IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON AI-DRIVEN BIG DATA PROCESSING: THEORY, METHODOLOGY, AND APPLICATIONS

Received March 31, 2019, accepted May 9, 2019, date of publication May 24, 2019, date of current version June 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918952

An Adaptive Density Peaks Clustering Method With Fisher Linear Discriminant

LIN SUN, RUONAN LIU, JIUCHENG XU, AND SHIGUANG ZHANG
College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China

Corresponding author: Lin Sun (linsunok@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772176, Grant 61402153, and Grant 61370169, in part by the China Postdoctoral Science Foundation under Grant 2016M602247, in part by the Plan for Scientific Innovation Talent of Henan Province under Grant 184100510003, in part by the Key Scientific and Technological Project of Henan Province under Grant 182102210362, in part by the Young Scholar Program of Henan Province under Grant 2017GGJS041, and in part by the Natural Science Foundation of Henan Province under Grant 182300410130.

ABSTRACT Clustering is one of the most important topics in data mining and machine learning. The density peaks clustering (DPC) algorithm is a well-known density-based clustering method that can efficiently and effectively deal with non-spherical clusters. However, the computational methods of the local density and the distance measure are simple and easily ignore the correlation and the similarity between samples, and the manual setting of parameters has a great influence on the clustering results; therefore, the clustering performance of DPC is poor on the high-dimensional datasets. To address these issues, this paper presents an adaptive DPC algorithm with Fisher linear discriminant for the clustering of complex datasets, called ADPC-FLD. First, the kernel density estimation function is introduced to calculate the local density of the sample points. Pearson correlation coefficient between samples as weight is employed to construct a weighted Euclidean distance function to measure the distance between samples. This considers both the spatial structure and the correlation of the samples. Then, a novel density estimation entropy is proposed, and based on the minimization of density estimation entropy, the density estimation parameters are adaptively selected according to the distribution characteristics of the data, which can efficiently eliminate the influence of manual setting. Third, an adaptive strategy of cluster center selection is designed to avoid the error caused by the noise data as the cluster centers and the uncertainty of manually selecting the cluster centers. Finally, Fisher linear discriminant algorithm is used to eliminate the irrelevant information and reduce the dimensionality of high-dimensional data, following on which an adaptive DPC method is implemented on six synthetic datasets, thirteen UCI datasets and seven gene expression datasets for comparing with other related algorithms. The experimental results on 26 datasets show that the proposed algorithm significantly outperforms several outstanding clustering approaches in terms of clustering accuracy and efficiency.

INDEX TERMS Density peaks clustering, Kernel density estimation function, density estimation entropy, Fisher linear discriminant, reduction.

I. INTRODUCTION

Recently, with the development of artificial intelligence (AI), the AI-driven big data processing technologies in many fields of pattern recognition, machine learning and deep learning, still face many challenges in the efficient clustering analysis for the large-scale heterogeneous complex datasets [1]. Clustering is a fundamental technique of utilizing information from the dataset or additional constraints to separate the objects into several groups in the real-world [2]. So it has been developed and widely applied to data mining, text analysis, information retrieval, image segmentation, biomedicine, and gene engineering [3]–[8]. Clustering methods divide a set of instances into several groups without any prior knowledge using the similarity of objects in which patterns in the same group have more similarities to each other [9]. In other words, the task of clustering is to find a set of groups that the similar objects are in the same group and the different objects are separated into different groups [10], [11].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma.

In the last few years, a great many clustering methods have been proposed, which can be roughly categorized into the following several categories: hierarchical, partitioning, density/neighborhood-based, and soft-computing methods [12]. Zheng *et al.* [13] proposed a hierarchical co-clustering approach to simultaneously group links and entity classes. Gullo *et al.* [14] presented a hierarchical clustering of uncertain objects by revising the key notions of cluster merging criterion and distance between uncertain cluster prototypes. The partitioning-based clustering methods are represented by k-modes and c-means [4]. Li *et al.* [15] adopted an interval kernel distance to calculate the distance of interval data and cluster prototypes for fuzzy c-means clustering incomplete datasets. Note that some density-based clustering methods are alike to the hierarchical-based clustering methods, the differences between them lie in the linkage criterion. Gallego *et al.* [16] used approximated similarity search as initial point to improve the efficiency issues at the expense of typically lowering the performance of k-nearest neighbor classifier. Ding and Song [17] developed the EM algorithm based on Gaussian copula models for the imputation of missing values. Fan and Chow [18] proposed sparse representation with missing entries and matrix completion method to deal with incomplete data subspace clustering and high-rank matrix completion. However, some of these methods usually measure the quality and diversity based on the cluster labels of base clusters while missing the information of the original data. To handle this drawback, Zhao *et al.* [19] presented a clustering ensemble selection algorithm for categorical data. Until now, there is no any clustering algorithm that performs the best for all types of data. Namely, for a given dataset, different clustering algorithms have their own strength and weakness, and cannot discover all types of cluster structures in the data. Therefore, it is difficult for users to decide which clustering algorithm would be an appropriate alternative for a given dataset [19]. Neither do these methods need the number of clusters using as input parameters, nor do they make assumptions about the underlying density or the variance within the groups which might exist in the datasets [7]. On the basis of the above analysis, our clustering algorithm is based on density method to efficiently and effectively address the issues of the complex dataset classification.

The density-based clustering approaches have gained popularity among researchers [3], [5], [7], [10], [11], [20]–[24] over the years. Density-based clustering is a nonparametric approach assuming that the points belonging to each cluster are drawn from a specific probability distribution [23]. The clusters of arbitrary shape can be discovered by the density-based methods, where the clusters are considered to be high density areas and separated from each other by contiguous regions with low density of objects [10]. Ester *et al.* [23] proposed density-based spatial clustering of applications with noise (DBSCAN), which discovers arbitrary shapes of clusters utilizing minimum domain knowledge about data. Points are classified as core objects or outliers with the density thresholds and the core objects are assigned to a cluster if they are closely packed together [10]. However, choosing an appropriate threshold can be nontrivial, and it is not fully deterministic for border points and could not perform well in overlapping densities [20]. Bai *et al.* [22] investigated a fast density clustering technique based on k-means algorithm. Rodriguez and Laio [5] developed a density peaks clustering (DPC) algorithm, which can figure out the cluster centers according to the decision graph and detect non-spherical clusters without specifying the number of clusters. Up to now, a large number of existing clustering algorithms and their variations have concerned density peaks [3], [5], [7], [10], [11], [25]–[29]. Du *et al.* [25] studied a DPC algorithm by using a similarity criterion to deal with the numerical, categorical, or mixed data. Liu *et al.* [10] proposed an adaptive clustering algorithm by introducing k-nearest neighbors to compute the global parameter and the local density of each point. However, most of all the above-mentioned algorithms are sensitive to initialization. Ding *et al.* [26] developed an entropy-based DPC algorithm for mixed type data by employing fuzzy neighborhood. Du *et al.* [27] presented density peaks clustering based on k-nearest neighbors and principal component analysis. Xu *et al.* [28] constructed an improved DPC algorithm with fast finding cluster centers, which improves the efficiency of DPC algorithm by screening points with higher local density. Jiang *et al.* [29] presented a DPC based on logistic distribution and gravitation to detect outliers and processed some datasets of varying densities and irregular shapes. However, some of the above models still have some shortcomings as follows. Euclidean distance is usually employed in the original DPC method and the extended DPC algorithms to calculate the distance between sample points; nevertheless, the Euclidean distance only considers the spatial structure of samples and does not take into account the correlation and the similarity between samples. The cutoff distance is often chosen by the minimum of 1% to 2% of the distance between all the sample points. It has a great influence on the calculation of local density. An appropriate parameter can be nontrivial and significantly effects on selecting the cluster centers and achieving the clustering results. In addition, the cluster centers are selected based on the two-dimensional decision graphs, which has certain subjective factors and causes the random error. In especial, when tackling the high-dimensional datasets, these DPC-based methods cannot yield the effective clustering results. Thus, this paper focuses on creating such a solution.

To solve these problems above, a novel adaptive DPC algorithm is investigated. First, a new local density of sample points is presented to calculate by using the Gaussian kernel density estimation function. Since the Euclidean distance easily ignores the correlation of the samples, a weighted Euclidean distance is proposed to measure the distance between samples, where the weight is the reciprocal of the absolute value of the Pearson correlation coefficient. It follows that the spatial structure of the samples and the correlation between samples are considered simultaneously.

Second, a density estimation entropy is defined to improve the density estimation parameter. The density estimation parameter is adaptively selected by minimizing the density estimation entropy, which avoids the influence of manual setting. Third, an adaptive strategy of cluster center selection is proposed to avoid the impact of manual selection and setting of parameters. Finally, to efficiently extract relevant and significant features from high-dimensional datasets and rapidly provide satisfactory clustering results, the Fisher linear discriminant (FLD) algorithm is introduced to reduce the dimensionality of the original high-dimensional datasets, and then an adaptive DPC algorithm with FLD (ADPC-FLD) is constructed to perform on many public standard datasets for obtaining the great clustering results.

The rest of this paper is structured as follows. Section II briefly reviews the basic work of the DPC algorithm. In Section III, the local density based on kernel density estimation function, the density estimation entropy and the adaptive strategy of cluster center selection are proposed, and then the ADPC-FLD algorithm is designed. The experimental results and analysis are described in Section IV. Finally, Section V summarizes this paper.

II. RELATED WORK

In 2014, clustering by fast search and find of density peaks is a new clustering method that was reported in Science [5]. It is a main density-based clustering method which is based on an assumption: The density of the cluster center is higher than the density of its neighbor sample points, and the cluster centers are generally far from each other. By calculating the distance from the nearest larger density point, the cluster centers are obtained, and the remaining points are sorted and divided into the categories according to the local density [28]. Obviously, the cluster center is a point that both the local density and the distance from the nearest larger density point are large, and then the number of cluster centers can be intuitively selected.

For each sample point, its two variables, the local density ρ_i and the distance from the nearest larger density point δ_i, are first calculated, and then for each sample point $x_i \in X = \{x_1, x_2, \cdots, x_n\}$, the local density ρ_i can be described as

$$\rho_i = \sum_{i \neq j} \chi(d_{ij} - d_c), \tag{1}$$

where d_{ij} is the distance between the samples x_i and x_j, and d_c is the cutoff distance. When $d_{ij} - d_c \leq 0$, $\chi(d_{ij} - d_c) = 1$; otherwise $\chi(d_{ij} - d_c) = 0$.

The high-density point nearest neighbor distance δ_i is expressed as

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}). \tag{2}$$

Obviously, there is no high-density nearest neighbor for sample points with high local or global density, and the distance from the nearest larger density point is simply described by

$$\delta_i = \max_j (d_{ij}). \tag{3}$$

From Eqs.(1) and (3), the cluster center is often a sample point which the local density ρ_i is large and the distance from the nearest larger density point δ_i is also large. Then, the DPC algorithm constructs the decision graphs by the ρ_i of the sample and the δ_i, selects both the larger local density ρ and the larger distance from the nearest larger density point δ of samples as the cluster centers, and identifies the other sample points into the nearest cluster centers.

III. ADAPTIVE DENSITY PEAKS CLUSTERING ALGORITHM WITH FISHER LINEAR DISCRIMINANT

For the complex and high-dimensional datasets, most of the traditional DPC algorithms are inefficient [3]. Therefore, it is very necessary to improve the DPC algorithm, with which the dimension reduction algorithm is combined to achieve the great clustering results. In this section, a new adaptive DPC algorithm with FLD is constructed well.

A. LOCAL DENSITY BASED ON KERNEL DENSITY ESTIMATION FUNCTION

The traditional DPC algorithm uses the cutoff distance to define the local density of sample points, which is only applicable to the discrete data points. For the continuous datasets, the local structural characteristics of the data are not always considered. Aiming at this problem, we introduce the Gaussian kernel density estimation function [30] to calculate the local density of each sample point, which can consider the local structural characteristics of the continuous datasets. Since the correlation of samples is easily neglected, a new weighted Euclidean distance is defined to solve this issue.

Definition 1: Based on the Gaussian kernel density estimation function, the local density formula of sample points is defined in [25] as

$$\rho_i = \sum_{i \neq j} e^{-\left(\frac{d_{ij}}{d_c}\right)^2}, \tag{4}$$

where d_{ij} is the distance between the samples x_i and x_j, usually calculating with the Euclidean distance, and d_c denotes the density estimation parameter and is equal to the window width of the kernel function.

The Pearson correlation coefficient [31] is a common criterion of the correlation between variables, and its formula is expressed as

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}, \tag{5}$$

where x and y represent two sample points, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ denotes the mean of all the feature values of the sample x, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ describes the mean of all the feature values of the sample y, and n represents the number of all the features of the sample.

From Eq. (5), the value of r_{xy} is between -1 and 1, and then one has $0 \leq |r_{xy}| \leq 1$. When $|r_{xy}| = 1$, the sample x is related to the sample y, and when $|r_{xy}| = 0$, the samples x and y are irrelevant. That is, the larger the absolute value of the correlation coefficient between the two samples is, the more relevant the two samples are, and the smaller the distance between the samples is; otherwise the less the relevance is, the greater the distance is.

Definition 2: The reciprocal of the absolute value of the Pearson correlation coefficient as the weight is introduced to redefine the distance between the two samples. Then, a weighted Euclidean distance function is defined as

$$d_{ij} = \frac{1}{|r_{ij}|} \sum (x_i - x_j)^2, \tag{6}$$

where x_i and x_j represent the features of sample data, i, $j = 1, 2, \cdots, n$ denote the number of the samples, and r_{ij} represents the Pearson correlation coefficient between the two samples.

Definition 3: Since the distance between the samples is calculated by the weighted Euclidean distance function, the formula of the local density of samples is described as

$$\rho_i = \sum_{i \neq j} e^{-\frac{\frac{1}{|r_{ij}|}\sum (x_i - x_j)^2}{{d_c}^2}}, \tag{7}$$

where $i, j = 1, 2, \cdots, n$ represent the number of all the features of the samples x and y respectively, r_{ij} denotes the Pearson correlation coefficient between the two samples, and d_c is the density estimation parameter.

Definition 4: The distance from the nearest larger density point δ_i is defined as

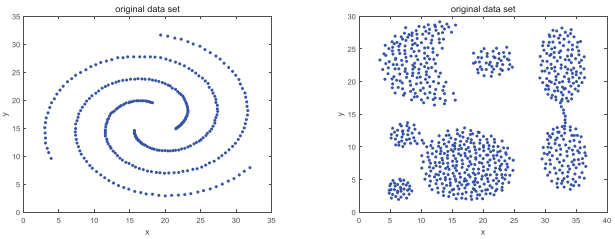
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}). \tag{8}$$

When the local density ρ_i of a sample point is the largest, $\delta_i = \max_j(d_{ij})$.

B. DENSITY ESTIMATION ENTROPY

In the traditional DPC algorithm, the cutoff distance d_c is often chosen by the minimum of 1% to 2% of the distance of all the sample points; however, in the practical experiments, it can be found that the value of the density estimation parameter d_c has a great influence on the clustering results. In order to more intuitively demonstrate the influence of d_c on the clustering results, the DPC algorithm is performed on two synthetic datasets (Spiral and Aggregation). The datasets are depicted in Fig. 1, where the Spiral dataset has 312 points around 3 clusters, and the Aggregation dataset includes 788 points around 7 clusters. Then, the clustering results of DPC on the two test datasets using the different d_c are shown in Figs. 2 and 3, respectively. Each sub-figure contains the two-dimensional decision graph and the clustering result graph. In the two-dimensional decision graphs, the abscissa is the value of ρ, and the ordinate is the distance from δ.

Fig. 2 shows that when d_c takes the minimum distance of 0.5% on the Spiral dataset, there are many sample points



(a) (b)

FIGURE 1. Description of the two synthetic datasets. (a) Spiral. (b) Aggregation.

with the small local density but the far distance from the nearest larger density point in two-dimensional decision graphs. This phenomenon leads to misclassification. Although the d_c takes the minimum distance of 2.3%, the appropriate cluster centers are displayed in Fig. 2 and they are far from the other points. Thus, the clustering effect is seriously poor, and it indicates that the d_c is seriously significant for the clustering result of the Spiral dataset. When the d_c takes the minimum distance of 1% to 2%, the correct clustering result can be obtained. The above results illustrate that the DPC method can get the results when the d_c is set to an appropriate value.

Fig. 3 shows that on the Aggregation dataset, when the d_c takes the minimum distance of 0.5%, it is difficult to determine the cluster centers from the two-dimensional decision graphs. When the d_c increases to the distance of 1%, the cluster centers can be determined well and the correct clustering result can be achieved. As the percentage of minimum distance of d_c increases, the clustering effect gradually decreases, and then many sample points are misclassified into noise. Hence, the experiments demonstrate that the traditional DPC algorithm cannot get the correct clustering results on the Aggregation dataset, and it further verifies that the d_c has a great influence on the clustering results of the Aggregation dataset. In summary, it is a challenge to set the optimal d_c for the different datasets.

To overcome this abovementioned drawback that the clustering result is very sensitive to d_c, an adaptive parameter optimization method based on the minimization of density estimation entropy is presented, and it can select an appropriate density estimation parameter d_c according to the distribution characteristics of the data and eliminates the errors of manual setting. It is well known that information entropy in information theory can effectively measure the uncertainty of random variables [32]. Inspired by the idea of information entropy, a concept of density estimation entropy is developed to optimize the density estimation parameter, and then the minimization method of density estimation entropy is used to adaptively select the density estimation parameter of the DPC algorithm.

Definition 5: Suppose that the local density of n sample points be $\rho_1, \rho_2, \cdots, \rho_n$, and the density estimation entropy is defined as

$$E = \sum_{i=1}^{n} \frac{\rho_i}{U} \log_2(\frac{1}{\rho_i} + 1), \tag{9}$$

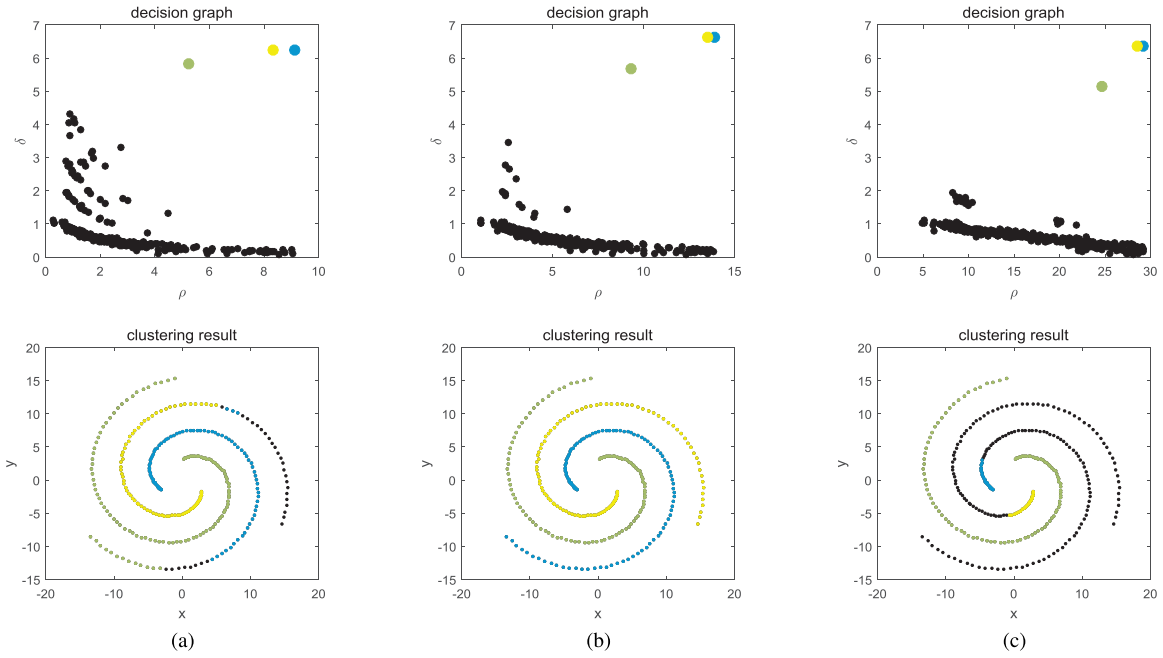


FIGURE 2. The clustering results for the different d_c on the spiral dataset. (a) $d_c = 0.9394(0.5\%)$. (b) $d_c = 1.7443(1.2\%)$. (c) $d_c = 3.7014(2.3\%)$.

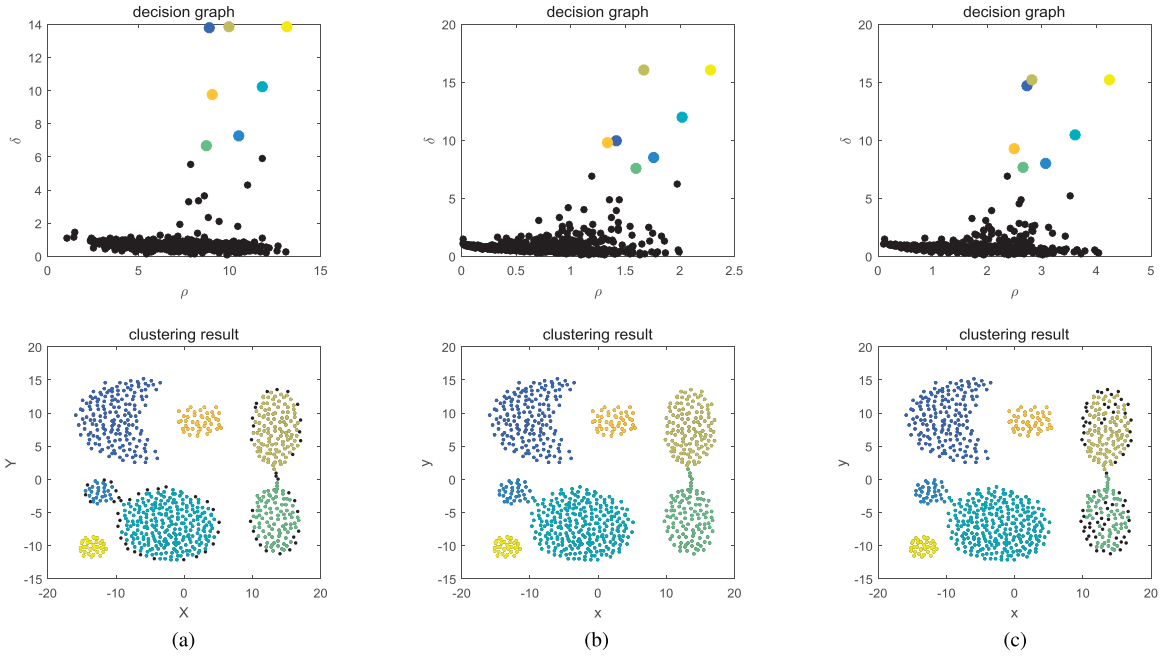


FIGURE 3. The clustering results for the different d_c on the aggregation dataset. (a) $d_c = 0.3202(0.5\%)$. (b) $d_c = 0.5099(1\%)$. (c) $d_c = 0.7018(1.6\%)$.

where ρ_i is the local density, $i = 1, 2, \cdots, n$, n denotes the number of all the sample points, and $U = \sum_{i=1}^{n} \rho_i$ represents the sum of the local density of all the sample points.

According to the distribution characteristics of the data, the local density estimation values of each sample point are different. If the local density values of each sample point are the same, it would be difficult to cluster. That is, the greater the uncertainty of the data distribution is, the larger the density estimation entropy is. When the local density of each sample point is different, the cluster centers are determined by the sample point with the high local density and then the other sample points are divided by the cluster centers. The more accurate the distribution is, the smaller the uncertainty is, and the less the density estimation entropy is.

Property 1: A range of the density estimation entropy is $0 \leq E \leq log_2(n)$.

Proof: The density estimation parameter d_c is changed from 0 to $+\infty$. From Eqs. (7) and (9), when $d_c \to 0$, E is close to $\log_2(n)$; as d_c increases, E decreases until reaching the minimum, and then gradually increases. When $d_c \to +\infty$, E is close to the maximum $\log_2(n)$. Thus, $0 \leq E \leq \log_2(n)$ can be obtained.

Definition 6: To optimize the density estimation parameter d_c*, the density estimation entropy is minimized by adjusting the* d_c*. Then, the minimization of density estimation entropy is described as*

$$\min_{d_c} E = \sum_{i=1}^{n} \frac{\rho_i}{U} \log_2(\frac{1}{\rho_i} + 1). \tag{10}$$

From Definition 6 and Property 1, by using the adaptive parameter optimization method based on Eq. (10), the optimal value of the d_c can be adaptively selected according to the distribution characteristics of the data.

C. ADAPTIVE STRATEGY OF CLUSTER CENTER SELECTION

For the DPC algorithm, the cluster centers are manually selected according to the two-dimensional decision graphs. The sample points with the large local density and distance from the nearest larger density point are selected as the cluster centers, but it has certain difficulties and subjectivity in practical application. Note that there are multiple density peaks points in a cluster. When the distribution of these points is similar, it is difficult to select the suitable number of cluster centers. This phenomenon can be easily observed from Fig. 3(a). To eliminate the error and the difficulty of selecting cluster centers manually, an adaptive strategy of cluster center selection based on the distribution of the data and the local density is developed.

Since the cluster centers should reflect the distribution of all data and contain useful information as much as possible, according to the local density and the distance from the nearest larger density point, the optimized adaptive strategy of cluster center selection is described as follows. On the one hand, the standard deviation of the distance from the nearest larger density point of all data is calculated, which is used as a measure of the statistical distribution and reflects the dispersion degree of the data to some extent. Then, it is considered to select sample as a cluster center, whose distance from the nearest larger density point is greater than or equal to the weighted standard deviation. On the other hand, in some cases, the DPC algorithm discovers the noises with the large distance from the nearest larger density point but small local density, which needs to be excluded from the cluster centers. Thus, the data points that the local density is greater than the mean of the local density of all data are considered as the cluster centers. By combining the above two steps, the conditions for selecting the cluster centers are described as

$$EC = \delta_i \geq \lambda\sigma(\delta_i), \tag{11}$$

$$RC = EC(\rho_i) \geq \mu(\rho_i), \tag{12}$$

where EC represents the expected cluster centers, δ_i is the distance from the nearest larger density point, $\sigma(\delta_i)$ is the standard deviation of the distance from the nearest larger density point of all data, λ is the weight, RC denotes the cluster centers after removing noises, ρ_i is the local density of each sample points, $EC(\rho_i)$ describes the local density of the expected cluster centers, and $\mu(\rho_i)$ is the mean of all the local densities.

The cluster centers selected by the two conditions above simultaneously ensure the large distance from the nearest larger density point and the large local density, which can avoid the errors caused by selecting the noises as cluster centers and guarantee the objectivity of clustering results.

D. DESCRIPTION OF ADPC-FLD ALGORITHM

To solve the problem that the DPC algorithm is difficult to deal with the large-scale and high-dimensional datasets, the dimension reduction method needs to be firstly employed to reduce the dimensionality of high-dimensional data. In recent years, the dimension reduction methods roughly include linear mapping and nonlinear mapping methods. The Fisher linear discriminant (FLD) method [33] is widely used in dimension reduction, which processes data by linear function values. Its basic idea is to project the sample onto a straight line by transforming the sample so that the projection of the sample can be best divided. Thus, FLD can select the features with the information classified, effectively eliminates the irrelevant and redundant features, and achieves the reduced dimensionality for high-dimensional datasets [34].

In what follows, the FLD is used to reduce the dimensionality of the high-dimensional datasets, and then an adaptive density peaks clustering (ADPC) algorithm is performed. For convenience of description, the ADPC algorithm with the FLD method (ADPC-FLD) is divided into two sub-algorithms: the dimension reduction algorithm based on FLD (Algorithm 1) and the ADPC algorithm (Algorithm 2). The special procedures of the ADPC-FLD algorithm are illustrated in Fig. 4.

The specific steps of the ADPC-FLD algorithm are described as Algorithms 1 and 2.

Suppose that the dataset has n sample points, and the number of cluster centers is m. For implementing the ADPC-FLD algorithm, the computational process of the distance matrix largely affects the time complexity and the space complexity of our cluster algorithm. Since the space complexity of the distance matrix is $O(n^2)$ [10], the time complexity of adjusting the density estimation parameter is $O(n)$, the complexity of calculating the local density is $O(n^2)$ and that of the distance from the nearest larger density point is $O(n^2)$, the overall time complexity of ADPC-FLD is $O(n^2)$, which mainly depends on calculating the distance matrix and is the same as that of DPC. Furthermore, the space complexity of ADPC-FLD is $O(n^2)$.

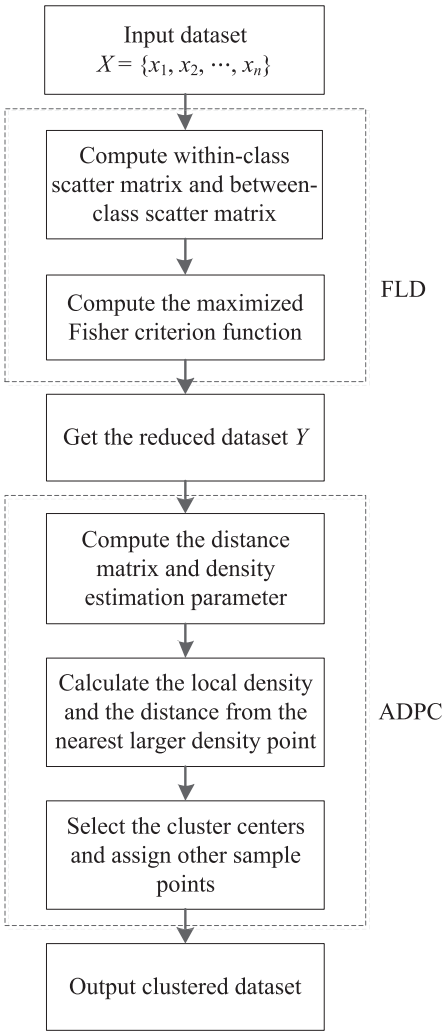


FIGURE 4. The flow chart of the ADPC-FLD algorithm.

Algorithm 1

Input: A dataset $X = \{x_1, x_2, \cdots, x_n\}$

Output: The reduced dataset Y

Step 1: Calculate the class mean vector of each class by

$$\mu_i = \frac{1}{N_i} \sum_{x_j \in X_i} x_j,$$

and then obtain the within-class scatter matrix in the sample space of the original dataset as follows:

$$S_i = \sum_{x_j \in X_i} (x_j - \mu_i)(x_j - \mu_i)^T,$$

where $i = 1, 2, \cdots, m$.

Step 2: Calculate the within-class scatter matrix by

$$S_W = \sum_{i=1}^{m} \sum_{x_j \in X_i} (x_j - \mu_i)(x_j - \mu_i)^T,$$

and compute the between-class scatter matrix by

$$S_B = \sum_{i=1}^{m} n_i(\mu_i - \mu)(\mu_i - \mu)^T.$$

Step 3: Calculate the maximized Fisher criterion function by

$$\max J_F(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_W \omega},$$

and use the Lagrange multiplier to solve the unconstrained extreme problem of the Lagrangian function. Then, one can obtain the optimal projection based on the Fisher discriminant criterion.

Step 4: Calculate $Y = \omega^T X$ to get the reduced dataset Y.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENT PREPARATION

To verify the clustering performance of the proposed ADPC-FLD algorithm, the experiment is divided into the following three parts: The effectiveness of the improved ADPC algorithm is validated on six synthetic datasets in terms of two-dimensional decision graph and three indices. Then, the clustering results of the ADPC-FLD algorithm are compared with those of the other related clustering algorithms on thirteen standard UCI datasets in terms of two-dimensional decision graph and seven indices. For the final part, the testing accuracy of the ADPC-FLD algorithm is evaluated on seven gene expression datasets in terms of two-dimensional decision graph and three indices. All the numerical experiments are implemented on a personal computer running Windows 7 with an Intel(R) Core(TM) i5-3470 CPU operating at 3.2 GHz and 8 GB of memory.

To effectively evaluate the clustering results of all the contrast algorithms, eleven kinds of evaluation indices [35]–[37], including the cluster number (*CN*), the silhouette index (*Sil*), the adjusted mutual information (*AMI*), the precision (*P*), the recall (*R*), the specificity (*S*), the accuracy (*AC*), the Jaccard coefficient (*JC*), the Fowlkes-Mallows index (*FMI*), the F-measure index (*FM*) and the adjusted rand index (*ARI*), are introduced to illustrate the effectiveness and efficiency of our algorithm. These indices are described as follows.

The *Sil* is denoted as

$$Sil(t) = \frac{[b(t) - a(t)]}{\max\{a(t), b(t)\}}, \tag{13}$$

where t denotes the number of all the samples of a dataset, $t = 1, 2, \cdots, n$, $a(t)$ is the average dissimilarity of t to all the other samples in a cluster C_i, $b(t) = min\{d(t, C_i)\}$, $d(t, C_i)$ is the average dissimilarity of t in C_j to all samples in another C_i, and $i, j = 1, 2, \cdots, k$ with $i \neq j$.

Suppose that $U = \{U_1, U_2, \cdots, U_R\}$ and $V = \{V_1, V_2, \cdots, V_C\}$ denote the true division and the division of the clustering result on a dataset $X = \{x_1, x_2, \cdots, x_n\}$, respectively. Then, the *AMI* is expressed as

$$AMI(U, V) = \frac{MI(U, V) - E\{MI(U, V)\}}{\max\{H(U), H(V)\} - E\{MI(U, V)\}}, \tag{14}$$

Algorithm 2

Input: The reduced dataset Y using Algorithm 1

Output: The clustering results

Step 1: Calculate the distance matrix $D = \{d_{ij}\}$ of each sample in Y with Eq. (6), where $i, j = 1, 2, \cdots, n$.

Step 2: Adjust the density estimation parameter d_c with Eq. (10).

Step 3: Calculate the local density ρ_i of each sample according to the D and the d_c in Eq. (7).

Step 4: Calculate the distance from the nearest larger density point δ_i of each sample according to the ρ_i in Eq. (8).

Step 5: Construct a two-dimensional decision graph by using the ρ_i and the distance from the δ_i.

Step 6: Select the cluster centers with Eqs. (11) and (12).

Step 7: Calculate the minimum distance of the other sample points and the cluster centers, and assign the other sample points into the nearest cluster center.

where $H(U)$ is the entropy of the original partition, $H(V)$ is the entropy of the partition of clustering results, $MI(U, V)$ is the mutual information between U and V, and $E\{MI(U, V)\}$ is the expected mutual information between U and V.

Suppose that a denotes the number of samples in the same class of U and V simultaneously; b represents the same class in U, but the different class in V; c denotes the different class in U, but the same class in V; and d denotes the number of samples in the different classes of U and V. Then, the indices are expressed as follows.

$$P = \frac{a}{a+b}, \tag{15}$$

$$R = \frac{a}{a+c}, \tag{16}$$

$$S = \frac{d}{b+d}, \tag{17}$$

$$AC = \frac{a+d}{a+b+c+d}, \tag{18}$$

$$JC = \frac{a}{a+b+c}, \tag{19}$$

$$FMI = \sqrt{\frac{a^2}{(a+b)(a+c)}}, \tag{20}$$

$$FM = \frac{2PR}{P+R}, \tag{21}$$

$$ARI = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}. \tag{22}$$

B. COMPARISONS OF CLUSTERING RESULTS ON SYNTHETIC DATASETS

In this subsection, our proposed ADPC algorithm is compared with the traditional DPC method [5] and the DBSCAN algorithm [23] on the six synthetic datasets selected from [5], [7], [11], which are widely used in the clustering algorithms. The description of the six synthetic datasets is shown in Table 1. The distributions of the sample points for each synthetic dataset are shown in Fig. 5.

TABLE 1. Description of the six synthetic datasets.

No.	Datasets	Samples	Features	Clusters
1	Spiral	312	2	3
2	4k2_far	400	2	4
3	R15	600	2	15
4	Aggregation	788	2	7
5	D31	3100	2	31
6	S1	5000	2	15

To determine the optimal parameters of the compared algorithms and obtain the more accurate performance, the following several experiments are performed on the six synthetic datasets. For the ADPC algorithm, the optimal d_c under the proposed density estimation entropy is adaptively selected. The density estimation parameter d_c of the traditional DPC algorithm is set to the minimum of 1% to 2% of the distance of all sample points. The DBSCAN has two parameters *eps* and *minpts*, where the *eps* is looped from 0.01 to 1 with a step size of 0.01, and the *minpts* is looped from 1 to 50. The optimal parameters involved in the compared algorithms are described in Table 2. The two-dimensional decision graphs of the ADPC algorithm on the six synthetic datasets are illustrated in Fig. 6, and the graphs of clustering with the three algorithms on the six synthetic datasets are shown in Figs. 7 to 12, respectively.

TABLE 2. The adjusted parameters for the three algorithms on the six synthetic datasets.

Datasets	ADPC	DPC	DBSCAN	
	d_c	d_c	*minpts*	*eps*
Spiral	2.5812	1.7	2	0.04
4k2_far	0.2699	0.154	2	0.58
R15	0.3694	0.2016	8	0.3
Aggregation	0.5025	0.495	6	0.04
D31	0.4712	0.1893	30	0.28
S1	0.1235	0.1639	17	0.62

It can be seen from Fig. 6 that the ADPC algorithm can select the appropriate cluster centers and the accurate number of clusters on the six synthetic datasets. From Figs. 7 to 12, on the Spiral and 4k2_far datasets, all the three compared algorithms can obtain the correct number of clusters and the good performance. On the R15 and Aggregation datasets, ADPC can obtain the correct number of clusters, but both DPC and DBSCAN cannot, which may lead to misclassification. On the D31 and S1 datasets, the ADPC algorithm still performs the great performance and obtains the correct number of clusters. Although the DPC algorithm can obtains the correct number of clusters, some sample points are misclassified to the other clusters. The DBSCAN algorithm mistakenly divides some sample points into noises, resulting in the poor clustering effect. In order to further illustrate the clustering performance of ADPC, the three indices (*AC*, *JC*, and *FMI*) of clustering results are employed to test the clustering performance. To ensure the objectivity of the clustering results and reduce the random errors, each method is

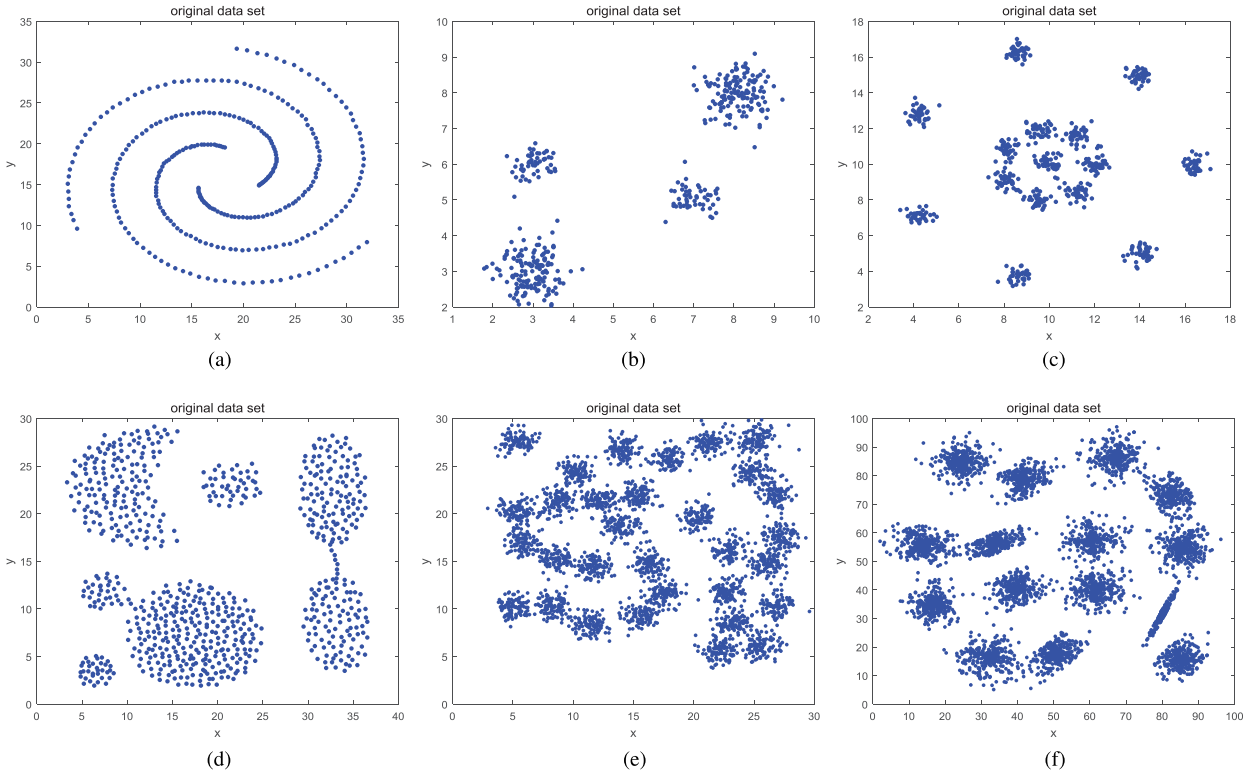


FIGURE 5. Distributions of the sample points for the six synthetic datasets. (a) Spiral. (b) 4k2_far. (c) R15. (d) Aggregation. (e) D31. (f) S1.

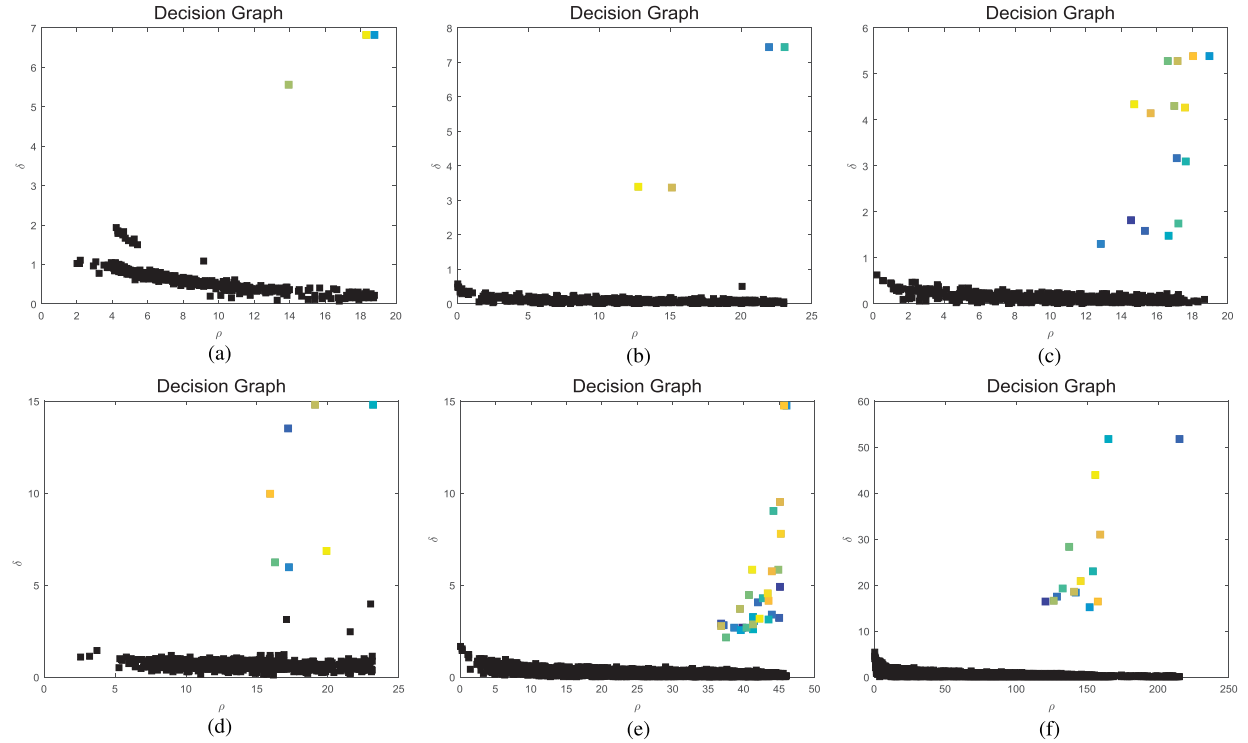


FIGURE 6. The two-dimensional decision graphs of the ADPC algorithm on the six synthetic datasets. (a) Spiral. (b) 4k2_far. (c) R15. (d) Aggregation. (e) D31. (f) S1.

run 10 times, and the results are the mean of the 10 evaluations of the cluster indices. The experimental results are illustrated in Table 3.

Table 3 shows the evaluated clustering results of the three compared algorithms on the six synthetic datasets. As can be seen from Table 3, the Spiral and 4k2_far datasets have

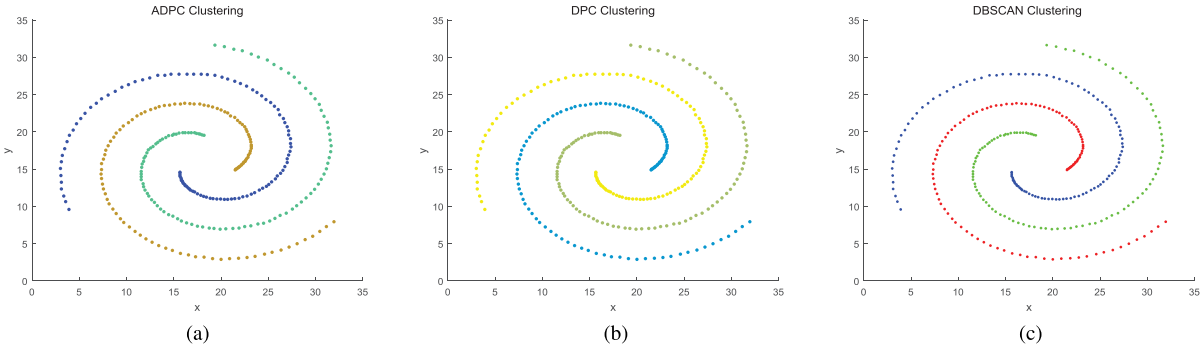


FIGURE 7. The graphs of clustering with the three algorithms on the Spiral dataset. (a) ADPC. (b) DPC. (c) DBSCAN.

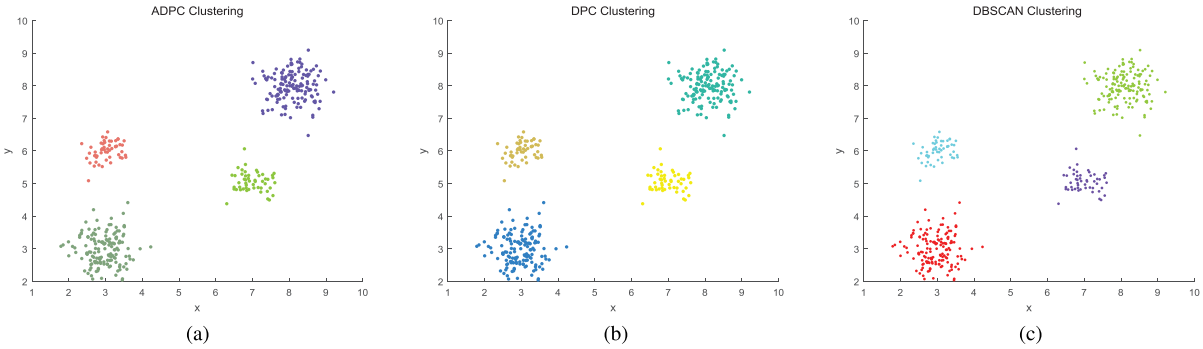


FIGURE 8. The graphs of clustering with the three algorithms on the 4k2_far dataset. (a) ADPC. (b) DPC. (c) DBSCAN.

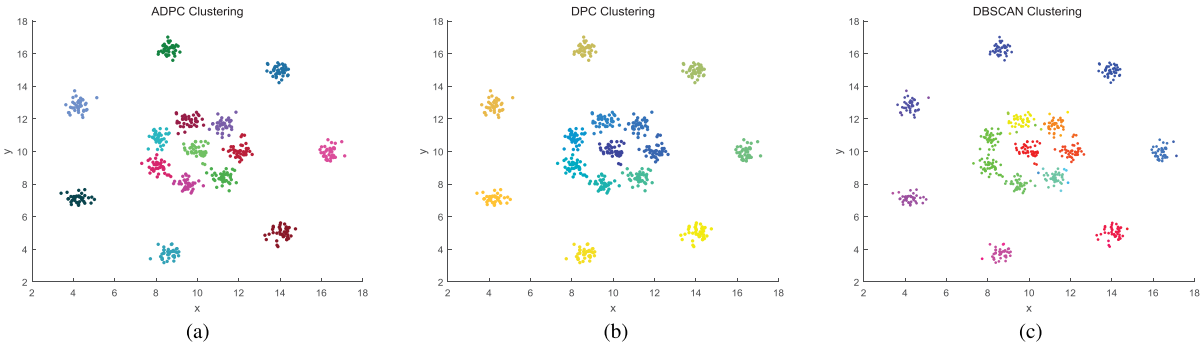


FIGURE 9. The graphs of clustering with the three algorithms on the R15 dataset. (a) ADPC. (b) DPC. (c) DBSCAN.

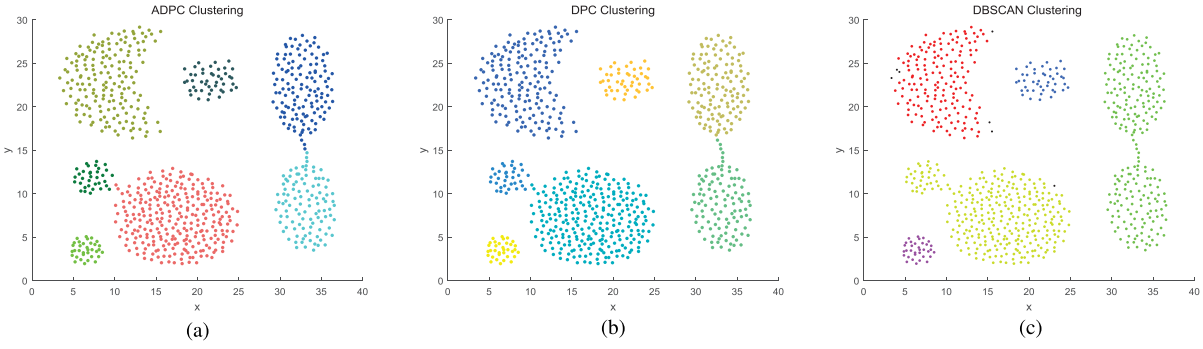


FIGURE 10. The graphs of clustering with the three algorithms on the Aggregation dataset. (a) ADPC. (b) DPC. (c) DBSCAN.

the characteristics of the small in-cluster distance and the large between-cluster distance, and the three index values of the three algorithms on the two datasets are all equal to 1, which indicates that all the three algorithms perform well. On the R15 dataset, the three evaluation indices of ADPC is 1, but due to the misclassification, the three indices of DPC and DBSCAN are smaller than that of ADPC. So, the results demonstrate that ADPC performs better than the other

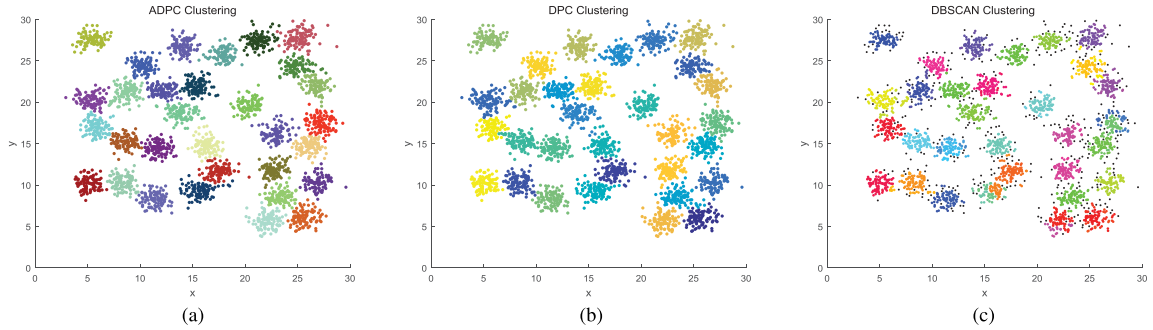


FIGURE 11. The graphs of clustering with the three algorithms on the D31 dataset. (a) ADPC. (b) DPC. (c) DBSCAN.

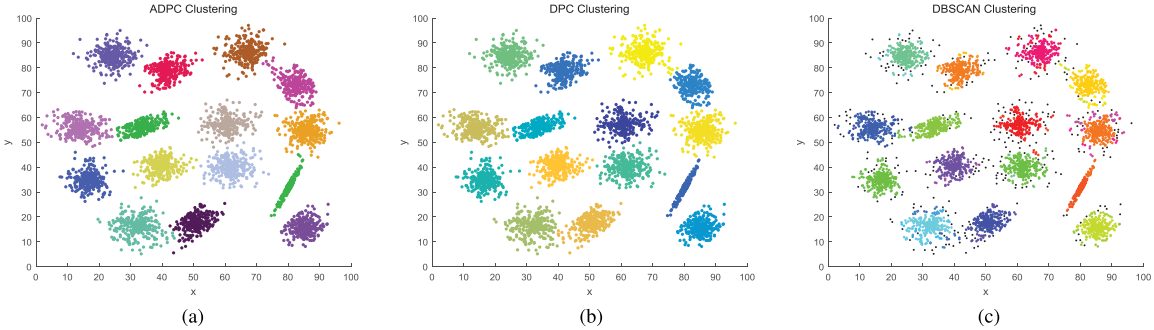


FIGURE 12. The graphs of clustering with the three algorithms on the S1 dataset. (a) ADPC. (b) DPC. (c) DBSCAN.

TABLE 3. The clustering results of the three algorithms on the six synthetic datasets.

Datasets	ADPC			DPC			DBSCAN		
	AC	*JC*	*FMI*	*AC*	*JC*	*FMI*	*AC*	*JC*	*FMI*
Spiral	1	1	1	1	1	1	1	1	1
4k2_far	1	1	1	1	1	1	1	1	1
R15	1	1	1	0.9987	0.9802	0.99	0.9905	0.8722	0.9318
Aggregation	0.9978	0.9898	0.9949	0.9628	0.8492	0.9185	0.9534	0.8229	0.9029
D31	0.9961	0.8851	0.94	0.9941	0.8326	0.9087	0.9908	0.7599	0.8636
S1	0.9994	0.9938	0.9969	0.9916	0.8881	0.9407	0.9463	0.7019	0.8248

two methods. On the Aggregation, D31 and S1 datasets, all the indices of ADPC are close to 1 and higher than those of DPC and DBSCAN. Therefore, it can be concluded that our proposed ADPC algorithm has better effectiveness and accuracy on all the two-dimensional synthetic datasets.

C. COMPARISONS OF CLUSTERING RESULTS ON STANDARD UCI DATASETS

This portion of the experiments is to test the feasibility and efficiency of ADPC-FLD on low dimensional UCI datasets. As we know, the experiment results on the real-world datasets with low-dimensionality for practical problems that are usually employed to evaluate the clustering performances of the clustering algorithms [38]. Then, to make the experiments more adequate, thirteen standard UCI datasets are downloaded from the UCI repository of machine learning databases (http://www.ics.uci.edu), and divided into two categories including ten general UCI datasets and three real-world imbalanced UCI datasets. The description of the ten general UCI datasets is shown in Table 4.

TABLE 4. Description of the ten UCI datasets.

No.	Datasets	Samples	Attributes	Clusters
1	Iris	150	4	3
2	Seeds	210	7	3
3	Ecoil	336	8	8
4	Wine	178	13	3
5	Dermatology	366	33	6
6	Segmentation	2310	19	7
7	Zoo	101	18	7
8	Pima	768	8	2
9	Chess	3196	36	2
10	Spambase	4601	57	2

The first part of this experiment is to verify the clustering results of ADPC-FLD in terms of the two-dimensional decision graphs on the six UCI datasets selected from Table 4. The two-dimensional decision graphs of ADPC-FLD on the six UCI are shown in Fig. 13. It can be seen from Fig. 13(a)-(e) that the ADPC-FLD algorithm can accurately select the cluster centers and the cluster number on the five datasets (Iris, Seeds, Ecoil, Wine and Dermatology), except for the Segmentation dataset. Since it is difficult to distinguish the

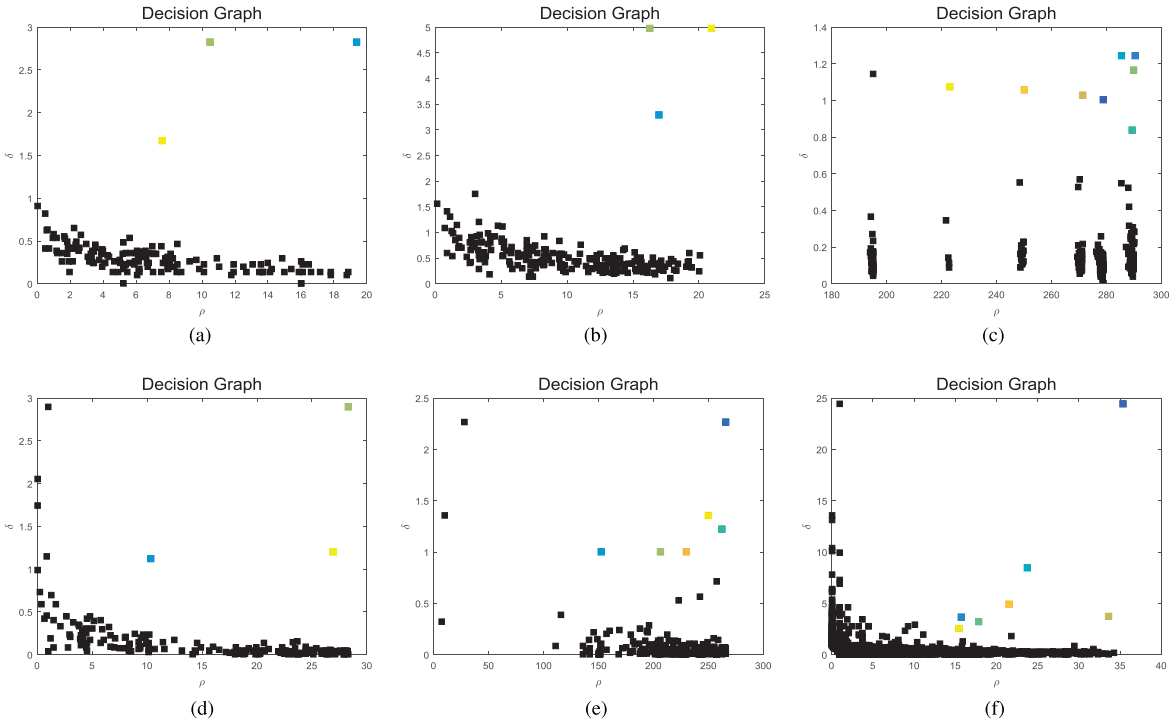


FIGURE 13. The two-dimensional decision graphs of the ADPC-FLD algorithm on the six UCI datasets. (a) Iris. (b) Seeds. (c) Ecoil. (d) Wine. (e) Dermatology. (f) Segmentation.

cluster centers and the other sample points from the two-dimensional decision graphs on the Segmentation dataset, ADPC-FLD shows the poor performance; however, our proposed algorithm can still select the correct number of clusters. In summary, the proposed ADPC-FLD algorithm has a good effect on selecting the cluster centers and the number of clusters.

The following part of this experiment continues testing the ADPC-FLD algorithm on the same six standard UCI datasets in terms of the three indices (*FMI*, *AMI* and *ARI*). ADPC-FLD is compared with five state-of-the-art clustering algorithms, which include: (1) the shared nearest neighbor-based clustering by fast search and find of density peaks algorithm (SNN-DPC) [39], (2) the fuzzy weighted k-nearest neighbors density peaks clustering algorithm (FKNN-DPC) [3], (3) the DPC algorithm [5], (4) the DBSCAN algorithm [23], and (5) the ordering points to identify the clustering structure algorithm (OPTICS) [21]. To ensure the objective of clustering results, the numerous experiments are performed to select the optimal parameters of all the compared algorithms. For the ADPC-FLD algorithm, the optimal d_c is adaptively selected by the proposed density estimation entropy. For the SNN-DPC and FKNN-DPC algorithms, the parameters are designed by Liu et al. [39]. For the traditional DPC algorithm, the value of the density estimation parameter d_c is set to the minimum of 1% to 2% of the distance of all sample points. For the DBSCAN and OPTICS algorithms, the *eps* is looped from 0.01 to 1 with a step size of 0.01, and the *minpts* is looped from 1 to 50. The optimal parameters involved in each compared algorithms are described in Table 5 in detail. Similar to this previous subsection, the six methods are executed 10 times, and the results of *FMI*, *AMI* and *ARI* are the mean of 10 clustering operations. The experimental results are shown in Tables 6 to 8.

Table 6 shows the *FMI* index of the six tested algorithms on the six general UCI datasets. As can be seen from Table 6, the ADPC-FLD algorithm obtains the best clustering accuracy than the other five algorithms on the Iris, Seeds, Ecoil, Wine and Dermatology datasets, where the *FMI* values are 1 on the Iris and Wine datasets. On the Iris and Seeds datasets, the performances of the three DPC algorithms (SNN-DPC, FKNN-DPC and DPC) are similar, and their clustering accuracies are higher than those of the DBSCAN and OPTICS algorithms. Compared with the other five clustering algorithms, the *FMI* index of DPC is the smallest on the Ecoil dataset, which indicates that the algorithm has the worst performance. The *FMI* values of the six algorithms on the Wine and Dermatology datasets further demonstrate that the four DPC algorithms (ADPC-FLD, SNN-DPC, FKNN-DPC and DPC) are superior to the two density-based clustering algorithms (DBSCAN and OPTICS). Since the difference of data categories is not obvious on the Segmentation dataset, the clustering accuracies of the six algorithms are slightly bad. Nevertheless, the ADPC-FLD algorithm still has the largest *FMI* value. In summary, the comparisons of the *FMI* index indicate that ADPC-FLD has the great clustering

TABLE 5. The adjusted parameters for the six algorithms on the six UCI datasets.

Datasets	ADPC-FLD	SNN-DPC	FKNN-DPC	DPC	DBSCAN		OPTICS	
	d_c	K	K	d_c	eps	$minpts$	eps	$minpts$
Iris	0.4472	15	22	0.2	0.12	5	0.15	5
Seeds	1.0226	6	9	0.7	0.24	16	0.81	5
Ecoil	0.1661	6	9	0.4	0.2	22	0.23	29
Wine	0.384	18	9	2	0.5	21	0.59	7
Dermatology	1.1426	19	35	1.6	0.99	3	0.99	1
Segmentation	0.4998	7	27	1.5	0.15	2	0.15	1

TABLE 6. The *FMI* index for the six algorithms on the six UCI datasets.

Datasets	ADPC-FLD	SNN-DPC	FKNN-DPC	DPC	DBSCAN	OPTICS
Iris	1	0.9479	0.9355	0.9233	0.7291	0.7868
Seeds	0.9625	0.8589	0.8276	0.8444	0.6711	0.635
Ecoil	0.9632	0.8243	0.6919	0.5775	0.6623	0.7515
Wine	1	0.933	0.8667	0.7835	0.7121	0.6296
Dermatology	0.9602	0.9021	0.8504	0.8221	0.5395	0.4563
Segmentation	0.76	0.6457	0.5581	0.673	0.5277	0.5361

TABLE 7. The *AMI* index for the six algorithms on the six UCI datasets.

Datasets	ADPC-FLD	SNN-DPC	FKNN-DPC	DPC	DBSCAN	OPTICS
Iris	1	0.9124	0.8831	0.8606	0.5692	0.4513
Seeds	0.933	0.7509	0.6971	0.7299	0.5302	0.3802
Ecoil	0.9258	0.6711	0.4755	0.4978	0.4516	0.426
Wine	1	0.8735	0.8038	0.7065	0.5484	0.3698
Dermatology	0.9601	0.8749	0.8355	0.784	0.5721	0.2934
Segmentation	0.8059	0.6725	0.583	0.6927	0.4965	0.4312

TABLE 8. The *ARI* index for the six algorithms on the six UCI datasets.

Datasets	ADPC-FLD	SNN-DPC	FKNN-DPC	DPC	DBSCAN	OPTICS
Iris	1	0.9222	0.9038	0.8857	0.612	0.6886
Seeds	0.944	0.789	0.7422	0.767	0.5291	0.419
Ecoil	0.9479	0.7547	0.5535	0.4465	0.5255	0.6642
Wine	1	0.8992	0.799	0.6724	0.5292	0.4119
Dermatology	0.9492	0.8689	0.8127	0.776	0.4165	0.343
Segmentation	0.7143	0.577	0.4367	0.6004	0.4543	0.46

performance on the six general UCI datasets than the other five clustering algorithms.

Table 7 shows the *AMI* index of the six tested algorithms for the six UCI datasets. As can be seen from Table 7, the four DPC algorithms (ADPC-FLD, SNN-DPC, ADPC-KNN, and DPC) perform generally better than the two density-based clustering algorithms (DBSCAN and OPTICS), especially our proposed ADPC-FLD algorithm. The values of the *AMI* index of ADPC-FLD are much larger than those of the other five methods on the six UCI datasets. Taking the traditional DPC algorithm as an example, the *AMI* values of ADPC-FLD on the six datasets are 0.14, 0.21, 0.42, 0.29, 0.18, and 0.11 higher than those of DPC, respectively. On the Iris and Wine datasets, the ADPC-FLD algorithm shows the highest *AMI* value, which indicates that ADPC-FLD has the best clustering effect on the two datasets. For the Seeds, Ecoil and Dermatology datasets, the proposed algorithm has the better clustering performance, and its *AMI* values are 0.93, 0.92 and 0.96, respectively. The OPTICS algorithm has the worst performance, because the *AMI* values are only 0.38, 0.43 and 0.29. The six algorithms have the smaller *AMI* values on the Segmentation dataset, but ADPC-FLD still shows the highest clustering accuracy. Therefore, all the *AMI* values demonstrate that ADPC-FLD has the better clustering performance on the six UCI datasets than the other five clustering algorithms.

Table 8 shows the *ARI* values of the six compared algorithms for the six UCI datasets. It can be observed from Table 8 that the *ARI* values of ADPC-FLD are significantly larger than those of the other five methods on the six UCI datasets. On the Iris and Seeds datasets, the clustering results of ADPC-FLD, SNN-DPC, FKNN-DPC and DPC are superior to DBSCAN and OPTICS. Meanwhile, the SNN-DPC, FKNN-DPC and DPC algorithms have the similar *ARI* on the Seeds dataset. However, for the Ecoil dataset, the FKNN-DPC, DPC and DBSCAN algorithms have the similar and poor performance in the *ARI* index, and the SNN-DPC and OPTICS algorithms demonstrate the better performance than the above three algorithms. The ADPC-FLD algorithm performs better than the other five clustering algorithms, and

TABLE 9. The adjusted parameters for the six algorithms on the four UCI datasets.

Datasets	ADPC-FLD	ADPC-KNN	DPC	DBSCAN		*K*-Means++	Single-link
	d_c	K	d_c	*eps*	*minpts*	*CN*	*CN*
Zoo	1	5	0.1	2.3	3	7	7
Pima	5.4074	22	0.03	1.4	4	2	2
Chess	1.7321	25	0.38	2.4	10	2	2
Spambase	3.3122	1000	0.01	17.3	1	2	2

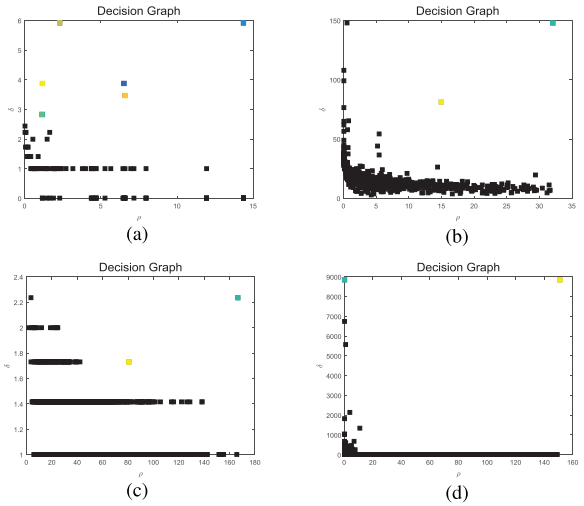


FIGURE 14. The two-dimensional decision graphs of the ADPC-FLD algorithm on the four UCI datasets. (a) Zoo. (b) Pima. (c) Chess. (d) Spambase.

its *ARI* values are much larger than those of the other five algorithms. On the Wine and Dermatology datasets, the *ARI* index of ADPC-FLD and SNN-DPC are significantly larger than the other four methods, and the performance of OPTICS is the worst. The performance of the six algorithms on the Segmentation dataset is still bad, but the clustering accuracy of ADPC-FLD is the highest, and reaches 0.7143. In general, the *ARI* values certify the validity and the clustering accuracy of our ADPC-FLD algorithm on all the six UCI datasets.

The third part of this experiment is to further evaluate the ADPC-FLD algorithm on the four general UCI datasets (Zoo, Pima, Chess and Spambase) selected from Table 4. The five state-of-the-art algorithms include: (1) the adaptive density peak clustering based on k-nearest neighbors algorithm (ADPC-KNN) [10], (2) the DPC algorithm [5], (3) the DBSCAN algorithm [23], (4) the K-Means++ algorithm [40], and (5) the Single-link algorithm [41]. In order to ensure the objective results and illustrate the best performance of all the compared algorithms, the parameters of each algorithm should be adjusted. The same as this previous portion, the optimal d_c of ADPC-FLD is adaptively selected by the proposed density estimation entropy, and then the two-dimensional decision graphs of the four UCI datasets, which is determined based on d_c, is shown in Fig. 14. For the ADPC-KNN and DBSCAN algorithms, the selection of parameters is designed by Liu et al. [10]. For the K-Means++ and Single-link algorithms, the parameter K is set as the real cluster number of each dataset. The optimal adjusted parameters of all the compared algorithms are denoted in Table 9. To verify the clustering effects of the six algorithms on the four UCI datasets, the experiments use the two indices (*CN* and *FM*) to evaluate the clustering results. Similarly, the results are also the mean of 10 evaluations of clustering accuracy to ensure the objectivity of the experimental results. The experimental results are shown in Table 10.

Fig. 14 shows the two-dimensional decision graphs of the ADPC-FLD algorithm on the four UCI datasets (Zoo, Pima, Chess and Spambase), where the color markers are the selected cluster centers. Table 10 states the comparisons of *CN* and *FM* with the six algorithms on the four UCI datasets. It can be seen from Fig. 14 and Table 10 that ADPC-FLD can select the cluster centers and determine the correct number of clusters. From Table 10, the ADPC-FLD, ADPC-KNN, DPC and DBSCAN algorithms obtains the number of correct clusters, but the cluster number of K-Means++ and Single-link algorithms is set as the real cluster number in advance. Although all algorithms can correctly identify the number of clusters, there are major differences in clustering accuracy. From the overall perspective, the clustering accuracies of the six algorithms on the Zoo dataset are higher than those on the other three datasets, where the accuracies of the ADPC-FLD and ADPC-KNN algorithms on the Zoo dataset are similar and higher than those of the other four algorithms. For the Pima dataset, the ADPC-FLD algorithm compared with the other five algorithms achieves the highest accuracy, and its *FM* value is 0.7. The ADPC-KNN performs as poorly as the Single-link. For the Chess dataset, the *FM* values of ADPC-FLD and Single-link are the same and higher than those of the other four algorithms. Thus, this shows that the clustering performance of ADPC-FLD and Single-link is better than the others, where the K-Means++ performs the worst result. For the Spambase dataset, the clustering accuracies of the six algorithms are flat. The *FM* values of the four algorithms (ADPC-KNN, DBSCAN, K-Means++ and Single-link) are the same and slightly lower than that of ADPC-FLD, and slightly higher than that of DPC. Hence, it is obvious that ADPC-FLD is effective. In summary, the results of the *CN* and *FM* indices on the four UCI datasets demonstrate that our ADPC-FLD algorithm outperforms the other five algorithms in most cases.

The fourth part of this experiment is to give more justifications of our ADPC-FLD algorithm in terms of the three indices (*CN*, *Sil*, and *FM*) by testing the Iris, Wine and Seeds datasets, selected from Table 4. The ADPC-FLD algorithm is compared with four state-of-the-art methods,

TABLE 10. The clustering results for the six algorithms on the four UCI datasets.

Datasets	Indices	ADPC-FLD	ADPC-KNN	DPC	DBSCAN	*K*-Means++	Single-link
Zoo	*CN*	7	7	7	7	7	7
	FM	0.9	0.87	0.69	0.7	0.82	0.65
Pima	*CN*	2	2	2	2	2	2
	FM	0.7	0.69	0.53	0.66	0.64	0.69
Chess	*CN*	2	2	2	2	2	2
	FM	0.67	0.65	0.65	0.61	0.51	0.67
Spambase	*CN*	2	2	2	2	2	2
	FM	0.69	0.68	0.67	0.68	0.68	0.68

which include: (1) the affinity propagation clustering algorithm using hybrid kernel function with locally linear embedding (HKAP-LLE) [38], (2) the adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity (SAAP-SS) [42], (3) the fireworks explosion optimization-based semi-supervised affinity propagation algorithm (FEO-SAP) [43], and (4) the affinity propagation (AP) algorithm [44]. In order to obtain the objective comparison, the optimal d_c of ADPC-FLD still is adaptively selected by the density estimation entropy, and the parameters of the other four algorithms are set by following the techniques designed by Sun et al. [38]. Here, alike to the above calculated results, all the compared methods need to be run 10 times, and then the clustering results are the mean of the 10 evaluations. The experimental results are shown in Table 11.

TABLE 11. The clustering results for the five algorithms on the three UCI datasets.

Datasets	Indices	ADPC-FLD	HKAP-LLE	SAAP-SS	FEO-SAP	AP
Iris	*CN*	3	3	3	3	12
	Sil	0.7478	0.6186	0.5233	0.5135	0.3848
	FM	1	0.95	0.94	0.9	0.84
Wine	*CN*	3	3	3	3	12
	Sil	0.6475	0.6223	0.589	0.559	0.368
	FM	1	0.91	0.89	0.84	0.72
Seeds	*CN*	3	3	3	3	17
	Sil	0.6669	0.5997	0.5424	0.4346	0.328
	FM	0.96	0.94	0.93	0.86	0.79

Table 11 shows the *CN*, *Sil* and *FM* values of the five clustering algorithms on the Iris, Wine and Seeds datasets. It can be observed from Table 11 that the ADPC-FLD, HKAP-LLE, SAAP-SS and FEO-SAP algorithms perform better in the number of clusters, but the *CN* values of the AP algorithm are far more than the real cluster number of the datasets, which leads to the poor accuracy. The *Sil* values of the five compared algorithms decrease on the three datasets in turn, where ADPC-FLD achieves the largest *Sil* value, but AP yields the smallest result. It indicates that the clustering performance of ADPC-FLD is the best, and AP has the worst performance. For the *FM* index, the ADPC-FLD algorithm has the largest values on the three UCI datasets, and these values of HKAP-LLE and SAAP-SS are similar and slightly lower than ADPC-FLD. The AP algorithm still performs the worst in terms of *FM*. In summary, the ADPC-FLD algorithm shows the great performance than the other four algorithms on the three general UCI datasets.

The last part of this experiment is to verify the effectiveness of the proposed ADPC-FLD algorithm on real-world imbalanced UCI datasets. Note that in many practical applications, imbalance occurs when a negative class contains many more patterns than dose a positive class [45]. The ADPC-FLD algorithm is applied to three imbalanced datasets, including Ecoli (im & others), Yeast (ME2 & others) and Pima-indians (1 & 0). The descriptions of the three imbalance datasets are shown in Table 12. Among them, the Ecoli dataset has 8 different categories, where the im class is used as the positive class, the other 7 classes are used as the negative class, and the imbalanced ratio is 23:77. The Yeast dataset includes 10 different categories, where the ME2 is used as the positive class, the other 9 categories are used as the negative class, and the imbalanced ratio is 3:97. There are two types of Pima-indians dataset, and its imbalanced ratio is 35:65.

TABLE 12. Description of the three imbalanced datasets.

No.	Datasets	Pos.	Neg.	Total	Attributes	Imbalanced ratio
1	Ecoli (im & others)	77	259	336	8	23:77
2	Yeast (ME2 & others)	51	1433	1484	8	3:97
3	Pima-indians(1 & 0)	268	500	768	8	35:65

On the three imbalanced UCI datasets, the ADPC-FLD algorithm is compared with six state-of-the-art algorithms, including (1) the HKAP-LLE algorithm [38], (2) the hybrid support vector machine algorithm (HSVM) [46], (3) the adaptive synthetic sampling with different error costs algorithm (ADASYN+DEC) [46], (4) the random under-sampling algorithm (RAMU) [47], (5) the adaptive synthetic sampling algorithm (ADASYN) [48], and (6) the different error costs algorithm (DEC) [49]. Following the experimental techniques designed by Liu et al. [46], we adjust the parameters of the six compared algorithms. The density estimation entropy of ADPC-FLD adaptively yields the optimal d_c, based on which, the two-dimensional decision graphs of three UCI datasets are illustrated in Fig. 15. To efficiently evaluate the clustering performance of the three imbalanced datasets, the *G-mean* index [46] is used and described as

$$G{-}mean = \sqrt{R \times S} = \sqrt{\frac{ad}{(a+c)(b+d)}}. \tag{23}$$

So do we all experiments running 10 times to reduce the random errors, and then their mean of the clustering accuracy are obtained. The experimental results are shown in Table 13.

Fig. 15 shows the two-dimensional decision graphs of ADPC-FLD on the three imbalanced datasets in Table 12. The

TABLE 13. The *G-mean* index for the seven algorithms on the three imbalanced datasets.

Datasets	ADPC-FLD	HKAP-LLE	HSVM	ADASYN+DEC	RAMU	ADASYN	DEC
Ecoli	92.74	87.92	88.65	86.53	89.08	86.35	85.33
Yeast	87.46	84.1	84.25	83.94	81.9	83.52	82.37
Pima-indians	72.23	73.56	72.67	72.16	72.1	67	69.58

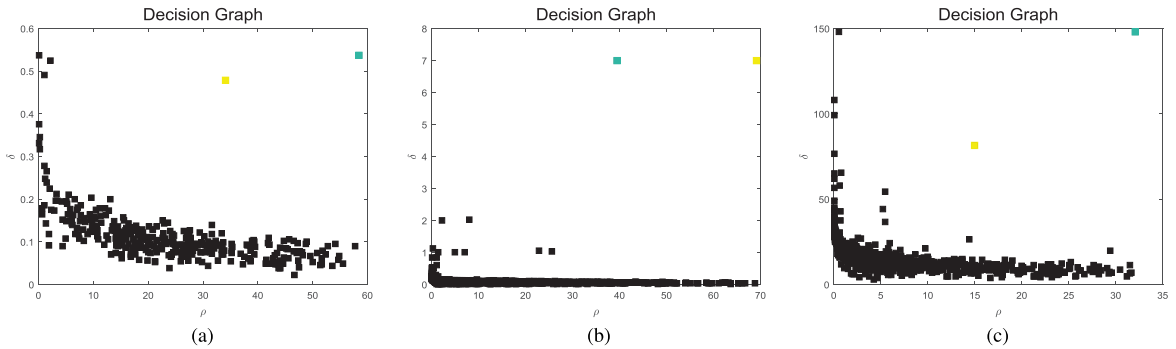


FIGURE 15. The two-dimensional decision graphs of the ADPC-FLD algorithm on the three imbalanced datasets. (a) Ecoli ($d_c = 0.2054$). (b) Yeast ($d_c = 0.2291$). (c) Pima-indians ($d_c = 5.4074$).

color markers are the selected cluster centers. As we can see from Fig. 15, the ADPC-FLD algorithm obtains the appropriate cluster centers and the accurate number of clusters on the three imbalanced datasets. Table 13 shows the *G-mean* index of the seven compared algorithms on the three imbalanced datasets. To more intuitively compare the seven algorithms, Fig. 16 shows the histogram of *G-mean* for the seven algorithms on the three imbalanced datasets. From the height of the histogram, the pros and cons of each compared algorithm can be clearly indicated. On the Ecoli dataset, the *G-mean* of ADPC-FLD is the largest, and RAMU exhibits the second best performance as HSVM. The *G-mean* values of HKAP-LLE, ADASYN+DEC, ADASYN and DEC are similar and smaller than those of ADPC-FLD, RAMU and HSVM. It can be easily concluded from the *G-mean* on the Yeast dataset, ADPC-FLD performs the best result, and RAMU achieves the worst accuracy. In detail, the other five algorithms shows slightly better performance than the RAMU algorithm, and slightly worse than the ADPC-FLD algorithm. On the Pima-indians dataset, the *G-mean* of ADPC-FLD is slightly lower than those of HKAP-LLE and HSVM, but larger than the other four methods. Although the clustering accuracy of ADPC-FLD is slightly lower than those of HKAP-LLE and HSVM on the Pima-indians dataset, it is superior to the other six algorithms on both the Ecoli and Yeast datasets. In general, the proposed ADPC-FLD algorithm is effective on the imbalanced UCI datasets.

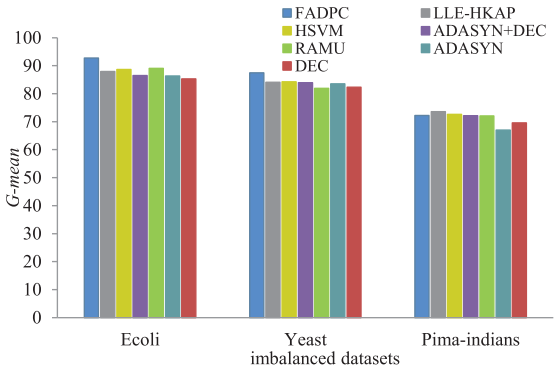


FIGURE 16. The *G-mean* index for the seven algorithms on the three imbalanced datasets.

D. COMPARISONS OF TESTING ACCURACY ON GENE EXPRESSION DATASETS

It is well known that the gene expression datasets have the common characteristics with high dimensionality and small samples [38], where include a large number of irrelevant and redundant features. Classification learning from gene expression data is a significant topic and inspiring many applications in cancer diagnosis [50], [51]. It follows that this portion concerning high-dimensional gene expression datasets is conducted to verify the clustering performance of the ADPC-FLD algorithm. Seven gene expression datasets are selected from http://csse.szu.edu.cn/staff/zhuzx/Datasets.html and http://www.gems-system.org/., respectively. The detailed description of the seven gene expression datasets is shown in Table 14. Note that, following the experimental techniques and parameters designed by Sun et al. [38], to ensure a fair comparison and reduce the random error, the 5-fold cross validation method is employed to test the clustering results in terms of three indices (R, S, and AC) on the seven gene expression datasets.

The first portion of this subsection is to evaluate the clustering performance of ADPC-FLD on the three gene expression datasets (Colon, Leukemia and Prostate) selected from Table 14. The two-dimensional decision graphs of the ADPC-FLD algorithm on the three gene expression datasets are shown in Fig. 17. The optimal d_c of each data is adaptively selected by the minimization method of density estimation

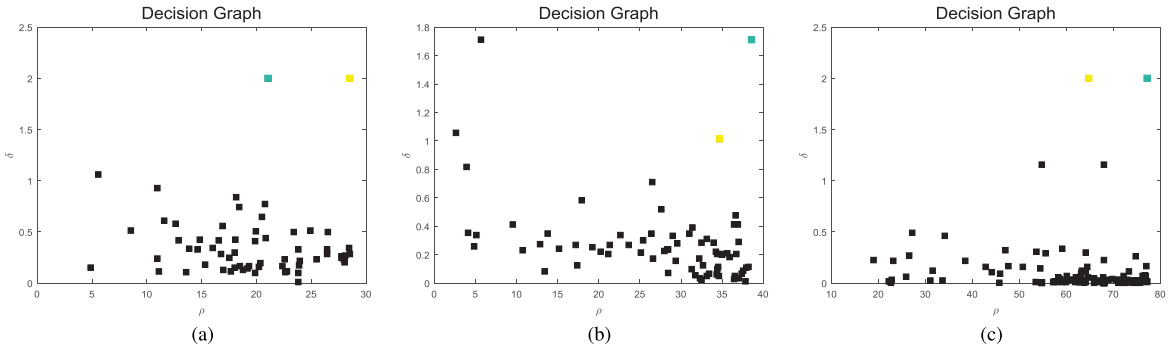


FIGURE 17. The two-dimensional decision graphs of the ADPC-FLD algorithm on three gene datasets. (a) Colon($d_c = 0.3492$). (b) Leukemia($d_c = 1.5351$). (c) Prostate($d_c = 1.3687$).

TABLE 14. Description of the seven gene expression datasets.

No.	Datasets	Samples	Attributes	Clusters
1	Colon	62	2000	2
2	Leukemia	72	7129	2
3	Prostate	136	12600	2
4	SRBCT	83	2308	4
5	Leukemia1	72	5327	3
6	9-Tumor	60	5726	9
7	Prostate1	102	10509	2

TABLE 15. The testing accuracy of the four algorithms on the three gene datasets.

Datasets	Indices	ADPC-FLD	HKAP-LLE	HMM	IG-SGA
Colon	R	1	0.9364	0.8147	0.883
	S	1	0.9522	0.9283	0.9489
Leukemia	R	0.9827	0.9786	0.9648	0.973
	S	0.9506	0.9947	0.9931	0.9978
Prostate	R	1	0.9829	0.9356	1
	S	1	0.9617	0.8884	1

entropy. The color markers are the selected cluster centers, and the d_c value is optimized by our proposed density estimation entropy. It can be observed from Fig. 17 that ADPC-FLD effectively selects the cluster centers and the number of clusters on the three gene expression datasets.

The second portion of this experiment is to evaluate the testing accuracy of ADPC-FLD on the three gene expression datasets (Colon, Leukemia and Prostate). Three compared state-of-the-art algorithms include: (1) the HKAP-LLE algorithm [38], (2) the hidden markov model (HMM) [50], and (3) the information gain and standard genetic algorithm (IG-SGA) [51]. The experimental results are shown in Table 15. To further illustrate the classification accuracy on the Colon and Leukemia datasets, five state-of-the-art algorithms, including (1) the HKAP-LLE algorithm [38], (2) the affinity propagation-based classifier ensemble selection algorithm (APCES) [52], (3) the ensemble gene selection algorithm by grouping (EGSG) [53], (4) the ensemble selection algorithm based on the random subspace method (RSM) [54], and (5) the random forest-based feature selection algorithm (RF) [55], are selected to compare with the ADPC-FLD algorithm. Following the experimental techniques and parameters designed by Sun et al. [38], all the above compared algorithms are adjusted to obtain the optimal performance. Similar to the last part, the optimal d_c of ADPC-FLD is adaptively selected by the proposed density estimation entropy. The experimental results on the Colon, Leukemia and Prostate datasets are evaluated with the four algorithms under the two indices (R and S), shown in Table 15. The results in terms of R, S and AC on the Colon and Leukemia datasets are illustrated in Table 16. The values of evaluation indices are between 0 and 1. The larger the value is, the better performance the algorithm shows.

Table 15 shows the comparisons of the R and S indices with the four algorithms on the three gene expression datasets. It can be observed from Table 15 that for the Colon dataset, the R and S indices of ADPC-FLD are much larger than the other three methods, and then the results indicate that the clustering performance of our algorithm on the Colon dataset is better than the other three algorithms. The HMM algorithm has the lowest clustering accuracy that leads to the worst performance on the Colon dataset. For the Leukemia dataset, the R value of ADPC-FLD is larger than the other three methods. The values of S for the HKAP-LLE, HMM and IG-SGA algorithms are similar, and the difference range is between 0.001 and 0.003. The S value of ADPC-FLD is slightly smaller than those of the other three methods, but its accuracy is pretty good and reaches 0.9506. On the Prostate dataset, the two indices of ADPC-FLD and IG-SGA are both 1, indicating that the two algorithms are much suitable for the Prostate dataset. More narrowly, it can be obviously observed from Table 15 that the values of R for HKAP-LLE and HMM are smaller than those of ADPC-FLD and IG-SGA, while the S index of HMM is much smaller than those of the other three algorithms on the Prostate dataset. In general, on the three gene expression datasets, the HMM algorithm is inferior to the other three algorithms, and the ADPC-FLD algorithm performs the best results, which indicate that ADPC-FLD is efficient on the three high-dimensional gene expression datasets.

Table 16 shows the comparison results of the R, S and AC indices for the six compared algorithms on the Colon and Leukemia datasets. As shown in Table 16, on the Colon dataset, the ADPC-FLD and HKAP-LLE algorithms perform well and achieve the great values for all three indices. Then, it can fully demonstrate the effectiveness of the two algo-

TABLE 16. The testing accuracy of the six algorithms on the colon and leukemia datasets.

Datasets	Indices	ADPC-FLD	HKAP-LLE	APCES	EGSG	RSM	RF
Colon	*R*	1	0.936	0.85	0.841	0.861	0.732
	S	1	0.952	0.822	0.802	0.788	0.903
	AC	1	0.925	0.84	0.83	0.814	0.83
Leukemia	*R*	0.983	0.979	0.958	0.914	0.945	0.996
	S	0.951	0.995	0.822	0.802	0.788	0.728
	AC	0.964	0.986	0.975	0.935	0.952	0.903

rithms on the Colon dataset. The APCES, EGSG and RSM algorithms have the similar values on R and S, but their values are smaller than those of the ADPC-FLD and HKAP-LLE algorithms. For the RF algorithm, the values of the R index are the smallest, but the values of the S index are larger than those of the APCES, EGSG and RSM algorithms. Under the AC index, the RSM algorithm achieves the smallest value, indicating that the effect is the worst. On the Leukemia dataset, the R and AC values of all algorithms are higher than 0.9. Among them, the RF algorithm has the largest R on the Leukemia dataset, the ADPC-FLD algorithm is the second, and the EGSG algorithm is the smallest. The AC value of ADPC-FLD is slightly smaller than those of HKAP-LLE and APCES, but larger than the other three methods. The S index of the ADPC-FLD and HKAP-LLE algorithms are much larger than the other four algorithms and reach above 0.95, while the S value of the RF algorithm is the smallest and less than 0.73. The above results and analysis illustrate that the ADPC-FLD algorithm is efficient on the Leukemia dataset. Therefore, the three evaluation indices efficiently show that ADPC-FLD can perform the better clustering results on the two gene expression datasets.

The final portion of this subsection denotes the clustering accuracy of our ADPC-FLD algorithm using the AC index, which is compared with eleven state-of-the-art clustering algorithms on the four gene expression datasets (SRBCT, Leukemia1, 9-Tumor and Prostate1), selected from Table 14. The eleven compared algorithms proposed in recent years include: (1) the HKAP-LLE algorithm [38], (2)-(4) three low rank projection least square regression (LPLSR) and subspace segmentation algorithms (LPLSR-2 and LPLSR-1) [56], (5)-(7) three subspace segmentation algorithms (LatLRR, LRR, and LSR) [57]–[60], (8)-(9) two non-negative matrix factorization-based algorithms (S-NMF and C-NMF) [61], (10) the K-means algorithm [62], and (11) the hierarchical clustering algorithm (HC) [63]. The values of the relevant parameters in each algorithm are designed by Salem *et al.* [51]. The two-dimensional decision graphs of the ADPC-FLD algorithm on the four datasets is illustrated in Fig. 18, and the clustering accuracies of the twelve algorithms on the four datasets are shown in Table 17.

Fig. 18 shows the two-dimensional decision graphs of the ADPC-FLD algorithm on the SRBCT, Leukemia1, 9-Tumor and Prostate1 datasets, where the color markers are the selected cluster centers, and the d_c value is optimized by the proposed density estimation entropy. According to Fig. 18, on the SRBCT, Leukemia1 and Prostate1 datasets,

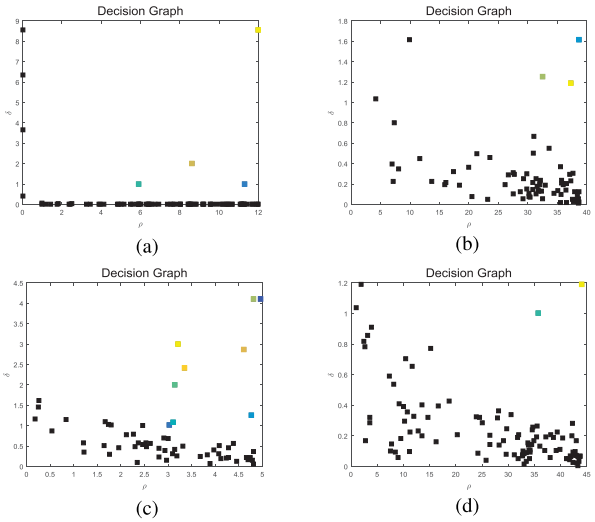


FIGURE 18. The two-dimensional decision graphs of the ADPC-FLD algorithm on the four gene datasets. (a) SRBCT ($d_c = 1.8273$). (b) Leukemia1 ($d_c = 2.0158$). (c) 9-Tumor ($d_c = 0.8914$). (d) Prostate1 ($d_c = 0.3642$).

TABLE 17. The *AC* of the twelve algorithms on the four gene datasets.

Algorithms	Datasets			
	SRBCT	Leukemia1	9-Tumor	Prostate1
HC	33.73	52.78	23.33	51.96
K-means	46.94	55.69	43.00	62.94
C-NMF	49.04	54.44	41.50	61.67
S-NMF	49.27	62.64	38.50	58.14
LSR	52.17	65.27	47.83	63.72
LRR	61.81	62.36	41.50	61.76
LatLRR	62.53	68.47	40.67	62.75
LPLSR-1	71.08	79.17	45.17	78.43
LPLSR-2	74.70	80.56	43.67	78.43
LPLSR	74.22	88.89	48.83	78.43
HKAP-LLE	83.78	89.53	70.40	83.59
ADPC-FLD	85.16	95.77	80.22	86.27

ADPC-FLD can select the appropriate cluster centers and the accurate number of clusters. But as shown in Fig. 18(c), it is difficult to determine the suitable cluster centers on the 9-Tumor dataset. Table 17 shows the comparisons of the AC values for the twelve algorithms on the four gene expression datasets. As we can see from Table 17, the four traditional clustering algorithms (S-NMF, C-NMF, K-means, and HC) have the lower clustering accuracies and the poor performance on the four datasets. The clustering accuracies of the subspace clustering algorithms (LPLSR, LPLSR-2, LPLSR-1, LatLRR, LRR, and LSR) are higher than those of the four traditional clustering algorithms. The clustering accuracies of LPLSR-1, LPLSR-2 and LPLSR are similar and higher than the other three subspace clustering algorithms. The ADPC-FLD and HKAP-LLE algorithms perform much better in terms of clustering accuracy than the other ten algorithms, especially the ADPC-FLD algorithm. Taking the LPLSR algorithm as an example, the clustering accuracy of the proposed algorithm on the SRBCT, Leukemia1, 9-Tumor and Prostate1 datasets is increased by about 10%, 7%, 31% and 8%, respectively, which fully indicates that our ADPC-FLD method is superior to the other eleven methods on these

datasets. Thus, the clustering results on the four gene datasets illustrate that the proposed ADPC-FLD algorithm is more valid and feasible than the other eleven algorithms. In summary, the ADPC-FLD algorithm is suitable for the high-dimensional gene expression datasets and performs better than the other recent clustering algorithms.

V. CONCLUSION

In order to solve the problems that the traditional DPC algorithm is difficult to deal with the complex datasets, the Euclidean distance only considers the spatial structure of the sample and does not take into account the correlation and the similarity between samples, and the manual setting of parameters affects the objectivity of clustering results. On the basis of the local structural characteristics of the data, this paper proposes an ADPC method with Fisher linear discriminant. The Pearson correlation coefficient is firstly introduced as the weight, and then the kernel density estimation function based on the weighted Euclidean distance is used to calculate the local density between the samples. Then, the density estimation entropy is presented to select the density estimation parameters. An adaptive strategy of cluster center selection is designed to construct a novel ADPC algorithm. Finally, the Fisher linear discriminant method is employed to reduce the dimensionality of the high-dimensional data, and then an ADPC-FLD algorithm is developed. The experimental results on synthetic datasets, standard UCI datasets and gene expression datasets indicate that the presented ADPC-FLD algorithm can obtain the more accurate cluster centers and the higher clustering accuracy, which proves that our method can effectively process the complex datasets.

REFERENCES

[1] Z. Chen, D. Chang, and Y. Zhao, "An automatic clustering algorithm based on region segmentation," *IEEE Access*, vol. 6, pp. 74247–74259, 2018.
[2] H. Wang and G. Liu, "Two-level-oriented selective clustering ensemble based on hybrid multi-modal metrics," *IEEE Access*, vol. 6, pp. 64159–64168, 2018.
[3] J. Xie, W. Xie, H. Gao, X. Liu, and P. W. Grantd, "Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors," *Inf. Sci.* vol. 354, pp. 19–40, Aug. 2016.
[4] M.-S. Yang, S.-J. Chang-Chien, and Y. Nataliani, "A fully-unsupervised possibilistic c-means clustering algorithm," *IEEE Access*, vol. 6, pp. 78308–78320, 2018.
[5] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[6] J. Gao, M. Kang, J. Tian, L. Wu, and M. Pecht, "Unsupervised locality-preserving robust latent low-rank recovery-based subspace clustering for fault diagnosis," *IEEE Access*, vol. 6, pp. 52345–52354, 2018.
[7] M. Du, S. Ding, and Y. Xue, "A robust density peaks clustering algorithm using fuzzy neighborhood," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 7, pp. 1131–1140, 2018.
[8] R. Zhang, F. Nie, and X. Li, "Self-weighted spectral clustering with parameter-free constraint," *Neurocomputing* vol. 241, pp. 164–170, Jun. 2017.
[9] H. Zhou, B. Xi, Y. Zhang, J. Li, and F. Zhang, "A graph clustering algorithm using attraction-force similarity for community detection," *IEEE Access*, vol. 7, pp. 13683–13692, 2019.
[10] L. Yaohui, M. Zhengming, and Y. Fang, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy," *Knowl.-Based Syst.* vol. 133, pp. 208–220, Oct. 2017.
[11] S. A. Seyedi, A. Lotfi, P. Moradi, and N. N. Qader, "Dynamic graph-based label propagation for density peaks clustering," *Expert Syst. Appl.*, vol. 115, pp. 314–328, Jan. 2019.
[12] H. Shao, P. Zhang, X. Chen, F. Li, and G. Du, "A hybrid and parameter-free clustering algorithm for large data sets," *IEEE Access*, vol. 7, pp. 24806–24818, 2019.
[13] L. Zheng, Y. Qu, X. Qian, and G. Cheng, "A hierarchical co-clustering approach for entity exploration over Linked Data," *Knowl.-Based Syst.*, vol. 141, pp. 200–210, Feb. 2018.
[14] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco, "An information-theoretic approach to hierarchical clustering of uncertain data," *Inf. Sci.*, vol. 402, pp. 199–215, Sep. 2017.
[15] T. Li, L. Zhang, W. Lu, H. Hou, X. Liu, W. Pedrycz, and C. Zhong, "Interval kernel fuzzy c-means clustering of incomplete data," *Neurocomputing* vol. 237, pp. 316–331, May 2017.
[16] A.-J. Gallego, J. Calvo-Zaragoza, J. J. Valero-Mas, and J. R. Rico-Juan, "Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation," *Pattern Recognit.*, vol. 74, pp. 531–543, Feb. 2018.
[17] W. Ding and P. X.-K. Song, "EM algorithm in Gaussian copula with missing data," *Comput. Statist. Data Anal.*, vol. 101, pp. 1–11, Sep. 2016.
[18] J. Fan and T. W. S. Chow, "Sparse subspace clustering for data with missing entries and high-rank matrix completion," *Neural Netw.* vol. 93, pp. 36–44, Sep. 2017.
[19] X. Zhao, J. Liang, and C. Dang, "Clustering ensemble selection for categorical data based on internal validity indices," *Pattern Recognit.*, vol. 69, pp. 150–168, Sep. 2017.
[20] R. Bie, R. Mehmood, S. Ruan, Y. Sun, and H. Dawood, "Adaptive fuzzy clustering by fast search and find of density peaks," *Pers. Ubiquitous Comput.* vol. 20, no. 5, pp. 785–793, Oct. 2016.
[21] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM Sigmod Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999.
[22] L. Bai, X. Cheng, J. Liang, H. Shen, and Y. Guo, "Fast density clustering strategies based on the k-means algorithm," *Pattern Recognit.*, vol. 71, pp. 375–386, Nov. 2017.
[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.
[24] Z. Liang and P. Chen, "Delta-density based clustering with a divide-and-conquer strategy: 3DC clustering," *Pattern Recognit. Lett.*, vol. 73, pp. 52–59, Apr. 2016.
[25] M. Du, S. Ding, and Y. Xue, "A novel density peaks clustering algorithm for mixed data," *Pattern Recognit. Lett.* vol. 97, pp. 46–53, Oct. 2017.
[26] S. Ding, M. Du, T. Sun, X. Xu, and Y. Xue, "An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood," *Knowl.-Based Syst.*, vol. 133, pp. 294–313, Oct. 2017.
[27] M. Du, S. Ding, and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," *Knowl.-Based Syst.*, vol. 99, pp. 135–145, May 2016.
[28] X. Xu, S. Ding, and Z. Shi, "An improved density peaks clustering algorithm with fast finding cluster centers," *Knowl.-Based Syst.*, vol. 158, pp. 65–74, Oct. 2018.
[29] J. Jiang, Y. Chen, D. Hao, and K. Li, "DPC-LG: Density peaks clustering based on logistic distribution and gravitation," *Phys. A, Stat. Mech. Appl.*, vol. 514, pp. 25–35, Jan. 2019.
[30] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Ann. Statist.*, vol. 38, no. 5, pp. 2916–2957, 2010.
[31] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing*, vol. 216, pp. 208–215, Dec. 2016.
[32] L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with Fisher score for tumor classification," *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, Apr. 2019.
[33] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
[34] W. Li, B. Liao, W. Zhu, M. Chen, Z. Li, X. Wei, L. Peng, G. Huang, L. Cai, and H. Chen, "Fisher discrimination regularized robust coding based on a local center for tumor classification," *Sci. Rep.*, vol. 8, no. 1, Jun. 2018, Art. no. 9152.
[35] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Jan. 2010.

[36] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Statist. Assoc.*, vol. 78, no. 383, pp. 553–569, 1983.
[37] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Proc. IEEE 8th Int. Conf. Data Mining*, Dec. 2008, pp. 63–72.
[38] L. Sun, R. Liu, J. Xu, S. Zhang, and Y. Tian, "An affinity propagation clustering method using hybrid Kernel function with LLE," *IEEE Access*, vol. 6, pp. 68892–68909, 2018.
[39] R. Liu, H. Wang, and X. Yu, "Shared-nearest-neighbor-based clustering by fast search and find of density peaks," *Inf. Sci.* vol. 450, pp. 200–226, Jun. 2018.
[40] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, Philadelphia, PA, USA: SIAM, Jan. 2007, pp. 1027–1035.
[41] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, 2012.
[42] L. Wang, Q. Ji, and X. Han, "Adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity," *Tehnicki Vjesnik/Tech. Gazette*, vol. 23, no. 2, pp. 425–435, 2016.
[43] W. Limin, H. Xuming, and J. Qiang, "Semi-supervised affinity propagation clustering algorithm based on fireworks explosion optimization," in *Proc. IEEE Int. Conf. Manage. E-Commerce, E-Government*, Oct./Nov. 2015, pp. 273–279.
[44] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.
[45] C. Zhu and Z. Wang, "Entropy-based matrix learning machine for imbalanced data sets," *Pattern Recognit. Lett.*, vol. 88, pp. 72–80, Mar. 2017.
[46] D. Q. Liu, Z. J. Chen, Y. Xu, and F. T. Li, "Hybrid SVM algorithm oriented to classifying imbalanced datasets," *Chin. Appl. Res. Comput.*, vol. 35, no. 4, pp. 1023–1027, 2018.
[47] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, Nov. 2016, Art. no. 31.
[48] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.* Jun. 2008, pp. 1322–1328.
[49] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 1999, pp. 55–60.
[50] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Hidden Markov models for cancer classification using gene expression profiles," *Inf. Sci.*, vol. 316, pp. 293–307, Sep. 2015.
[51] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput.*, vol. 50, pp. 124–134, Jan. 2017.
[52] J. Meng, H. Hao, and Y. Luan, "Classifier ensemble selection based on affinity propagation clustering," *J. Biomed. Informat.*, vol. 60, Apr. 2016, pp. 234–242.
[53] H. W. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 81–87, Feb. 2010.
[54] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[56] X. Chen, B. Xiao, and L. Lin, "Low rank projection least square regression subspace segmentation for gene expression data," *Pattern Recognit. Artif. Intell.*, vol. 30, no. 2, pp. 106–116, Feb. 2017.
[57] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.* Jun. 2010, pp. 663–670.
[58] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[59] C. Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy: Springer-Verlag, Oct. 2012, pp. 347–360.
[60] G. Liu and S. Yan, "Latent Low-Rank Representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 1615–1622.
[61] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.
[62] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
[63] F. Nielsen, "Hierarchical clustering," *Revista Mexicana De Astronomía Y Astrofísica*, vol. 43, no. 2. Universidad Nacional Autónoma De México, Instituto De Astronomía, Mexico City, Mexico, 1999, pp. 59–67.

LIN SUN received the M.S. degree in computer science and technology from Henan Normal University, in 2007, and the Ph.D. degree in pattern recognition and intelligent systems from the Beijing University of Technology, in 2015. He became a Postdoctoral Researcher with the Medical and Biological Engineering Research Group, Henan Normal University, China, in 2016, where he is currently an Associate Professor with the College of Computer and Information Engineering. He has received funding from ten grants from the National Natural Science Foundation of China, the China Postdoctoral Science Foundation, the Plan for Scientific Innovation Talent of Henan Province, and the Key Scientific and Technological Project of Henan Province. He has authored or coauthored for more than 70 articles. His current research interests include granular computing, cluster analysis, big data mining, and intelligent information processing. He has received the title of Henan's Distinguished Young Scholars for Science and Technology Innovation Talents. He has served as a Reviewer in several prestigious peer-reviewed international journals.

RUONAN LIU received the B.Sc. degree in computer science and technology from Henan Normal University, in 2016, where she is currently pursuing the master's degree in computer science and technology with the College of Computer and Information Engineering. Her current research interests include granular computing, cluster analysis, and data mining.

JIUCHENG XU received the M.S. and Ph.D. degrees in computer science and technology from Xi'an Jiaotong University, in 1995 and 2004, respectively. He is currently a Professor with the College of Computer and Information Engineering, Henan Normal University. He has received funding from grants from the National Natural Science Foundation of China, the Key Scientific Research Project of Higher Education of Henan Province, and the Key Scientific and Technological Project of Henan Province. He has published more than 100 articles. His research interests include granular computing, data mining, intelligent information processing, and pattern recognition. He has received the title of Henan's Distinguished High Profile Professional. He has served as a Reviewer in several prestigious peer-reviewed international journals.

SHIGUANG ZHANG received the M.S. degree in mathematics from Guangxi University for Nationalities, in 2007, and the Ph.D. degree in applied mathematics from Hebei Normal University, in 2014. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tianjin University, Tianjin, China. He is also with the College of Computer and Information Engineering, Henan Normal University, China. He has authored more than 10 peer-reviewed papers. His research interests include knowledge discovery and machine learning. He has served as a Reviewer in several prestigious peer-reviewed international journals.