A Novel Fuzzy C-means Clustering Algorithm Based on Local Density

Jian-Jun Liu, Jian-Cong Fan

A Novel Fuzzy C-means Clustering Algorithm Based on Local Density

Jian-jun Liu[1], Jian-cong Fan[1, 2, 3]

[1] College of Computer Science and Engineering, Shandong University of Science and Technology, China
[2] Provincial Key Lab. for Information Technology of Wisdom Mining of Shandong Province, Shandong University of Science and Technology, China
[3] Provincial Experimental Teaching Demonstration Center of Computer, Shandong University of Science and Technology, China
fanjiancong@sdust.edu.cn

Abstract. Fuzzy C-means (FCM) clustering algorithm is a fuzzy clustering algorithm based on objective function. FCM is the most perfect and widely used algorithm in the theory of fuzzy clustering. However, in the process of clustering, FCM algorithm needs to randomly select the initial cluster center. It is easy to generate problems such as multiple clustering iterations, low convergence speed and unstable clustering. In order to solve the above problems, a novel fuzzy C-means clustering algorithm based on local density is proposed in this paper. Firstly, we calculate the local density of all sample points. Then we select the sample points with the local maximum density as the initial cluster center at each iteration. Finally, the selected initial cluster center are combined with the traditional FCM clustering algorithm to achieve clustering. This method improved the selection of the initial cluster center. The comparative experiment shows that the improved FCM algorithm reduces the number of iterations and improves the convergence speed.

Keywords: Fuzzy C-means Algorithm, Local Density, Clustering.

1 Introduction

Cluster analysis [1-5] is an important function of data mining, and the clustering algorithm is the core of current research. Clustering is to divide the data set into multiple clusters or classes based on the similarity between a set of unlabeled data objects. A good clustering algorithm should be able to produce high-quality clustering results: clusters. These clusters must have two characteristics: (1) high intra-cluster similarity; (2) low inter-cluster similarity. The quality of the clustering results depends not only on the similarity evaluation method and its specific implementation, but also on whether the method can find some hidden patterns or all hidden patterns [6].

Fuzzy C-means (FCM) clustering algorithm [7-10] is one of the widely used fuzzy clustering algorithms. FCM algorithm belongs to the category of fuzzy clustering algorithms based on objective functions. However, the traditional FCM algorithm has some disadvantages. One of the issues is that the number of clusters needs to be determined manually, and the algorithm is sensitive to the initial cluster center. In addition, the FCM algorithm is prone to problems such as multiple clustering iterations, low convergence speed and local optimal solution. Many algorithms have been proposed to improve the FCM algorithm. Wang et al. [11] systematically improved the traditional fuzzy clustering algorithm. They proposed a new method by combining PSO(particle swarm optimization) and fuzzy C-means algorithm. By a simple and effective particle encoding method, the best initial cluster center and fuzzy weighting exponent were both searched in the process of PSO. Li et al. [12] proposed a scheduling algorithm based on fuzzy clustering and two-level scheduling mode. Geweniger et al. [13] combined the median c-means algorithm with the fuzzy c-means method to improve the accuracy of the algorithm. Median clustering is a powerful methodology for prototype based clustering of similarity/dissimilarity data. The approach is only applicable for vector (metric) data in its original variant. Wang et al. [14] presented a rough-set [15, 16] based measurement for the membership degree of fuzzy C-means algorithm, and take the advantage of the positive region set and the boundary region set of rough set. Lai et al. [17] presented a rough k-means clustering algorithm by minimizing the dissimilarity to solve the divergence problem of the original approaches that the cluster centers may not be converged to their final positions. Cai et al. [18] proposed a novel initial cluster centroids selection algorithm, called WLV-K-means (weighted local variance K-means). The WLV-K-means algorithm employs the weighted local variance to measure the

density of each sample, which can find samples with higher density. This algorithm also uses the improved max-min method to select cluster centroid heuristically. Liu et al. [19] proposed to combine the FCM algorithm and DPC (Clustering by fast search and find of density peaks) algorithm. Firstly, DPC algorithm is used to automatically select the center and number of clusters, and then FCM algorithm is used to realize clustering. The comparison experiments show that the improved FCM algorithm has a faster convergence speed and higher accuracy. Khan and Ahmad [20] proposed a new cluster center initialization algorithm (CCIA). By clustering the samples in each dimension, we find that the K' ($K' > K$) clusters have the same pattern points, and get the center points of the K' clusters. Then we use the data compression method in reference [21] to merge the neighborhood of high-density samples, and finally get the K initial center points. In this paper, we fully consider the constraints of cluster centers in the process of cluster center selection and optimization. Firstly, the initial cluster center is selected by calculating the local density of each sample point. Then the selected initial cluster center is combined with the traditional FCM algorithm to cluster the data. Therefore, we propose a novel fuzzy C-means clustering algorithm based on local density (LD-FCM).

The rest of this paper are organized as follows. In Section 2, the concept of fuzzy clustering and Fuzzy C-means clustering algorithms are briefly reviewed. Some important preliminary knowledge used in our proposed approaches are stated. In Section 3, we present the algorithms proposed in this paper, and some theories and analysis necessary in it. In Section 4, experimental studies are conducted to verify the effectiveness of our proposed algorithm. Section 5 concludes the paper.

2 Preliminaries

2.1 Fuzzy Cluster Analysis

The concept of fuzzy clustering was firstly proposed by Professor Ruspiniv [22]. Fuzzy clustering is an algorithm combining fuzzy mathematics with clustering methods. Fuzzy clustering determines the fuzzy relationship between the samples by method of fuzzy mathematics. In other words, the clustering results are blurred, so that the problem of data attribution in the real world can be described objectively from multiple angles. Therefore, fuzzy clustering analysis has become one of the mainstream directions of clustering research.

Fuzzy clustering [23] calculates the similarity between different data samples by using some distance measurement method. Each data sample is divided into different clusters according to the similarity between data samples. For any number of data sample subsets k ($1 \le k \le C$), where C is the number of clusters, the data sample X_i ($1 \le i \le N$) (N is the number of samples) will belong to this cluster with a fuzzy membership degree, which is similar to a probability value. The fuzzy clustering will obtain membership matrix $[u_{ik}]$ ($1 \le k \le C, 1 \le i \le N$) and cluster center $V = \{v_1, v_2, \cdots v_c\}$. And then the membership matrix is judged by hardening matrix technology to determine the final attribution result of data samples. The membership matrix is composed of the fuzzy degree that each data sample belongs to a subset. The value range of each element in the membership matrix is [0,1]. In other words, if the membership degree of data sample to a subset is greater than that of other subsets, it means that the sample is more likely to belong to the subset. When $u_{ik} = 1$, it means that x_i belongs completely to the k-th cluster, while when $u_{ik} = 0$, it means that x_i does not belong to the k-th cluster at all.

2.2 Fuzzy C-means Clustering Algorithm

Fuzzy C-means (FCM) clustering algorithm [8] is an improvement of the common C-means algorithm. The common C-means algorithm is hard for data classification, while FCM is a soft fuzzy division. Many of the discussions in this paper are based on the FCM algorithm. Therefore, the FCM algorithm is described in detail.

Supposed that the data sample set $X=\{x_1, x_2, \ldots\ldots, x_n\} \subset R^s$ is an s-dimensional data set in Euclidean space, n is the number of samples. Where x_i contains the s dimensions, which is expressed as $x_i = \{x_i^1, \ldots, x_i^d, \ldots, x_i^s\}$. FCM algorithm divides X into C classes ($2 \le C \le n$), and has C cluster centers $V = \{v_1, v_2, \ldots\ldots, v_c\}$. Thus, FCM algorithm can be expressed as the following mathematical programming matters:

$$Minimize\ J(X, U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^m \, ||x_j - v_i||^2 \qquad (1)$$

And satisfied

$$\sum_{i=1}^{c} u_{ij} = 1 \ (j = 1,2,\cdots,n) \quad (2)$$

Where u_{ij} is the membership degree of data sample x_j belonging to a certain class i. $U = (u_{ij})_{c\times n}$ is the fuzzy partition matrix. The value of membership degree of each data sample relative to each cluster can be found from the fuzzy partition matrix. The similarity between the data sample x_j and the class center of the class i is calculated by Euclidean distance, which is recorded as $d_{ij} = ||x_j - v_i||$. m is the fuzzy weight index, also known as the fuzzy factor. m is mainly used to adjust the fuzzy degree of the fuzzy partition matrix.

The specific steps of the algorithm are as follows.

Step1: We set the number of clusters *C* and the fuzzy factor m (usually 1.5 to 2.5), We initialize the membership matrix $U^{(\gamma)}(\gamma = 0)$, and make it satisfy the Eq. (2);

Step2: The cluster center $V^{(\gamma+1)} = \{v_1, v_2, \ldots\ldots, v_c\}$ is updated according to Eq. (3);

$$v^{(\gamma+1)}{}_i = \frac{\sum_{j=1}^{n}(u^{(\gamma)}{}_{ij})^m \cdot x_j}{\sum_{j=1}^{n}(u^{(\gamma)}{}_{ij})^m}, i = 1,2,\ldots\ldots,c \quad (3)$$

Step3: The membership matrix $U^{(\gamma+1)} = (u_{ij})_{c\times n}$ is updated according to Eq. (4);

$$u^{(\gamma+1)}{}_{ij} = [\sum_{k=1}^{c}\left(\frac{||x_j - v_i||^{2(\gamma)}}{||x_j - v_k||^{2(\gamma)}}\right)^{\frac{2}{m-1}}]^{-1}, i = 1,2,\ldots.c; j = 1,2,\ldots.n \quad (4)$$

Step4: We calculate $e = ||U^{(\gamma+1)} - U^{(\gamma)}||$. If $e \le \eta$ (η is the threshold, generally 0.001 to 0.01), then the algorithm stops; Otherwise, $\gamma = \gamma + 1$, go to Step2.

Step5: The samples are classified and output according to the final membership matrix *U*. If the sample x_j satisfies $u_{ij} > u_{kj}$, then x_j is classified into the *i*-th cluster, where u_{ij} represents the membership degree of the sample x_j to the cluster center v_i.

The FCM algorithm [24] is a point-by-point iterative clustering algorithm based on the sum of squared errors as a criterion function. This iterative process starts from a random cluster center. In order to find the minimum value of the objective function $J(X,U,V)$, the cluster center *V* and the membership matrix *U* are iteratively calculated by Eq. (3) and Eq. (4). Therefore, the value of the objective function $J(X,U,V)$ is continuously reduced until the value is minimized. Generally, the convergence condition of the algorithm is that the difference between the objective functions of two iterations is less than the threshold η, or the specified number of iterations is reached. When the objective function is minimized, the final clustering result of the data samples is obtained, that is, the cluster center V and the membership matrix U after the fuzzy division. Then the purpose of determining the sample category is achieved.

3 Fuzzy C-means Clustering Algorithm Based on Local Density (LD-FCM)

3.1 Selection of Initial Cluster center

The traditional FCM clustering algorithm is a sort of local optimal search algorithm. FCM algorithm has some imperfections, such as being sensitive to the initial cluster center and tending to be trapped in the local optimal solution. The random selection of the initial cluster center of the FCM algorithm will lead to the unstable clustering results that are different each time. Therefore, the effect of the clustering may not be best every time, which limits the application of the algorithm. In order to solve the above problems, the paper improve the selection of initial cluster centers in the FCM algorithm. We propose a novel fuzzy C-means clustering algorithm based on local density (LD-FCM). LD-FCM algorithm calculates the local density ρ_i of all sample points in the algorithm and select the sample point with the local maximum density as the initial cluster center by using the distance matrix *D* and the distance threshold α in each iteration. In this way, the selection of the initial cluster center not only ensures the compactness of the objects in the same cluster, but also ensures the separation of the cluster centers [25]. The specific improvements on the selection of the initial cluster center are as follows:

Supposed that the data sample set $X=\{x_1,x_2,\ldots\ldots,x_n\} \subset R^s$ is an *s*-dimensional data set in Euclidean space, *n* is the number of samples. Where x_i contains the *s* dimensions, which is expressed as $x_i = \{x_i^1, \ldots, x_i^d, \ldots, x_i^s\}$. LD-FCM algorithm divides *X* into *C* classes ($2 \le C \le n$), and has *C* cluster centers $V = \{v_1, v_2, \ldots\ldots, v_c\}$.

Step1: Calculate the distance between any two samples according to Euclidean Distance Equation (5), and generate a distance matrix *D*

$$d_{ij} = \sqrt{\sum_{d=1}^{s}(x_i^d - x_j^d)^2} \quad (5)$$

Step2: Calculate the local density ρ_i of the data object x_i according to Equation (6)

$$\rho_i = \sum_{x_j \in X} \chi(d_{ij} - d_c) \quad (6)$$

Where d_c represents the truncation distance. $\chi(x) = \begin{cases} 1, & x \leq 0 \\ 0, & x > 0 \end{cases}$, the meaning of this equation is to count the number of data points whose distance to the *i*-th data point is less than the truncation distance d_c, and use it as the local density of the *i*-th data point.

Step3: Arrange the local density of each sample point from large to small: $\rho_i > \rho_j > \rho_k > \cdots > \rho_n$, and take the sample point with the local maximum density ρ_i as the first cluster center v_1.

Step4: Select the distance threshold α, then find all samples whose distance from the first cluster center v_1 is greater than α by using the distance matrix *D*. And select the sample point with the highest local density among these samples as the second cluster center v_2.

Step5: Similarly, find all samples whose distance from the found sample points is greater than α in the remaining samples, and select the sample point with the highest local density among these samples as the third cluster center v_3.

Step6: Repeat Step5 until *C* clusters are found. In this way, *C* initial cluster centers will be obtained.

3.2 The LD-FCM Algorithm

LD-FCM clustering algorithm is divided into two stages. In the first stage, the method of local maximum density and the distance threshold α are used to select the initial cluster center. In the second stage, the FCM algorithm is performed with the initial cluster center that obtained in the first stage. The specific steps of the LD-FCM algorithm are as follows.

Step1: Calculate the distance between any two samples according to Euclidean Distance Eq. (5), and generate a distance matrix *D*;

Step2: Calculate the local density ρ_i of the data object x_i according to Eq. (6). Arrange the local density of each sample point from large to small: $\rho_i > \rho_j > \rho_k > \cdots > \rho_n$, and take the sample point with the highest local density ρ_i as the first cluster center v_1;

Step3: Select the distance threshold α ($\alpha > 0$), then find all samples whose distance from the first cluster center v_1 is greater than α by using the distance matrix *D*. And select the sample point with the highest local density among these samples as the second cluster center v_2;

Step4: Similarly, find all samples whose distance from the found sample points is greater than α in the remaining samples, and select the sample point with the highest local density among these samples as the third cluster center v_3;

Step5: Repeat Step4 until *C* clusters are found. In this way, *C* initial cluster centers $v_i(k), (i = 1, 2, \ldots, C)$ iterations will be obtained;

Step6: Set the number of iterations k = 1, and use the result of Step 5 as the initial cluster center $v_i(k), (i = 1, 2, \ldots, C)$;

Step7: The membership matrix $U^{(\gamma+1)} = (u_{ij})_{c \times n}$ (i = 1,…,c, j = 1,…,n) is updated according to initial cluster center $v_i(k)$ and Eq. (4);

Step8: The cluster center $V^{(\gamma+1)} = \{v_1, v_2, \ldots\ldots, v_c\}$ is updated according to Eq. (3);

Step9: We calculate $e = ||U^{(\gamma+1)} - U^{(\gamma)}||$. If $e \leq \eta$ (η is the threshold, generally 0.001 to 0.01), then the algorithm stops; Otherwise, $\gamma = \gamma + 1$, go to Step7.

Step10: The samples are classified and output according to the final membership matrix *U*. If the sample x_j satisfies $u_{ij} > u_{kj}$, then x_j is classified into the *i*-th cluster, where u_{ij} represents the membership degree of the sample x_j to the cluster center v_i.

4 Results

In order to test the effect of the LD-FCM algorithm, we have used several artificial datasets and real datasets in UCI for experiments. We also compared the LD-FCM algorithm proposed in this paper with FCM algorithm, K-means algorithm, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm and DP-FCM (Density peak-based FCM) algorithm. The K-means algorithm and the

DBSCAN algorithm are classic algorithms in the partition-based clustering algorithm and the density-based clustering algorithm, respectively. In the K-means [26] algorithm, each cluster is represented by the mean of the objects in the corresponding cluster. However, K-means algorithm can only obtain local optimal solution by adopting iterative algorithms. In addition, K-means algorithm work well when analyzing small and medium-sized data sets to find circular or spherical clusters. But K-means algorithm perform poorly when analyzing large-scale data sets or complex data types, so they need to be extended [27-29]. DBSCAN [30] algorithm derives the maximum density connected set according to the density reachability relationship, which is a category or cluster for our final clustering. DBSCAN algorithm can divide areas with enough density into clusters, and find clusters of arbitrary shape in the spatial database with noise [31]. The DP-FCM algorithm is proposed by Liu et al. [19]. Firstly, DPC algorithm is used to automatically select the center and number of clusters, and then FCM algorithm is used to realize clustering.

4.1 The Datasets and Evaluation Indexes of Experiment

The information of the experimental datasets is shown in Table 1 and Table 2, where Table 1 is artificial datasets and Table 2 is a real datasets in UCI [32]. These datasets have some discrimination in the number of attributes and the number of clusters.

Table 1. Description of the artificial datasets.

Dataset	Size	Attribute	Number of class
Set	5000	2	15
R15	600	2	15
Shape	1000	2	4
Sizes	1000	2	4
Twenty	1000	2	20
Target	770	2	6

Table 2. Description of the real datasets in UCI.

Dataset	Size	Attribute	Number of class
Iris	150	4	3
Aggregation	788	2	7
Wine	178	13	3
Pima	768	8	2
Compound	399	2	6
Seeds	210	7	3
Wingnut	1016	2	2
Glass	214	10	7

The evaluation indexes of the experimental results are Accuracy, NMI (Normalized Mutual Information) and ARI (Adjusted Rand index) [33,34]. Accuracy is the number of right samples divided by the total number of samples. NMI is an information measure in information theory, and its value range is [0,1]. ARI is the goodness of fit which measures the distribution of two data, and its value range is [-1,1]. The larger the values of the three evaluation indexes are, the better the clustering result is. Their definitions are as follows.

$$ACC = \frac{\sum_{i=1}^{k} a_i}{|U|} \tag{7}$$

Where K is the number of clusters, a_i is the number of samples correctly classified into C_i, and U is the all samples .

$$I(X,Y) = \sum_{h=1}^{k(a)} \sum_{l=1}^{k(b)} n_{h,l} log(\frac{n n_{h,l}}{n_h^{(a)} n_l^b}) \tag{8}$$

$$H(X) = \sum_{h=1}^{k(a)} n_h^{(a)} log \frac{n_h^{(a)}}{n} \tag{9}$$

$$H(Y) = \sum_{l=1}^{k(b)} n_l^{(b)} log \frac{n_l^{(b)}}{n} \tag{10}$$

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{11}$$

Where *X* and *Y* are the random variables, *I(X;Y)* represents the mutual information of two variables, and *H(X)* is the entropy of *X*.

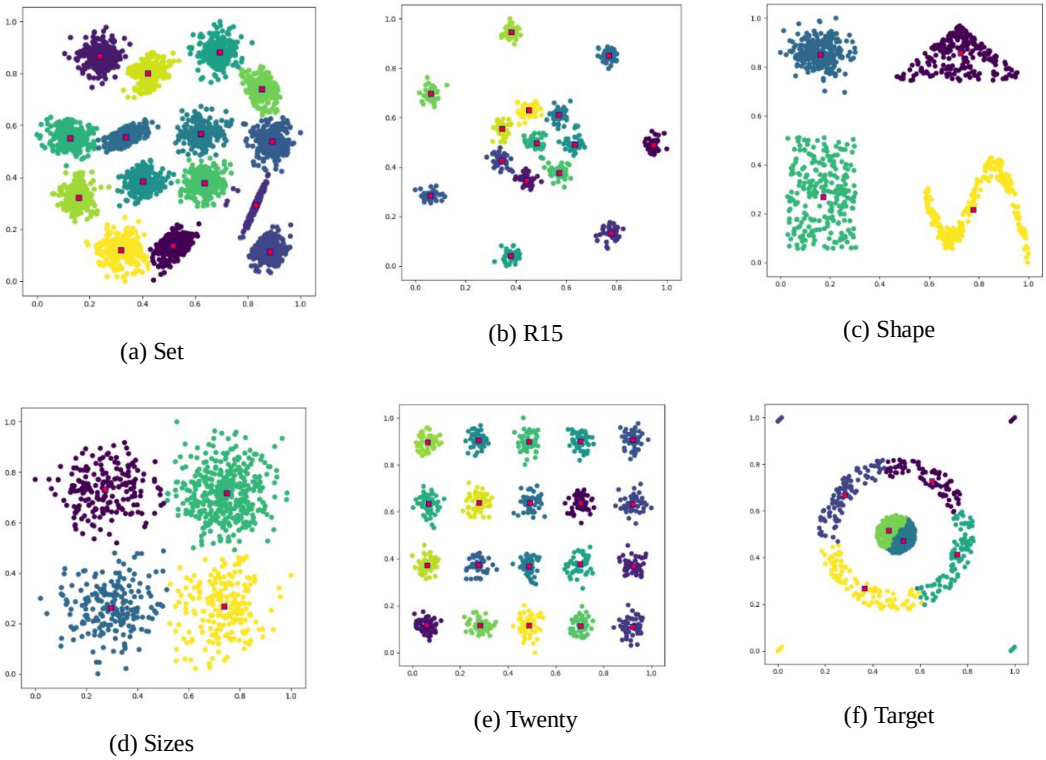
$$RI = \frac{a+b}{C_2^{n_{samples}}} \tag{12}$$

$$ARI = \frac{RI - E|RI|}{max(RI) - E|RI|} \tag{13}$$

Where *C* is the actual category information. *K* is the clustering result. *a* represents the logarithms of elements of the same categories in both *C* and *K*. *b* represents the logarithms of elements of the different categories in both *C* and *K*. $C_2^{n_{samples}}$ represents the logarithm that can be formed in the datasets. *RI* represents the Rand index. *E* is the expectation. *max()* is the function to find the maximum value.

4.2 Results on the Artificial Datasets

As shown in Fig. 1 (a)-(f), it shows the clustering results of the LD-FCM algorithm on six different artificial datasets. The datasets including Set, R15, Shape, Sizes, Twenty and Target. Fig. 1 shows that the LD-FCM algorithm can correctly cluster the datasets with spherical or elliptical shapes. The experimental results show that the LD-FCM algorithm is very effective in seeking clusters with any shape, density, distribution and number. LD-FCM algorithm solves the disadvantages of the original algorithm. LD-FCM algorithm can reasonably select the initial cluster center, then correctly calculate the membership of each sample, and each clustering result is relatively stable.



(a) Set (b) R15 (c) Shape

(d) Sizes (e) Twenty (f) Target

Fig. 1. Clustering result graph.

4.3 Results on the Real Datasets in UCI

The experimental results of the real datasets in UCI are shown in Table 3, Table 4 and Table 5. These tables show the Accuracy, NMI and ARI of each clustering result in every algorithm. The experimental results are showed by the way of percentages. The numbers highlighted in bold in the three tables indicate the best performance in the evaluation aspect.

Table 3. Comparison of Accuracy of clustering algorithms.

Dataset	K-means(%)	DBSCAN(%)	FCM(%)	DP-FCM(%)	LD-FCM(%)
Iris	74.56	63.45	89.33	91.33	**91.69**
Aggregation	69.25	77.23	64.34	**92.21**	80.25
Wine	67.93	75.74	70.79	70.79	**95.50**
Pima	65.84	65.21	66.66	65.12	**69.34**
Compound	65.77	63.34	61.74	65.91	**66.79**
Seeds	**90.56**	62.85	87.22	87.27	90.47
Wingnut	93.24	92.58	92.08	**98.43**	95.29
Glass	83.42	80.33	79.85	85.49	**85.83**

Table 4. Comparison of NMI of clustering algorithms.

Dataset	K-means(%)	DBSCAN(%)	FCM(%)	DP-FCM(%)	LD-FCM(%)
Iris	70.93	72.19	70.21	**74.96**	73.03
Aggregation	76.22	79.69	73.35	82.99	**83.15**
Wine	79.04	80.31	71.20	38.33	**84.35**
Pima	1.84	2.06	1.29	**4.05**	3.97
Compound	69.22	66.15	67.23	70.03	**70.12**
Seeds	**70.03**	56.83	60.52	61.51	67.35
Wingnut	81.92	**86.22**	67.88	85.73	73.92
Glass	63.22	**65.83**	29.14	40.27	47.32

Table 5. Comparison of ARI of clustering algorithms.

Dataset	K-means(%)	DBSCAN(%)	FCM(%)	DP-FCM(%)	LD-FCM(%)
Iris	70.83	62.32	70.02	**74.64**	72.49
Aggregation	70.04	76.21	57.15	**83.62**	71.34
Wine	76.35	60.06	79.22	37.21	**84.97**
Pima	3.97	3.22	5.93	9.32	**10.69**
Compound	50.94	32.72	50.01	51.99	**52.72**
Seeds	70.65	48.67	70.40	64.29	**71.62**
Wingnut	**92.29**	81.03	73.21	90.87	77.01
Glass	37.76	**42.22**	16.67	33.73	34.11

Based on the three evaluation indicators, it can be found that the clustering results of the LD-FCM algorithm are generally better than the other four algorithms. LD-FCM algorithm can effectively cluster the data set and produces improved clustering results. For datasets with a large number of clusters and high dimensions, the LD-FCM algorithm can still effectively cluster the datasets. In addition, when there are many noise samples in the data set and the boundaries between clusters are not clear, the LD-FCM algorithm is more suitable, which also proves the effectiveness of the LD-FCM algorithm.

Table 6. Comparison of running time of clustering algorithms (seconds/s).

Dataset	K-means(%)	DBSCAN(%)	FCM(%)	DP-FCM(%)	LD-FCM(%)
Iris	0.059	0.565	0.148	0.884	0.032
Aggregation	0.309	1.016	2.602	3.068	0.059
Wine	0.098	0.832	0.168	1.203	0.032
Pima	0.223	1.457	0.372	1.776	0.099
Compound	0.984	5.223	0.564	4.251	0.090
Seeds	0.122	0.973	0.164	0.932	0.074
Wingnut	0.159	1.006	0.439	2.041	0.197
Glass	8.722	10.021	0.313	0.806	0.912

Table 6 shows the time performance of each algorithm on different datasets. From the experimental results, we can see that the algorithm in this paper inherits the time advantage of FCM algorithm, and running time takes less time than other algorithms, so it has some advantages in both running time and memory consumption.

5 Conclusion

Firstly, this paper introduces the concepts of fuzzy clustering. The principle of fuzzy C-means clustering algorithm is explained. Considering that the selection of the initial cluster center of the traditional FCM clustering algorithm is random, which will lead to the unstable clustering results and the clustering effect may not be the best each time. In order to solve the above problems, a novel fuzzy C-means clustering algorithm based on local density is proposed in this paper. This algorithm uses the local maximum density of sample points to improve the selection of the initial cluster center, which reduces the number of iterations and avoids falling into a local optimal solution. Through specific experiments, better clustering results are obtained by using artificial datasets. The real datasets in UCI were used to analyze the algorithm from three evaluation indexes about Accuracy, NMI, ARI and running time of real datasets. The experimental results show that the LD-FCM algorithm is better than FCM algorithm, K-means algorithm, DBSCAN algorithm and DP-FCM algorithm. Running time of LD-FCM algorithm takes less time than other algorithms. Therefore, it can be concluded that the LD-FCM algorithm has good effectiveness and robustness.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by Shandong Provincial Natural Science Foundation of China under Grant ZR2018MF009, ZR2019MF003, The State Key Research Development Program of China under Grant 2017YFC0804406, National Natural Science Foundation of China under Grant 91746104, 61976126, the Special Funds of Taishan Scholars Construction Project, and Leading Talent Project of Shandong University of Science and Technology.

References

1. He ZH, Lei YJ.: Research on intuitionistic fuzzy C-means clustering algorithm. Control and Decision 26(6), 847–856 (2011).
2. Zhang Y, Li ZM, Zhang H, Yu Z, Lu TT.: Fuzzy c-means clustering-based mating restriction for multiobjective optimization. International Journal of Machine Learning and Cybernetics 9, 1609–1621 (2018).
3. Wang XZ, Xing HJ, Li Y.: A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning. IEEE Trans Fuzzy Syst 23(5), 1638–1654 (2015).
4. Wang R, Wang XZ, Sam K, Chen X.: Incorporating diversity and informativeness in multiple-instance active learning. IEEE Trans Fuzzy Syst 25(6), 1460–1475 (2017).
5. Liu CS, Xu QL.: A fuzzy C-means clustering algorithm based on density peak algorithm optimization. Computer Engineering and Applications 54(14), 153–157 (2018).
6. Fan JC, Niu ZH, Liang YQ, Zhao ZY.: Probability model selection and parameter evolutionary estimation for clustering imbalanced data without sampling. Neurocomputing 211, 172–181 (2016).
7. Xie XL, Beni G.: A validity measure for fuzzy clustering. IEEE Trans Pami 13(13), 841–847 (1991).
8. Bezdek JC.: Pattern recognition with fuzzy objective function algorithms. Adv Appl Pattern Recognit 22(1171), 203–239 (1981).
9. Kosko B.: Fuzzy systems as universal approximators. IEEE Trans Comput 43(11), 1329–1333 (1994).
10. Fan JC.: OPE-HCA: an optimal probabilistic estimation approach for hierarchical clustering algorithm. Neural Computing and Application 31(7), 2095–2105 (2019).
11. Wang ZH, Liu ZJ, Chen DH.: Research of PSO-based fuzzy C-means clustering algorithm. Computer Science 39(9), 166–169 (2012).
12. Li WJ, Zhang QF, Ping LD, Pan XZ.: Cloud scheduling algorithm based on fuzzy clustering. Journal on Communications 33(3), 147-154 (2012).
13. Geweniger T, Zülke D, Hammer B, Villmann T.: Median fuzzy c-means for clustering dissimilarity data. Neurocomputing 73, 1109–1116 (2010).
14. Wang ZH, Fan JC.: A rough-set based measurement for the membership degree of fuzzy C-means algorithm. In: Proceedings of SPIE-the international society for optical engineering, 3rd international workshop on pattern recognition (2018).
15. Pawlak Z.: Rough sets. Int J Comput Inf Sci 11(5), 341–356 (1982).
16. Fan JC, Li Y, Tang LY, Wu GK.: RoughPSO: rough set-based particle swarm optimisation. Int J Bio-inspired Comput 12, 245–253 (2018).
17. Lai JZC, Juan EYT, Lai FJC.: Rough clustering using generalized fuzzy clustering algorithm. Pattern Recognition 46, 2538–2547 (2013).

18. Cai YH, Liang YQ, Fan JC, Li X, Liu WH.: Optimizing Initial Cluster Centroids by Weighted Local Variance in K-means Algorithm. Journal of Frontiers of Computer Science and Technology 10(5), 732–741 (2016).
19. Liu XY, Fan JC, Chen ZW.: Improved fuzzy C-means algorithm based on density peak. Int. J. Mach. Learn. & Cyber 11, 545–552 (2020).
20. Khan SS, Ahmad A.: Cluster center initialization algorithm for K-means clustering. Pattern Recognition Letters 25(11), 1293–1302 (2004).
21. Mitra P, Murthy CA, Pal SK.: Density-based multiscale data condensation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(6), 734–747 (2002).
22. Ruspini EH.: A new approach to clustering. Inf Control 15(1), 22–32 (1969).
23. Li Y, Fan J, Pan J-S, Mao G, Wu G.: A novel rough fuzzy clustering algorithm with a new similarity measurement. J Internet Technol 20(4), 1145–1156 (2019).
24. Xue Z, Shang Y, Feng A.: Semi-supervised outlier detection based on fuzzy rough C-means clustering. Mathematics and Computers in Simulation 80(9), 1911–1921 (2010).
25. Xia YY, Liu Y, Huang YD.: Community discovery based on improved clustering algorithm with central constraints. Computer Engineering and Applications 54(8), 265–270 (2018).
26. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24(7), 881–892 (2002).
27. Yang SL, Li YS, Hu XX, Pan RY.: Optimization study on K value of K-means algorithm. Systems Engineering-Theory and Practice 2, 97–101 (2006).
28. Zhang YF, Mao JL, Xiong ZY.: An improved K-means algorithm. Computer Applications 23(8), 31–60 (2003).
29. Wang Z, Liu GJ, Chen EH.: A K-means algorithm based on optimized initial center points. Pattern Recognition and Artificial Intelligence 22(2), 299–304 (2009).
30. Sander J, Ester M, Kriegel HP, Xu XW.: Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. Data Min Knowl Discov 2(2), 169–194 (1998).
31. Wang G, Lin GY.: Improved Adaptive Parameter DBSCAN Clustering Algorithm. Computer Engineering and Applications 1–8 (2019).
32. UCI Homepage, http://archive.ics.uci.edu/ml/index.php, last accessed 2019/09/17.
33. Fahad A, Alshatri N, Tari Z.: A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans Emerg Top Comput 2(3), 267–279 (2014).
34. Bie R, Mehmood R, Ruan S.: Adaptive fuzzy clustering by fast search and find of density peaks. Pers Ubiquitous Comput 20(5), 785–793 (2016).