, **9**, 52-56.

Saad, M. F. and Alimi, A.M., 2016. Selecting parameters of the fuzzy possibilistic clustering algorithm. *Communications on Applied Electronics*, **5**, 42-52.

Singhal, R. and Deepika, N., 2016. Classification of words: using PFCM clustering. *International Journal of Computer Science and Mobile Computing*, **5**, 114-117.

Şahinli, F., 1999. Kümeleme analizine fuzzy set teorisi yaklaşımı. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 119.

Şanlı, K. and Apaydın, A., 2006. Robust kümeleme yöntemleri. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, **7**, 33-39.

Timm, H, Borgelt, C., Doring, C. and Kruse, R., 2004. An extension to possibilistic fuzzy cluster analysis. *Fuzzy Sets and Systems*, **147**, 3-16.

Zadeh, L., 1965. Fuzzy sets. *Information and Control*, **8**, 338-353.

Internet resources

https://archive.ics.uci.edu/ml/index.php (07.03.2019)