Comparison of Different Fuzzy Clustering Algorithms: A Replicated Case Study

Tusharika Singh and Anjana Gosain

Abstract Fuzzy clustering partitions data points of a dataset into clusters in which one data point can belong to more than one cluster. In the literature, a number of fuzzy clustering algorithms have been proposed. This paper reviews various fuzzy clustering algorithms such as Fuzzy C-Means (FCM), Possibilistic C-Means (PCM), Possibilistic Fuzzy C-Means (PFCM), Intuitionistic Fuzzy C-Means (IFCM), Kernel Fuzzy C-Means (KFCM), and Density-Oriented Fuzzy C-Means (DOFCM). We have demonstrated the experimental performance of these algorithms on some standard and synthetic datasets which include—Bensaid, Square (DUNN), D15, and D45 dataset. Then, the results are analyzed and compared to see the effectiveness of these algorithms in presence of noise and outliers.

Keywords Fuzzy clustering · FCM · PCM · PFCM · IFCM
KFCM · DOFCM · Outliers

1 Introduction

The concept of separating data elements into groups or clusters in a way that elements in the same cluster are homogeneous and elements held by different clusters are disparate is called as clustering [1]. Clustering can be either hard clustering or fuzzy clustering (soft clustering). When data points of a dataset are divided into different clusters and each data point can belong to only one cluster, then this type of clustering is called hard clustering, whereas fuzzy clustering allows each data point to belong to multiple clusters [2]. Fuzzy clustering uses the concept

T. Singh (✉) · A. Gosain
University School of Information and Communication Technology,
Guru Gobind Singh Indraprastha University, Dwarka 110078, Delhi, India
e-mail: tusharikasingh170@gmail.com

A. Gosain
e-mail: anjana_gosain@hotmail.com

S. C. Satapathy et al. (eds.), *Information and Decision Sciences*,
Advances in Intelligent Systems and Computing 701,
https://doi.org/10.1007/978-981-10-7563-6_28

of fuzzy logic where a membership grade is associated with each data point, between a range of 0 and 1 [3].

Mostly used algorithm for fuzzy clustering was proposed by Bezdek [4], called Fuzzy C-Means (FCM) [5]. FCM algorithm works well on most noise-free data but fails to detect data and segment images corrupted by noise and outliers [5]. Algorithms like Possibilistic C-Means (PCM) and Possibilistic Fuzzy C-Means (PFCM) perform better in presence of noise as compared to FCM but PCM fails to find global optimal cluster and leads to the generation of coincident clusters. PFCM lacks to give accurate results for datasets consisting of two clusters which are highly unequal in size with outliers given in it [6].

FCM, PCM, and PFCM are not effective in clustering nonspherical clusters [7]. Hence, a new algorithm, KFCM (Kernel Fuzzy C-Means) was proposed to deal with these types of nonspherical data. When the data sets contain one or more very large outliers, KFCM is more desirable than FCM and PCM [7].

In order to improve accuracy and effectiveness of clustering algorithms, a new algorithm called Intuitionistic Fuzzy C-Means (IFCM) was proposed by Chaira [8]. IFCM uses intuitionistic fuzzy set theory and converges to a more desirable location as compared to the cluster centers obtained using FCM, PCM, PFCM, and KFCM. IFCM works well for hyper-spherical data but is not suitable to cluster nonlinearly separable data. Another algorithm DOFCM [9] was proposed. Density-Oriented Fuzzy C-Means (DOFCM) is a robust technique which uses density of dataset to identify outliers before applying clustering.

In this paper, we have analyzed and compared the performance of these algorithms to see their effectiveness in presence of noise and outliers.

Paper is organized as follows: Sect. 2 briefly discusses various algorithms used in our work. Section 3 presents experimental result of datasets with respect to algorithms used in the form of figures and tables. Section 4 concludes this paper with a short summary.

2 Literature Review

2.1 Fuzzy C-Means (FCM)

FCM fragments a set of data points into a number of clusters "c", which are assumed to be known for a dataset and minimizes the objective function as expressed in Eq. (1) [5]:

$$J_{FCM} = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^{m} d_{ik}^{2} \tag{1}$$

With respect to membership function u_{ik} of a data point x_i in cluster k. d_{ik} is the Euclidean distance between data point, x_i and cluster center, v_k. "m" is fuzzifier.

2.2 Possibilistic C-Means (PCM)

PCM proposed by Krishnapuram and Keller [10] overcomes the FCM's problem of noise points which are equidistant from two clusters. Equation (2) shows objective function:

$$J_{PCM} = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^{m} d_{ik}^{2} + \sum_{k=1}^{c} \sigma_k \sum_{i=1}^{n} (1 - u_{ik})^{m} \tag{2}$$

σ_k are suitable positive numbers.

2.3 Possibilistic Fuzzy C-Means (PFCM)

The fuzzy approach of FCM and possibilistic approach of PCM was combined by Pal et al. [11]. Hence, it has two types of memberships, i.e., a fuzzy membership (u_{ik}) and possibilistic membership (t_{ki}). Objective function is given in Eq. (3):

$$J_{PFCM} = \sum_{k=1}^{c} \sum_{i=1}^{n} \left(a u_{ik}^{m} + b t_{ki}^{n}\right) d_{ik}^{2} + \sum_{k=1}^{c} \Upsilon_k \sum_{i=1}^{n} (1 - t_{ki})^{n} \tag{3}$$

2.4 Intuitionistic Fuzzy C-Means (IFCM)

IFCM uses Intuitionistic Fuzzy Set (IFS) theory. IFS theory considers both membership and nonmembership functions [12]. This algorithm merged the hesitation degree (an uncertainty factor) with membership degree as shown in Eq. (4):

$$J_{IFCM} = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^{*m} d_{ik}^{2} + \sum_{k=1}^{c} \eta_k^{*} e^{1-\eta_k^{*}}, \tag{4}$$

where u_{ik}^{*} denotes the intuitionistic fuzzy membership of the ith data in kth cluster.

2.5 Kernel Fuzzy C-Means (KFCM)

KFCM uses a new kernel-induced metric in the data space, instead of conventional Euclidean norm metric in FCM, in order to deal with high-dimensional data set. By swapping the conventional distance measure with a suitable "kernel" function, without increasing the number of parameters, a nonlinear mapping can be

performed to a high-dimensional feature space [7]. Objective function is determined by Eq. (5):

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^{m} \|\Phi(x_i) - \Phi(v_k)\|^2, \tag{5}$$

where $\|\Phi(x_i) - \Phi(v_k)\|^2 = K(x_i, x_i) + K(v_k, v_k) - 2K(x_i, v_k)$.

2.6 Density-Oriented Fuzzy C-Means (DOFCM)

DOFCM separates noise into different clusters, i.e., it finds "n" noiseless clusters and one cluster which is not valid consisting of all the outliers present in a dataset, resulting in total of "$n + 1$" clusters [13]. Neighborhood membership is defined as in Eq. (6):

$$M_{neighborhood}^{i}(X) = \frac{\eta_{neighborhood}^{i}}{\eta_{\max}} \tag{6}$$

3 Result and Simulation

We have implemented FCM, PCM, PFCM, IFCM, KFCM, and DOFCM in MATLAB Version 7.10 on core i3 processor, 1.70 GHz with 4 GB RAM. For all datasets, we have considered m = 2, $\varepsilon = 0.0001$ and maximum number of iterations as 100.

3.1 Bensaid Dataset

Dataset: Bensaid [14]
Number of clusters: 3
Number of data points in respective clusters: 6, 25, 16
Number of outliers: 2

Figure 1 depicts the clustering result of discussed algorithms. Symbols "x", "o" and ">" represent the three clusters. Centroids of three clusters are plotted by "*" and outliers are plotted using symbol "o" in blue. We examined that FCM is slightly affected with the presence of outliers. PCM fails to give appropriate result due to unequal sized clusters in the dataset and hence only two clusters are obtained. PFCM results into two overlapping clusters with three centroids.

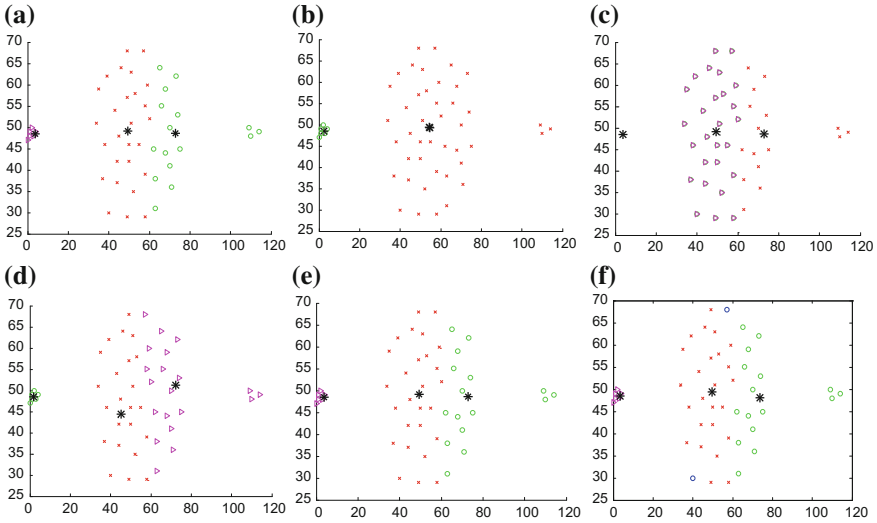

Fig. 1 Clustering result of **a** FCM, **b** PCM, **c** PFCM, **d** IFCM, **e** KFCM, **f** DOFCM on Bensaid dataset

Similar to FCM, presence of noise and outliers affects the performance of IFCM and KFCM. IFCM could not detect original clusters whereas KFCM detects clusters but results in faulty centroid locations. DOFCM detects outliers, revealing original clusters.

3.2 Square Dataset

Dataset: Square [15]
Number of clusters: 2
Number of data points in respective clusters: 53, 81
Number of outliers: 21

Figure 2 depicts the clustering result of discussed algorithms. Symbols "x" and "o" represent the two clusters. Centroids of two clusters are plotted by "*" and outliers are plotted using symbol "o" in blue. We examined that outliers affect the performance of FCM, PCM, PFCM, and IFCM. However, unlike Bensaid dataset, PCM and PFCM are able to detect actual number of clusters. KFCM obtains centroid of the clusters but fails to find outliers. DOFCM results in original clusters as it distinguishes outliers from the cluster's data points.

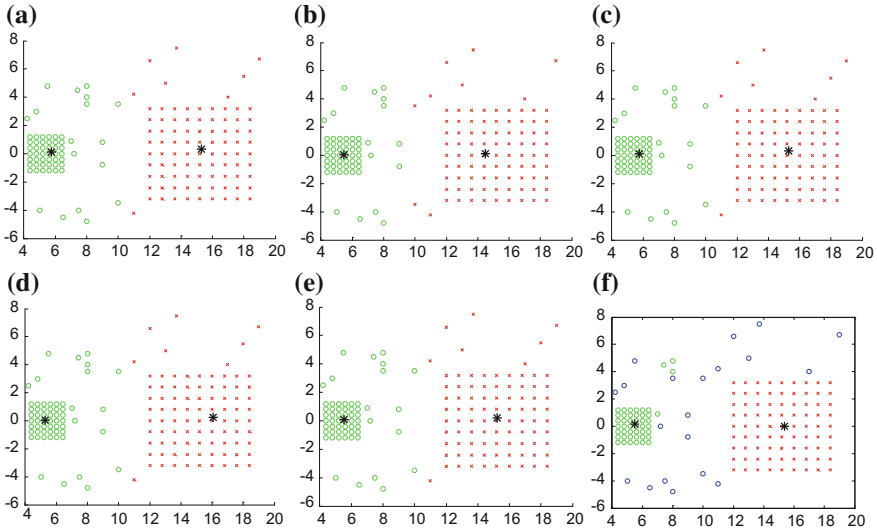

Fig. 2 Clustering result of **a** FCM, **b** PCM, **c** PFCM, **d** IFCM, **e** KFCM, **f** DOFCM on Square dataset

3.3 D15 Dataset

Dataset: D15 [15]
Number of clusters: 2
Number of data points in respective clusters: 6, 5
Number of outliers: 4

Figure 3 depicts the clustering result of discussed algorithms. We examined and observed that PCM cannot find appropriate number of clusters due to its unequal size (11 and 4 data points in two clusters) and hence provide only one cluster. FCM and PFCM produce centroids which are more attracted toward the outliers. IFCM could not detect original clusters and its performance is badly affected by noise. KFCM and DOFCM give centroid locations which preclude attraction of centroids toward outliers but compared to KFCM, DOFCM gives more accurate result and detects outliers, which KFCM lacks.

3.4 D45 Dataset

Dataset: D45 [16]
Number of clusters: 2

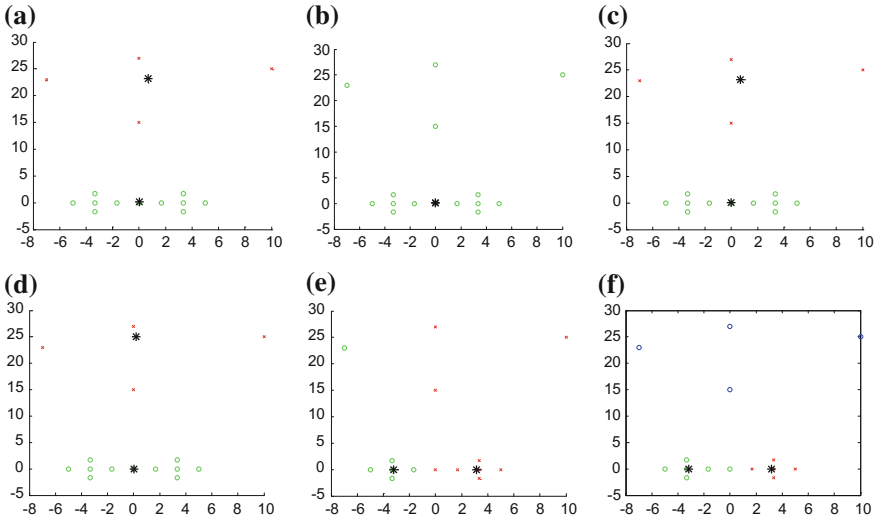

Fig. 3 Clustering result of **a** FCM, **b** PCM, **c** PFCM, **d** IFCM, **e** KFCM, **f** DOFCM on D15 dataset

Number of data points in respective clusters: 18, 18
Number of outliers: 9

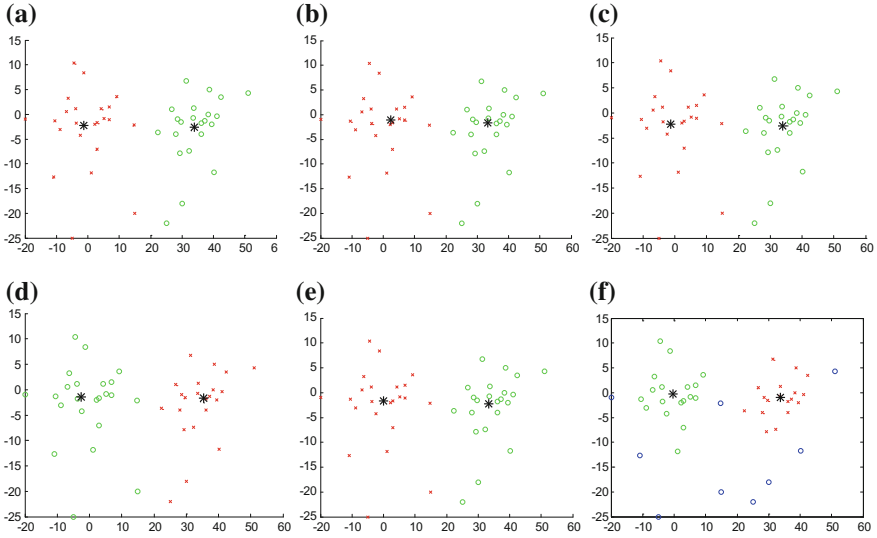

Fig. 4 Clustering result of **a** FCM, **b** PCM, **c** PFCM, **d** IFCM, **e** KFCM, **f** DOFCM on D45 dataset

Table 1 Centroid coordinates produced on standard and synthetic datasets

Dataset	Algorithms					
BENSAID	*FCM*		*PCM*		*PFCM*	
	49.3104	49.0970	54.5093	49.2560	49.3100	49.0910
	3.6201	48.4843	2.4704	48.5296	72.9868	48.5630
	72.9866	48.5565	54.3423	49.4217	3.6202	48.4851
	IFCM		*KFCM*		*DOFCM*	
	1.8103	48.5128	4.3509	48.4742	73.7985	48.0879
	45.2492	44.4176	49.7620	49.1125	49.5533	49.4391
	72.1888	51.2069	73.2312	48.5238	3.3955	48.5460
SQUARE	*FCM*		*PCM*		*PFCM*	
	15.3149	0.3322	14.5058	0.1195	15.3093	0.3286
	5.7652	0.1165	5.4570	0.0456	5.7667	0.1186
	IFCM		*KFCM*		*DOFCM*	
	5.2717	0.0418	15.2015	0.1724	5.4870	0.1719
	16.0559	0.2383	5.4841	0.0784	15.3848	0.0086
D15	*FCM*		*PCM*		*PFCM*	
	0.6757	23.1738	0.0033	0.0222	0.0040	0.1050
	0.0047	0.1227	−0.0016	0.1454	0.6756	23.1720
	IFCM		*KFCM*		*DOFCM*	
	0.0123	−0.0141	−3.2205	0.0033	−3.1672	0.0000
	0.1796	25.0757	3.1256	0.0026	3.1675	0.0000
D45	*FCM*		*PCM*		*PFCM*	
	−1.2111	−2.2085	2.3730	−1.1031	−1.1982	−2.2039
	33.8600	−2.6449	33.4284	−1.6086	33.8562	−2.6481
	IFCM		*KFCM*		*DOFCM*	
	−2.5559	−1.4287	33.3498	−2.2699	−0.4359	−0.3126
	35.3518	−1.6149	0.0353	−1.7038	33.8011	−0.9544

Figure 4 depicts the clustering result of discussed algorithms. D45 is a synthetic dataset consisting of two clusters which are represented by symbols “x” and “o”. Centroids are plotted by “*” and outliers are plotted using symbol “o” in blue. FCM, PCM, PFCM, and IFCM performance is highly affected by outliers. KFCM detects centroids which are more attracted toward data points of the actual clusters instead of outliers. DOFCM provides original clusters, excluding the outliers from consideration. Table 1 shows the centroid coordinates produced by each algorithm on various datasets.

4 Conclusion

In this paper, we have analyzed some of the fuzzy clustering algorithms on standard and synthetic datasets considering noise and outliers in the datasets. We observed that FCM does not perform well in presence of noise and outliers whereas performance of PCM and PFCM improves over FCM but not significantly. IFCM could not detect original clusters and KFCM exhibits attraction of centroid toward outlier so the clustering accuracy is affected. DOFCM, compared to other algorithms, gives foremost cluster centroid location as it first detects outliers and then applies clustering technique on outlier-free clusters. In future, we will try to come up with an algorithm which optimizes existing clustering results.

References

1. Hung, C.C., Kulkarni, S., Kuo, B.: A new weighted fuzzy c-means clustering algorithm for remotely sensed image classification. IEEE J. Sel. Top. Signal Process. **5**(3), 543–553 (2011)
2. Grover, N.: A study of various fuzzy clustering algorithms. Int. J. Eng. Res. (IJER) **3**(3), 177–181 (2014)
3. Gosain, A., Dahiya, S.: Performance analysis of various fuzzy clustering algorithms: a review. Proc. Comput. Sci. 100–111 (2016)
4. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum (1981)
5. Gong, M., et al.: Fuzzy c-means clustering with local information and kernel metric for image segmentation. IEEE Trans. Image Process. **22**(2), 573–584 (2013)
6. Sharma, S., Goel, M., Kaur, P.: Performance comparison of various robust data clustering algorithms. Int. J. Intell. Syst. Appl. **5**(7), 63–71 (2013)
7. Zhang, D., Chen, S.C.: Kernel-based fuzzy and possibilistic c-means clustering. In: Proceedings of the International Conference Artificial Neural Network (2003)
8. Chaira, T.: A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images. Appl. Soft Comput. **11**, 1711–1717 (2011)
9. Kaur, P., Lamba, I.M.S., Gosain, A.: DOFCM: a robust clustering technique based upon density. IACSIT Int. J. Eng. Technol. **3**(3), 297–303 (2011)
10. Krishnapuram, R., Keller, J.M.: The possibilistic c-means algorithm: insights and recommendations. IEEE Trans. Fuzzy Syst. **4**(3), 385–393 (1996)
11. Pal, N.R., et al.: A possibilistic fuzzy c-means clustering algorithm. IEEE Trans. Fuzzy Syst. **13**(4), 517–530 (2005)
12. Kaur, P., et al.: Novel intuitionistic fuzzy C-means clustering for linearly and nonlinearly separable data. WSEAS Trans. Comput. **11**(3), 65–76 (2012)
13. Gosain, A., Singh, T.: DKFCM: kernelized approach to density-oriented clustering. In: Accepted in 4th International Conference on Computational Intelligence in Data Mining (ICCIDM-2017), Odisha, India, Nov 2017
14. Bensaid, A.M., et al.: Validity-guided (re) clustering with applications to image segmentation. IEEE Trans. Fuzzy Syst. **4**(2), 112–123 (1996)
15. Kaur, P., Soni, A.K., Gosain, A.: Robust kernelized approach to clustering by incorporating new distance measure. Eng. Appl. Artif. Intell. **26**(2), 833–847 (2013)
16. Rehm, F., Klawonn, F., Kruse, R.: A novel approach to noise clustering for outlier detection. Soft. Comput. **11**(5), 489–494 (2007)