




Improving the performance of the FCM algorithm in clustering using the DBSCAN algorithm[†]

S. Barkhordari Firozabadi, S.A. Shahzadeh Fazeli^{*,}, J. Zarepour Ahmadabadi and S.M. Karbassi

Abstract

The fuzzy-C-means (FCM) algorithm is one of the most famous fuzzy clustering algorithms, but it gets stuck in local optima. In addition, this algorithm requires the number of clusters. Also, the density-based spatial of the application with noise (DBSCAN) algorithm, which is a density-based clustering algorithm, unlike the FCM algorithm, should not be pre-numbered. If the clusters are specific and depend on the number of clusters, then it can determine the number of clusters. Another advantage of the DBSCAN clustering algorithm over FCM is its ability to cluster data of different shapes. In this paper, in order to overcome these limitations, a hybrid approach for clustering is proposed, which uses FCM and DBSCAN algorithms. In

*Corresponding author

Received 11 May 2023; revised 9 July 2023; accepted 27 July 2023

Saeideh Barkhordari Firozabadi

PhD candidate, Department of Computer Science, Yazd University, Yazd, Iran.
e-mail: s.barkhordari@stu.yazd.ac.ir

Seyed Abolfazl Shahzadeh Fazeli

Parallel Processing Lab, Department of Computer Science, Yazd University, Yazd, Iran.
e-mail: fazeli@yazd.ac.ir

Jamal Zarepour Ahmadabadi

Department of Computer Science, Yazd University, Yazd, Iran.
e-mail: zarepourjamal@yazd.ac.ir

Seyed Mehdi Karbassi

Department of Applied Mathematics, Faculty of Mathematical Sciences, Yazd University, Yazd, Iran.
e-mail: smkarbassi@yazd.ac.ir

[†] This article was suggested by the scientific committee of the 5th national seminar on control and optimization for publication in IJNAO, which was accepted after independent review.

this method, the optimal number of clusters and the optimal location for the centers of the clusters are determined based on the changes that take place according to the data set in three phases by predicting the possibility of the problems stated in the FCM algorithm. With this improvement, the values of none of the initial parameters of the FCM algorithm are random, and in the first phase, it has been tried to replace these random values to the optimal in the FCM algorithm, which has a significant effect on the convergence of the algorithm because it helps to reduce iterations. The proposed method has been examined on the Iris flower and compared the results with basic FCM algorithm and another algorithm. Results shows the better performance of the proposed method.

AMS subject classifications (2020): 68T10, 62H30.

Keywords: Clustering; Fuzzy clustering; DBSCAN.

1 Introduction

Clustering is one of the important techniques of knowledge discovery in databases. Density-based clustering algorithms are one of the main methods for clustering in data mining. The density-based spatial of application with noise (DBSCAN) algorithm is a clustering method that is based on density. This algorithm has the ability to discover clusters of different sizes and shapes from a large amount of data and performs well against noise [3, 6]. Another method that has received a lot of attention is the fuzzy method. In these methods, unlike deterministic clustering that, any data belongs to exactly one cluster; the data can belong to several clusters [7]. Although the approach adopted by both algorithms is widely accepted to deal with clustering problems, due to the weakness in each and in order to achieve a better method for data clustering, various methods have been used. In all methods, it has been tried to find values that are as close as possible to an exact answer.

2 Related works

Wei and Xie [10], after a better analysis of the slower convergence speed, introduced a new competitive learning-based rival checked fuzzy c-means clustering algorithm. In the method proposed by Xue, Shang, and Feng [12], a fuzzy rough semi-supervised outlier detection is used, which is able to minimize the sum squared errors of the clustering. Maraziotis [4], for gene expression profile clustering, proposed a novel semi-supervised fuzzy clustering algorithm (SSFCA). Abdellahoum et al. [1] presented a new version of fuzzy clustering based on the ABC algorithm, namely ABC – SFCM. For detecting the malicious behavior in wireless sensor networks, Shamshirband

et al. [8] presented a hybrid clustering method, namely a density-based fuzzy imperialist competitive clustering algorithm. Mekhmoukh and Mokrani [5] introduced an improved fuzzy-C-means (FCM) using particle swarm optimization based on the outlier rejection and level set. The results of this method were compared with related works, which showed more effectiveness. Zhang et al. [13] proposed a variant of FCM for image segmentation, which has reduced the complexity of the algorithm compared to similar types. Alomoush et al. [2] proposed a method for choosing cluster centers to avoid getting stuck in local optimum.

3 Preliminaries and definitions

Here we present some necessary algorithms.

3.1 Clustering

Clustering is a process by which a set of objects can be separated into distinct groups. Each release is called a cluster. Members of each cluster, according to characteristics which, are very similar to each other, but instead, the degree of similarity between the clusters is the lowest [9]. Although most clustering algorithms or methods have the same basis, there are differences in the method of measuring similarity or distance, as well as choosing labels for objects in each cluster. There are methods: for example, discriminative clustering, hierarchical clustering, model-based clustering, fuzzy clustering and density-based clustering. Here We deal with the last two methods: fuzzy clustering and density-based clustering. By combining these methods, we try to provide a new method for clustering.

3.2 FCM clustering algorithm

As mentioned in fuzzy clustering, unlike classical clustering, where each input sample belongs to only one cluster, one sample can belong to more than one cluster. Actually the basic idea in fuzzy clustering is to assume that each element can be placed in several clusters with different degrees of membership [7]. As a result, we can have clusters that are more consistent with reality.

One of the basic fuzzy clustering algorithms is FCM. In the FCM algorithm, we try to optimize the following objective function [11]:

$$J_m = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m d_{ik}^2 = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|x_k - v_i\|^2,$$

where m is a real number greater than one. Moreover, u_{ik} is the degree of membership of the k th data in the i th cluster, d_{ik} is the measure of similarity in the next n space, x_k represents the k th data, and v_i is the center of the i th cluster. The complete procedure of the algorithm is as follows:

Algorithm 1 FCM algorithm

Input: Data set, number of clusters, $max - iter$, $threshold$ (minimum objective function improvement value), and m (the value for exponentiation of matrix U)

Output: Cluster centers, objective function values and matrix U

1. Initialization: Randomly determine the value of each data belonging to the desired cluster, put it in the matrix U , set the value of the $iter$, and the value of the objective function to zero.
 2. Calculate new centers for each cluster.
 3. Calculate the distance from the data to the cluster centers.
 4. Calculate the value of the objective function in terms of distance values.
 5. Calculate the matrix U in terms of the values obtained from the previous steps.
 6. Calculate imp (the difference between the value of the objective function in the new step and in the previous step) and set $iter = iter + 1$, if $imp \geq threshold$ and $iter \leq max - iter$, then repeat step 2; otherwise, the algorithm terminates.
-

As mentioned, FCM clustering performs well when working with overlapping data and performs well with noise-free data. Since they cannot distinguish between data points and noise, it leads to the center, which may gravitate toward the outliers. Also, it may be located at a local optimum. To improve the algorithm, we use the DBSCAN algorithm. In which follows, the DBSCAN algorithm is presented.

3.3 DBSCAN algorithm

In density-based clustering algorithms, points with high density are identified and placed in a cluster. One of the famous algorithms cited in this field to DBSCAN, which was presented by Ester and colleagues in 1996. This algorithm has the ability to identify remote points [3]. In the DBSCAN algorithm, there are two parameters, the radius (Epsilon) and the minimum points in a cluster (MinPoints). Each data point has a distance from other

points. Any point whose distance to an assumed point is less than Epsilon is considered a neighbor of that point. Any given point that has *MinPoints* of neighbors is the center of the cluster.

The way that the algorithm works is that the algorithm first selects a sample (which is a point in the vector space) and according to the radius Epsilon, the neighbor looks for this point in space. If the algorithm is able to find at least as many points as *MinPoints* within the specified Epsilon radius, then all those points together belong to a cluster. The algorithm then looks for one of the points adjacent to the current point to look again at that point with the Epsilon radius. The other neighbor points are searched, and if the number of serious new neighbor points is found again, then this algorithm again places all those new points in the same cluster with the previous points. If it does not find a new point in the neighborhood, then this cluster is complete. To find other clusters at other points, it randomly selects another point and starts finding neighbors and forming a new cluster for that point. If the algorithm is within the desired radius of a point but does not find enough samples, then the DBSCAN algorithm identifies this point as outlier data and does not assign it to any cluster. It should make all the clusters and check all the points to be able to identify whether it is an outlier or not. The algorithm continues in the same way to find other clusters that have at least as much as *MinPoints* in their radius and are clustered. Finally, those that are not assigned to any cluster are identified as outlier data. This continues until all points have been checked [3, 6].

4 Proposed method

To improve the fuzzy clustering method, changes are made in three phases:

1. In the FCM clustering method, as stated, the value of each data belonging to the desired cluster is randomly determined and placed in the matrix U . In the proposed method, first, the data set is clustered through the DBSCAN algorithm. Since the number of clusters must be given to the FCM algorithm as input, the initial cluster number is determined in this way. Then, the distance between each data to the centers of the clusters obtained from the DBSCAN algorithm is calculated. In the next step, these distances are reversed and normalized. To help improve the convergence of the algorithm, the above values are placed in the matrix U to determine the value belonging of each data to the clusters. The points that are closer to the cluster centers get more value. As a result, better convergence is achieved, and the number of iterations also becomes less.
2. Similar to the idea of the simulated annealing method, changes are applied to the number of clusters. In this way, in the range of $+k/2$ and $-k/2$, a value is randomly selected and added to the number of

clusters. If the number of clusters increases, then centers are randomly selected from the data, and if the number of clusters decreases, then some centers are randomly deleted. In the event that the objective function is improved by changing the number of clusters, results will be updated.

3. The cluster centers are moved to find the optimal centers. In the proposed method, the primary centers are obtained with the DBSCAN algorithm. Considering the criteria of the DBSCAN algorithm, in data density clustering, several data may be close to each other, but according to the changes in the value of data dimensions from the first to the last data, it is more appropriate that this data should not be placed in a cluster. For this reason, in the second and third phases of the proposed algorithm, the number and location of the cluster centers are changed to reach the optimal centers. These changes increase in the first iteration and decrease in subsequent iterations. The process is as follows: in the first iteration, based on the data diameter and the angle that is randomly determined, the transfer value is determined. In the next iterations, a coefficient from the diameter of the data determines that the amount of displacement is based on this coefficient takes place, and this displacement will be reduced. At each stage, based on the new centers, the matrix U and the objective function are calculated, and if improved, results will be replaced.

In the proposed method, different aspects of clustering and different ways of improving these methods were studied and investigated. Then, according to the weaknesses of the FCM algorithm, based on the changes made in three phases in the proposed method, the algorithm was improved with new methods from three points of view. In each point of view, different aspects of clustering are considered:

1. In the first phase of improvement, combine the algorithm with DBSCAN algorithm. In addition to solving the basic problem of the algorithm in determining the number of clusters, it is tried to make the initialization of the matrix U in a completely intelligent and accurate way by making changes. Because by conducting tests, we found that the initial values have a significant effect on the convergence and accuracy of the algorithm result, and if this value is done with the random method used in the FCM algorithm, then the number of repetitions will increase.
2. In the second phase, it was tried to find the optimal value for the number of clusters with a creative method. In the FCM algorithm, due to the unknown number of clusters, this value should be given as an initial parameter to the algorithm.
3. In this method, the initial value for the location of the centers is done according to the criteria of the DBSCAN algorithm. According to this

fact, in the third phase of the proposed method, it is tried to find the best place for the centers of the clusters with a new method, which has a significant effect on reaching the optimal solution.

The proposed method is summarized in Figure 1. Here, IMP is the improvement value of the objective function, NC is the number of clusters, and C is the centers of the clusters.

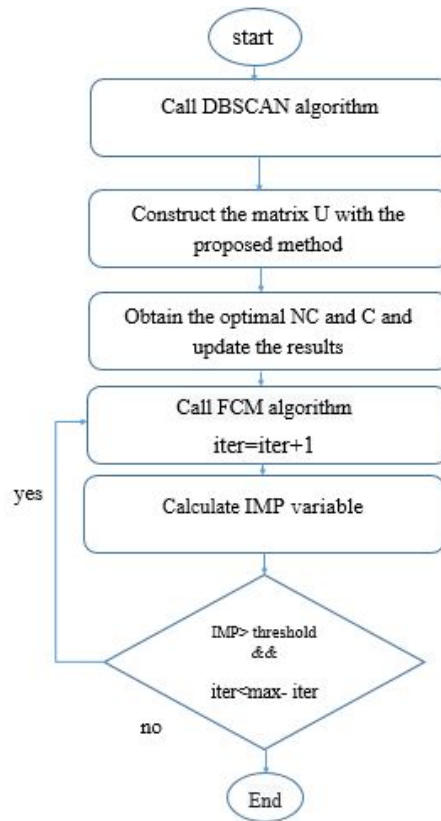


Figure 1: The proposed method

The general routine of the algorithm is given in Algorithm 2 as follows:

Algorithm 2 Proposed algorithm MODFCM

Input: Data set, number of clusters, $max - iter$, threshold (minimum objective function improvement value), and m (the value for exponentiation of matrix U)

Output: Cluster centers, objective function values, and matrix U

1. Initialization:
 - (a) Call the DBSCAN algorithm and determine the number of clusters in the data set.
 - (b) Calculate the distance of each data to the centers obtained from the DBSCAN algorithm and construct the matrix U with the proposed method.
 - (c) Set $iter$ and the value of the objective function to zero.
 2. Obtain the optimal number of clusters using the second phase of the proposed method, update the new results and go to the next step.
 3. Initialize f by calculating the diameter of the data set.
 4. Initialize s by randomly choosing an angle.
 5. Set $d = f * \cos(s)$ and displace all centers by d .
 6. Calculate the distance of the data to the centers of the new clusters and the value of the objective function in terms of the distance values.
 7. Calculate the values of the matrix U according to the values obtained from the previous steps.
 8. If the objective function is improved, then update new results and go to the next step; otherwise, go to step 10.
 9. Set $f = .9f$. If $f \geq 5$, then go to step 4; otherwise, go to the next step.
 10. Calculate the new centers for each cluster, the distance of the data to the centers of the clusters, and the value of the objective function in terms of the distance values.
 11. Calculate the values of the matrix U according to the values obtained from the previous steps.
 12. Calculate imp (the difference between the value of the objective function in the new step and in the previous step), and set $iter = iter + 1$. If $imp \geq threshold$ and $iter \leq max - iter$, then repeat step 10; otherwise, the algorithm terminates.
-

5 Experimental results

Two sets of tests have been performed on the FCM algorithm and the proposed MODFCM algorithm on the Iris flower with four features. Different similarity measures in the solution clustering problems are used. Here, the Euclidean distance criterion is used due to its high efficiency. Also, the evaluation of the results obtained from the clustering of the data set with the DBSCAN algorithm and the direct transfer of the results to the FCM algorithm was performed. The algorithm was named DBSCAN – FCM. We analyze the convergence and the iterative process of algorithms. The convergence and the iterative process for these algorithms are shown in Table 1. As we can see from Table 1, the convergence speed of the proposed MODFCM algorithm is faster than the FCM algorithm and DBSCAN – FCM algorithm. This shows that the proposed MODFCM algorithm improves the convergence speed. Further analysis reveals that the proposed MODFCM algorithm can reduce the required clustering time effectively and improve the efficiency of the data processing.

Table 1: Comparison table of algorithms

Algorithm	Objective function	Number of iterations
FCM	12.469286	15
DBSCAN – FCM	14.169376	19
MODFCM	9.589957	19
MODFCM	9.589961	15

In the second experiment, the GENETIC algorithm and RAND index were used, and the performance of two algorithms was evaluated. The results show that although both algorithms reach the final solution, the speed of the proposed algorithm is increased because the algorithm is converged in fewer iterations. In addition, the RAND index in the first population generated was evaluated for both algorithms. It was about 0.67 for the proposed algorithm in most iterations, but the same amount for FCM was about 0.41. The evaluation diagram of two algorithms, FCM and MODFCM, are given in Figures 2 and 3, respectively. As a result shows, the proposed algorithm MODFCM has a better performance in achieving the desired clustering.

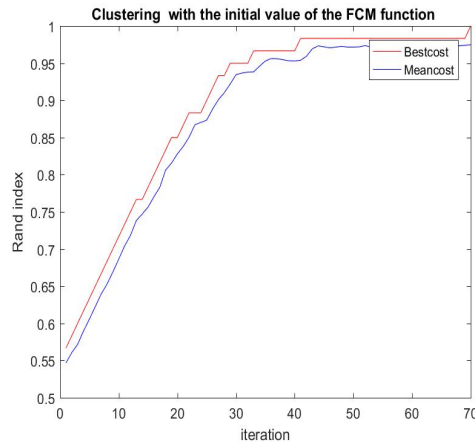


Figure 2: Evaluation of the performance of the FCM algorithm with RAND index

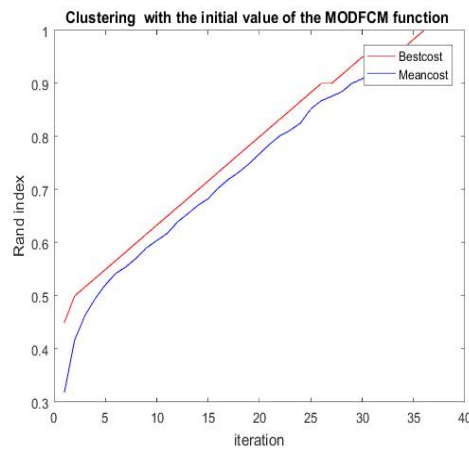


Figure 3: Evaluation of the performance of the MODFCM algorithm with RAND index

6 Conclusion

Today, there are many methods for data clustering, each of which has advantages and disadvantages. Methods can be achieved by combining algorithms to improve the results by covering each other's weaknesses. In this article, the FCM algorithm and the DBSCAN algorithm were combined. One of the advantages of the proposed algorithm in all the experiments compared

to FCM is that we do not face the problem of determining the number of clusters. The algorithm was improved by determining the optimal number of clusters. In addition, to increase the quality of clustering, optimal centers were also obtained. In total, by making these changes in the proposed method, it was found that by evaluating the objective function in both algorithms, the improvement of the objective function in the proposed algorithm with the same number of iterations has better performance than the FCM algorithm and the speed of convergence increases.

References

- [1] Abdellahoum, H., Mokhtari, N., Brahim, A. and Boukra, A. *CSFCM: An improved fuzzy C-Means image segmentation algorithm using a cooperative approach*, Expert Syst. Appl. 166 (2021), 114063.
- [2] Alomoush, W., Khashan, O.A., Alrosan, A., Houssein, E.H., Attar, H., Alweshah, M. and Alhosban, F. *Fuzzy clustering algorithm based on improved global best-guided artificial bee colony with new search probability model for image segmentation*, Sensors 22(22) (2022), 8956.
- [3] Ester, M., Kriegel, H. P. and Sander, J. *A density-based algorithm for discovering clusters in large spatial databases with noise*, kdd 96(34) (1996), 226–231.
- [4] Maraziotis, I.A. *A semi-supervised fuzzy clustering algorithm applied to gene expression data*, Pattern Recognit. 45(1) (2012), 637–648.
- [5] Mekhmoukh, A. and Mokrani, K. *Improved fuzzy C-Means based particle swarm optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation*, Comput. Methods Programs Biomed. 122(2) (2015), 266–281.
- [6] Parimala, M., Lopez, D. and Senthilkumar, N.C. *A survey on density based clustering algorithms for mining large spatial databases*, Int. J. Adv. Sci. Technol. 31(1) (2011), 59–66.
- [7] Ruspini, E.H., Bezdek, J.C. and Keller, J.M. *Fuzzy clustering: A historical perspective*, IEEE Comput. Intell. Mag. 14(1) (2019), 45–55.
- [8] Shamsirband, S., Amini, A., Anuar, N.B., Kiah, M.L.M., Teh, Y.W. and Furnell, S. *D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks*, Measurement 55 (2014), 212–226.
- [9] Singh, T., Saxena, N., Khurana, M., Singh, D., Abdalla, M. and Alshazly, H. *Data clustering using moth-flame optimization algorithm*, Sensors 21(12) (2021), 4086.

- [10] Wei, L.M. and Xie, W.X. *Rival checked fuzzy c-means algorithm*, ACTA ELECTONICA SINICA 28(7) (2000), 79.
- [11] Xu, R. and Wunsch, D. *Survey of clustering algorithms*, IEEE Trans. Neural Netw. 16(3) (2005), 645–678.
- [12] Xue, Z., Shang, Y. and Feng, A. *Semi-supervised outlier detection based on fuzzy rough C-means clustering*, Math. Comput. Simul. 80(9) (2010), 1911–1921.
- [13] Zhang, H., Li, H., Chen, N., Chen, S. and Liu, J. *Novel fuzzy clustering algorithm with variable multi-pixel fitting spatial information for image segmentation*, Pattern Recognit. 121 (2022), 108201.

How to cite this article

Barkhordari Firozabadi, S., Shahzadeh Fazeli, S.A., Zarepour Ahmadabadi, J. and Karbassi, S.M., Improving the performance of the FCM algorithm in clustering using the DBSCAN algorithm. *Iran. J. Numer. Anal. Optim.*, 2023; 13(4): 763-774. <https://doi.org/10.22067/ijnao.2023.82361.1260>