IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 19, 2018, accepted October 27, 2018, date of publication November 9, 2018, date of current version December 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2880271

An Affinity Propagation Clustering Method Using Hybrid Kernel Function With LLE

LIN SUN[1], RUONAN LIU[1], JIUCHENG XU[1], SHIGUANG ZHANG[1,2], AND YUN TIAN[3]

[1]College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China
[2]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[3]College of Information Science and Technology, Beijing Normal University, Beijing 100875, 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, Grant 61370169, and Grant 61472042, 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 and Grant 182102210078, in part by the Young Scholar Program of Henan Province under Grant 2017GGJS041, in part by the Key Scientific and Technological Project of Xinxiang City under Grant CXGG17002, in part by the Natural Science Foundation of Henan Province under Grant 182300410130 and Grant 182300410368, and in part by the Ph.D. Research Foundation of Henan Normal University under Grant qd15132 and Grant qd15129.

ABSTRACT Cluster analysis is important in data mining and clustering algorithms and has gained much attention during the last decade. However, it is a challenge to extract significant features from high-dimensional data and to rapidly provide satisfactory clustering results. This paper presents a new affinity propagation (AP) clustering method based on a hybrid kernel function with locally linear embedding, called LLE-HKAP, for the classification of gene expression datasets and standard UCI datasets. First, the locally linear embedding algorithm is used to reduce the dimension of the original dataset. Then, a novel AP clustering method based on a similarity measure with the hybrid kernel function is proposed. In this method, a new global kernel is defined that has high generalization ability. Meanwhile, a hybrid kernel function that linearly combines the proposed global kernel and the Gaussian kernel is defined to further enhance the learning ability of the global kernel. Moreover, the novel hybrid kernel is introduced to define a similarity measure and construct a similarity matrix of the AP clustering. Finally, the improved AP clustering algorithm is implemented on eight public gene expression datasets and eight standard UCI datasets for comparison with other related algorithms. The experimental results validate that our proposed clustering algorithm is efficient in terms of clustering accuracy and outperforms the currently available approaches with which it is compared.

INDEX TERMS Granular computing, cluster, reduction, affinity propagation, kernel function.

I. INTRODUCTION

Cluster analysis is one of the most important unsupervised learning techniques for extracting information from data. This type of analysis is widely used in a number of fields such as granular computing, data mining, machine learning, and bioinformatics [1], [2]. Granular computing as an important mathematical method has been introduced in cluster analysis to reveal the uncertainty of structured datasets, and it groups similar objects in clusters called information granules [3]. The aim of clustering analysis is to cluster datasets into categories that contain different information granules by minimizing the similarity between clusters and maximizing the similarity within clusters [4]. The purpose of clustering is to effectively eliminate redundancy in high-dimensional data, so that the most significant data can be identified [3], [5].

With the development of clustering techniques, many clustering algorithms have been reported. Kumar and Reddy [6] introduced an efficient initial seed selection method for improving the performance of the K-means filtering method by locating the seed points in dense areas of the dataset and ensuring that they are well separated. Xu *et al.* [7] performed a self-adaptive extreme learning machine algorithm based on rough set theory and affinity propagation (AP) clustering for finding the appropriate and universal number of hidden nodes. Truong *et al.* [8] developed an advanced fuzzy possibilistic C-means clustering method based on granular computing to select features as a preprocessing method for clustering problems. Pagnuco *et al.* [9] studied a method based on cluster validation indices and hierarchical clustering for identifying sets of coexpressed genes. Xu *et al.* [10]

calculated a hierarchical clustering method based on density peaks for directly generating clusters in each possible clustering layer. Denoeux *et al.* [11] developed an evidential clustering algorithm based on the iterative row-wise quadratic programming method for problems of dissimilar data. Ding *et al.* [12] presented an entropy-based density peaks clustering algorithm for dealing with numerical, categorical, and mixed type data. However, when tackling complex high-dimensional data, these methods cannot yield accurate clustering results. At present, many cluster ensemble algorithms can effectively handle high-dimensional data [13]. Hu *et al.* [14] constructed cluster ensemble models based on knowledge granulation and rough set theory for solving the cluster ensemble problem as well as an ensemble learning application of knowledge granulation. Meng *et al.* [15] offered a classifier ensemble selection method based on AP clustering. Zhao *et al.* [16] investigated a clustering ensemble selection algorithm for improving the quality of clustering results and retaining the information of the original data. However, some of these methods still have shortcomings; for example, the number of clusters and the initial cluster centers must be determined in advance. Although ensemble clustering algorithms can perform well when using high-dimensional datasets, a large number of experiments need to be implemented to determine the number and the quality of the base classifiers, which has a great effect on the clustering results [16].

To solve the problem of dimensionality and achieve satisfactory clustering results, it is necessary to combine dimension reduction with cluster analysis. Commonly used dimension reduction methods include principle component analysis [17], linear discriminant analysis [18], isometric feature mapping [19], Laplacian eigenmap [20], and local linear embedding (LLE) [21]. LLE, a dimension reduction algorithm for high-dimensional data, was proposed by Roweis and Sual [21]. The LLE algorithm is used to reduce the dimensionality of nonlinear data [22]. Globally nonlinear data are converted into locally linear data, and global structure information is obtained by overlapping local areas. After linear dimension reduction of each local area, low-dimensional global coordinates are obtained by combining the results, according to certain rules [23], [24]. Therefore, the LLE algorithm has some advantages; for example, it is easy to operate, and the processed low-dimensional data can maintain the original topology [25].

The AP presented by Frey and Dueck [26] is an exemplar-based clustering method. In contrast to traditional clustering methods, AP simultaneously considers all data points as potential cluster centers, and does not require the specification of initial cluster centers; the number of clusters is unknown [3], [27]–[29]. In recent years, many improved methods and extensions of AP algorithms have been applied in many domains [3]. Givoni *et al.* [30] performed a hierarchical AP algorithm for solving hierarchical clustering problems. Shang *et al.* [31] described a fast AP clustering approach that simultaneously considered both local and global structure information in datasets. Wang and Chen [32] developed an AP algorithm to control the number of clusters and identify multiple exemplars that can represent each cluster automatically. Zhang and Gu [33] introduced an adaptive AP clustering algorithm for clustering mixed datasets. Hang *et al.* [34] designed a transfer AP algorithm for identifying the appropriate number of clusters. However, both AP and extended AP suffer from poor performance in clustering accuracy when processing complex high-dimensional data. Therefore, in this study, an improved AP algorithm is combined with the LLE algorithm for dimension reduction to increase efficiency while guaranteeing clustering performance. Another shortcoming of the AP algorithm is that data point similarity is defined as a negative squared error, i.e., Euclidean distance. Thus, similarity is larger if the distance between points is smaller. However, in practice, similarity can be defined according to specific issues without satisfying Euclidean space constraints [35]. The Euclidean distance can also lead to misclassification and reduce clustering performance. To solve this issue, in this study, a new kernel function is introduced to develop a similarity measure for the AP algorithm.

The kernel-based learning method developed from statistical theory is an essentially nonlinear information processing tool [36]. Compared with other learning methods, the kernel-based learning methods have many advantages for addressing complex high-dimensional pattern recognition tasks [37]. In recent years, because single kernel functions possess only strong generalization ability or strong learning ability, hybrid kernel functions consider the global and local properties of base kernels and can provide both strong generalization ability and strong learning ability [38]. Yeh *et al.* [39] provided a two-stage multiple-kernel learning algorithm based on a linear combination of the radial basis kernel function with different hyperparameters to decrease the amount of time and space required. Wang *et al.* [40] defined a kernel function that combines the global and local information of base kernels and presented an alternative algorithm with proven convergence to identify multiple kernel coefficients. Thus, this paper focuses on creating a new hybrid kernel function. To overcome the challenge of extracting relevant and significant features from high-dimensional data and to rapidly provide satisfactory clustering results, an improved AP algorithm that uses a hybrid kernel function with LLE is proposed. First, the LLE algorithm is used to map high-dimensional data into a low-dimensional space for linear dimension reduction. It is effective to reduce the dimensionality of high-dimensional datasets and retain the potential information in the data. Then, to overcome the problem of misclassification caused by the Euclidean distance and maintain the original structure of the data, a new global kernel function is defined and a novel hybrid kernel function that provides both strong generalization ability and strong learning ability is constructed by linearly combining the proposed global kernel and the Gaussian kernel. The proposed hybrid kernel function is introduced into a similarity measure of the AP algorithm to form a new similarity matrix. Finally, the proposed hybrid kernel function AP (HKAP) algorithm is used to cluster low-dimensional data. The experimental results pertaining to several gene expression datasets and standard UCI datasets demonstrate that the proposed method has better classification accuracy and effectiveness than the other related methods.

The rest of this paper is structured as follows. Section 2 briefly reviews the basic theories of the AP clustering algorithm and LLE-based dimension reduction. In Section 3, a novel global kernel function and a hybrid kernel function are developed, and a similarity measure and its similarity matrix are constructed. Then, the LLE-HKAP algorithm is presented. The experimental results and analysis are described in Section 4. Finally, Section 5 presents the conclusion.

II. RELATED WORK

In this section, the basic notions of the AP clustering algorithm and LLE-based dimension reduction are briefly reviewed [15], [27], [35], [41], [42].

A. AP CLUSTERING ALGORITHM

The AP clustering algorithm takes the similarities between pairs of data points as its input. The similarity matrix, denoted as $S_{N\times N}$, where the similarity measure $s(i,j) = -\parallel x_i - x_j \parallel^2$ is described by the Euclidean distance between two data points, is fundamental to AP.

Responsibility and availability are two significant factors of AP; the former is denoted as $r(i,k) = s(i,k) - \max\limits_{k'\neq k}\{a(i,k') + s(i,k')\}$ and the latter is expressed as

$$a(i,k) = \begin{cases} \min\{0, r(k,k) + \sum\limits_{i'\notin\{i,k\}} \max\{0, r(i',k)\}\}, \\ \quad \text{if } i \neq k \\ \sum\limits_{i'\neq k} \max\{0, r(i',k)\}, \quad \text{if } i = k \end{cases}$$

The AP algorithm searches for clusters through an iterative process until a high-quality set of exemplars and corresponding clusters are assembled. Meng *et al.* [15] introduced a damping factor λ into iterations to overcome the problems of oscillation and convergence failure. The iterative formulas are respectively described as

$$r^t(i,k) = (1-\lambda) * r^t(i,k) + \lambda * r^{t-1}(i,k), \tag{1}$$

$$a^t(i,k) = (1-\lambda) * a^t(i,k) + \lambda * a^{t-1}(i,k), \tag{2}$$

where t indicates the tth iteration.

B. LLE-BASED DIMENSION REDUCTION

The LLE maps a dataset $X \in R^N$ globally to a dataset $Y \in R^M$, where $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_m\}$. The basic principle of LLE is to minimize the reconstruction error of the set of all local neighborhoods in the dataset [41]. An original data point x_{ij} with D-dimension is input, where $1 \leq i \leq n, 1 \leq j \leq n$ and the distance measure of k neighboring points for every sample point is calculated using the Euclidean distance. This measure is denoted by

$$\begin{cases} d_{ij} = \sqrt[2]{\sum_{k=1}^{D} |x_{ik} - x_{jk}|^2}, & \text{if } i \neq j \\ d_{ij} = 0, & \text{if } i = j, \end{cases} \tag{3}$$

where $1 \leq i \leq n$, $1 \leq j \leq n$, $1 \leq k \leq D$, and k is set according to experience to a value that is greater than the output dimension of the samples. Then, the k nearest neighbors $N_k(x_i)$ of the ith point are selected to calculate the reconstructed weight vectors [42].

For each input point, the optimal linearly reconstructed weight vectors can be calculated by

$$\begin{aligned} \varepsilon_i(w) &= \min ||x_i - \sum_{j=1, j\neq i}^{k} w_{ij}x_j||^2 \\ &= \min \| \sum_{j,j\neq i} w_{ij}(x_i - x_j)\|^2 \\ &= \sum_{j,k} w_j w_k G_{jk}, \end{aligned} \tag{4}$$

where ε_i is an error function of the linear reconstruction between x_i and k neighboring points $x_1, x_2, \ldots, x_k$, and $G_{jk} = (x_i - x_j)^T(x_i - x_k)$ is a local Gramian matrix, and w_{ij} is a linearly reconstructed weight that is subject to the following constraint condition: $\sum w_{ij} = 1$, where $w_{ij} = 1$ if x_j is a neighboring point of x_i and $w_{ij} = 0$ otherwise. Additionally, an optimal weight w_j is calculated using the Lagrange multiplier approach, i.e., $w_j = \frac{\sum_k G_{jk}^{-1}}{\sum_{bm} G_{bm}^{-1}}$, where b, $m = 1, 2, \cdots, k$.

These reconstructed weights are used to find the low-dimensional embedding matrix Y, which is defined as

$$\begin{aligned} \varepsilon(Y) &= \min \sum_{i=1}^{k} ||y_i - \sum_{j=1}^{k} w_{ij}y_j||^2 \\ &= \min ||Y(I - W^T)||^2 \\ &= \min\ trYMY^T, \end{aligned} \tag{5}$$

where $i, j = 1, 2, \ldots, k$, $y_i \in Y$, y_i satisfies $\sum_{i=1}^{N} y_i = 0$ and $\frac{\sum_{i=1}^{N} y_i y_i{}^T}{N} = I$, where I is a $d \times d$ unit matrix, and M is a sparse symmetric positive semi-definite matrix and that equals $(I - W)^T(I - W)$.

III. HYBRID KERNEL FUNCTION-BASED AP CLUSTERING METHOD WITH LLE

For high-dimensional and large-scale data, the traditional distance-based clustering method cannot effectively avoid the curse of dimensionality. Thus, it is necessary to combine dimension reduction methods with cluster analysis to achieve better clustering results. In this study, the improved AP algorithm is combined with the LLE algorithm to solve this problem. First, the LLE algorithm is used to map the original high-dimensional dataset into a low-dimensional space for dimension reduction. Then, a new global kernel function is defined, and a novel hybrid kernel function is obtained by combining the global kernel with the Gaussian kernel to develop a similarity measure. A similarity matrix of the AP algorithm is constructed using the similarity measure. Thus, the LLE-HKAP algorithm is presented.

A. GLOBAL KERNEL FUNCTION

The kernel function $K(x, y)$ is defined as a dot product of the feature space: $\Phi(x)*\Phi(y) = K(x, y)$. The symmetric kernel $K(x, y)$ must satisfy the Mercer condition. The kernel functions include the global kernel function and the local kernel function. If the global kernel function allows distant data to significantly impact the value of the kernel function, then such functions have stronger generalization ability but weaker learning ability. If the local kernel function allows nearby data to significantly impact the value of the kernel function, then such functions have stronger learning ability but weaker generalization ability [38]. The two types of kernel functions are combined so that they complement each other and achieve better results than traditional kernel functions. With the advantages of the global kernel function, a new global kernel function is introduced. Here, a global function is given as follows:

Definition 1: A new global function is defined as

$$f(x) = \frac{1}{1+\exp(-\frac{x^2}{a^2})}, \tag{6}$$

where x, $a \in R$.

If a function satisfies the conditions of Mercer's theorem, then it can be called as a kernel function. Mercer's theorem [43] is described as follows.

Proposition 1: Suppose that $k(x, x')$ is a continuous symmetric function. Then $k(x, x')$ is a support vector machine (SVM) kernel function if and only if $\int_{R^d} g^2(\xi)d_\xi < \infty$ for $\forall g \neq 0$, when $\iint k(x, x')g(x)g(x')dxdx' \geq 0$ holds.

If a kernel function is translation-invariant (for example, $k(x, x') = k(x - x')$), then it is challenging to prove that the function satisfies Mercer's theorem. The following lemma provides necessary and sufficient conditions for a translation-invariant kernel function.

Lemma 1: Let $k(x)$ be a translation-invariant kernel function whose Fourier transform is $F[k(\omega)] = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-j\omega x)k(x)dx$. Then, $k(x)$ is an SVM kernel function if and only if the Fourier transform satisfies $F[k(\omega)] \geq 0$.

Definition 2: Given a dataset $X \in R^N$, where $X = \{x_1, x_2, \cdots, x_n\}$, a global kernel function is defined as

$$K(x_i, x_j) = \frac{1}{1+\exp(-\frac{(x_i-x_j)^2}{a^2})}, \tag{7}$$

where x_i, $x_j \in X$, $1 \leq i \leq n$, $1 \leq j \leq n$, and $a \in R$.

From Definitions 1 and 2, the following proposition can be obtained.

Proposition 2: The global kernel function is an SVM kernel function.

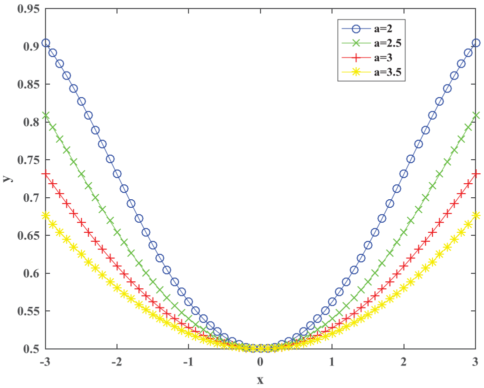


FIGURE 1. The functional behavior of the global kernel.

Proof: The Fourier transform of the global kernel function is as follows:

$$F[k(\omega)] = (2\pi)^{-\frac{d}{2}} \int_{R^d} \exp(-j\omega x) \frac{1}{1+\exp(-\frac{x^2}{a^2})} dx \geq 0.$$

This expression can be rewritten as

$$\begin{aligned} F[k(\omega)] &= (2\pi)^{-\frac{d}{2}} \int_{R^d} \frac{\exp(-j\omega x)}{1+\exp(-\frac{x^2}{a^2})} dx \\ &= (2\pi)^{-\frac{d}{2}} \int_{-\infty}^{+\infty} \frac{\exp(-j\omega x)}{1+\exp(-\frac{x^2}{a^2})} dx. \end{aligned}$$

It follows that $\exp(-j\omega x) > 0$ and $1 + \exp(-\frac{x^2}{a^2}) > 0$. Then, $F[k(\omega)]$>0. When x is close to ∞, the integral is approximately 0. Thus, the integral is an improper integral, and its value is 0; that is, $F[k(\omega)] = 0$. In conclusion, the Fourier transform of the global kernel satisfies $F[k(\omega)] \geq 0$. Hence, the proposed global kernel can be called an SVM kernel function.

The functional behavior of our proposed global kernel is shown in Fig. 1, where the parameter is set to different values (i.e., $a = 2$, 2.5, 3, and 3.5). It can be seen from Fig. 1 that our proposed global kernel function has better generalization ability, and its generalization ability that changes with the size of parameter a.

B. HYBRID KERNEL FUNCTION

As mentioned in Section 3.1, the proposed global kernel function has strong generalization ability and the local kernel function has strong learning ability. A new hybrid kernel function that combines the global kernel function with the local kernel function has both strong generalization ability and strong learning ability. The linear combination of the global kernel and the local kernel is also called a kernel function. The hybrid kernel functions consider the global and local properties of the base kernels and can provide both strong generalization ability and strong learning ability. Yeh *et al.* [39] and Wang *et al.* [40] proposed several hybrid kernel functions and made good applications, respectively. Then, in this study, by borrowing that idea of hybrid kernel

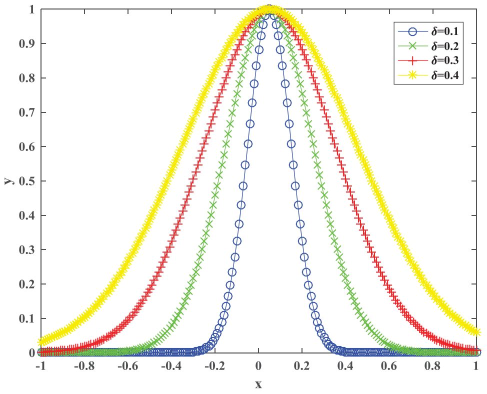


FIGURE 2. The functional behavior of the Gaussian kernel.

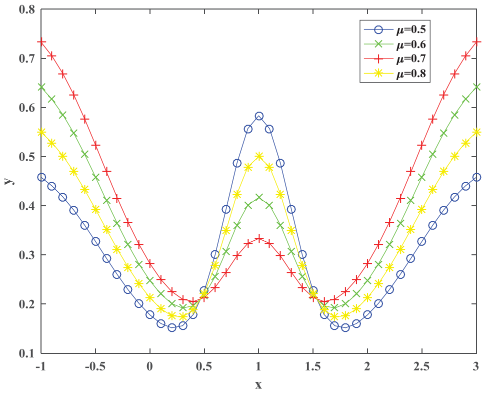


FIGURE 3. The functional behavior of the hybrid kernel.

functions in [39] and [40], the global kernel function and the local kernel function of the new hybrid kernel function can be given different linear weights. The Gaussian kernel function can map the input space to a feature space with infinite dimension, and this function has a simple structure, fast convergence and strong learning ability [36]; the Gaussian kernel $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. Because of the above advantages, the Gaussian kernel function is selected as the local kernel function. The functional behavior of the Gaussian kernel is shown in Fig. 2.

Definition 3: By linearly combining the global kernel function with the Gaussian kernel function, a new hybrid kernel function K_H is defined as

$$K_H = \mu K_{GK} + (1-\mu)K_G, \tag{8}$$

where K_{GK} is the proposed global kernel, K_G is the Gaussian kernel, and μ is used to adjust the effects of the two kernels, with $0 \leq \mu \leq 1$.

When $\mu = 1$, the hybrid kernel becomes the global kernel. When $\mu = 0$, the hybrid kernel becomes the local kernel. Fig. 3 shows the functional behavior of the hybrid kernel for various values of μ.

Then, following the computation approach to similarity measure in [30], [33], and [34], the proposed hybrid kernel function K_H is used as a new similarity measure $s(i, j)$ of AP algorithm, i.e., $s(i, j) = K_H$.

Algorithm 1 LLE

```
Input: Original dataset $X = \{x_1, x_2, \cdots, x_n\}$, the number of nearest neighbor points $K$, and the dimensions $d$
Output: Low-dimensional embedding matrix $Y$
for each data point $x_i$ in $X$ do
    Find $K$ nearest neighbors $N_k(x_i)$
    Compute the weights that best reconstructed from $N_k(x_i)$ using Eq. 4
end for    //compute the reconstructed weight vectors
Find the weight matrix $w$ using the reconstructed weights
Compute the low-dimensional embedding vector $Y$ using Eq. 5
return $Y$
```

Definition 4: A similarity measure $s(i, j)$ in the AP algorithm is defined as

$$s(i,j) = \frac{\mu}{1+\exp(-\frac{(x_i-x_j)^2}{a^2})} + (1-\mu)\exp\left(-\frac{\| x_i - x_j\|^2}{2\sigma^2}\right), \tag{9}$$

where $x_i, x_j \in X$, $1 \leq i \leq n$, $1 \leq j \leq n$, $0 \leq \mu \leq 1$, and a, $\sigma \in R$.

From the above analysis, since the kernel function and its parameters directly determine the nonlinear mapping of feature space, the performance of the kernel depends on its parameters. If the kernel function parameters are improperly chosen, poor results will be obtained. Therefore, it is important to determine the adjustable parameters μ, a, and σ correctly.

Definition 5: A similarity matrix $S_{n\times n}$ in the AP algorithm is defined as

$$S = [s(i,j)]_{n\times n}, \tag{10}$$

where $s(i, j)$ denotes a similarity measure, $1 \leq i \leq n$, and $1 \leq j \leq n$.

C. THE LLE-HKAP ALGORITHM

The goal of this section is to combine the proposed AP algorithm with the LLE algorithm for the analysis of high-dimensional and large-scale data. Fig. 4 illustrates the special procedures of the LLE-HKAP algorithm. As we can see from Fig. 4, the LLE algorithm is first used to map the original high-dimensional dataset into low-dimensional space for dimension reduction; the specific steps of the LLE algorithm are shown in Algorithm 1. Then, a new hybrid kernel function is used as a similarity measure, and a similarity matrix is constructed. Based on the similarity matrix, HKAP updates the responsibility and availability for each point, selects the cluster centers, and allocates the other data points based on the nearest cluster centers; the specific steps of the LLE-HKAP algorithm are depicted in Algorithm 2.

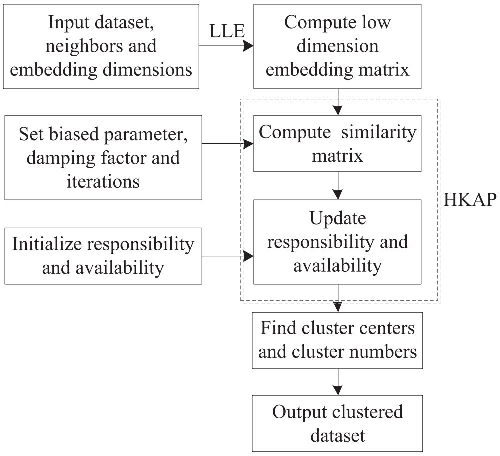

Input dataset, neighbors and embedding dimensions
LLE
Compute low dimension embedding matrix
Set biased parameter, damping factor and iterations
Compute similarity matrix
Update responsibility and availability
HKAP
Initialize responsibility and availability
Find cluster centers and cluster numbers
Output clustered dataset

FIGURE 4. Flowchart of the LLE-HKAP algorithm.

Algorithm 2 LLE-HKAP

```
Input: Original dataset $X = \{x_1, x_2, \cdots, x_n\}$, the biased parameter $p$, the damping factor $\lambda$, and the number of iterations $t$
Output: Clustered $C$.
Initialize $r(i, k) = 0$ and $a(i, k) = 0$
Compute the low-dimensional embedding matrix $Y$ using Algorithm 1    //reduce dimensionality based on LLE
for each data point $y_i$ in $Y$ do
    Compute the distance between two points
end for
Compute the similarity matrix $S$ using Eq. 10    //construct the similarity measure
for $t = 1{:}1000$ do
    Based on matrix $S$, compute $r(i, k)$ and $a(i, k)$ using Eqs. 1 and 2, respectively    //update responsibility and availability
    Compute the value of $r(i, k) + a(i, k)$
    Find the cluster centers, and compute the number of cluster centers as paper [15]
    if converge then
        break
    end if
end for    //cluster based on HKAP
if correct cluster number then
    break
else
    Change the value of the biased parameter $P$
    Repeat
    Until obtain correct cluster number
end if    //adjust the biased parameter
return Clustered $C$
```

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENT PREPARATION

In this section, the performances of our proposed LLE-HKAP algorithm described in Section 3.3 is demonstrated by evaluating our algorithm in terms of the number of clusters, the number of iterations, the operation time, and the clustering accuracy. Our experiments can be divided into three parts as follows. The LLE-HKAP algorithm is run on eight public gene expression datasets to evaluate the performances of the LLE algorithm, the kernel-function-based AP algorithms and the related classification algorithms. For the final part, LLE-HKAP is run on eight standard UCI datasets to evaluate classification accuracy and precision. These experiments are performed on a personal computer running Windows 7 with an Intel(R) Core(TM) i5-3470 CPU operating at 3.2 GHz and 4 GB of memory.

Five indices [15], [44], [45], i.e., silhouette index (*Sil*), precision (*P*), specificity (*S*), F-measure (*FM*), and accuracy (*AC*), are introduced to evaluate the clustering effect of the LLE-HKAP algorithm. Their formulas are expressed as follows

$$Sil(t) = \frac{[b(t) - a(t)]}{\max\{a(t), b(t)\}}, \tag{11}$$

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

$$P = \frac{TP}{TP + FP}, \tag{12}$$

$$R = \frac{TP}{TP + FN}, \tag{13}$$

$$S = \frac{TN}{TN + FP}, \tag{14}$$

$$FM = \frac{2 \times P \times R}{P + R}, \tag{15}$$

$$AC = \frac{TP + TN}{TP + FP + FN + TN}, \tag{16}$$

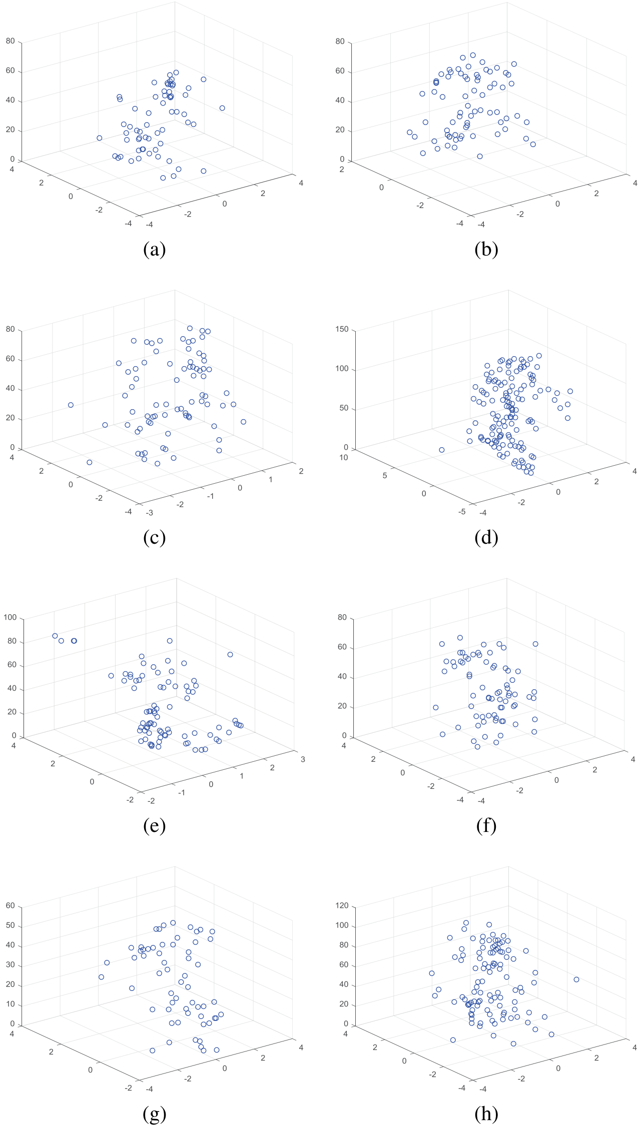
where True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*) are metrics.

B. THE EFFECT OF LLE AND THE HYBRID KERNEL FUNCTION

It is known that accurate cancer classification directly using original gene expression profiles remains challenging due to the intrinsic high-dimensional features and the small number of samples [28]. The objective of the following experiments is to show the clustering efficiency of the proposed general framework on eight types of public gene expression datasets. To evaluate the effectiveness of the dimension reduction of the LLE algorithm on high-dimensional gene expression datasets, the Colon, Leukemia, DLBCL and Prostate datasets are selected from http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, and the SRBCT, Leukemia1, 9-Tumor and Prostate1 datasets are selected from http://www.gems-system.org/. The basic information of these eight datasets is described in Table 1.

TABLE 1. Description of the eight gene expression datasets.

No.	Datasets	Samples	Genes	Clusters
1	Colon	62	2000	2
2	Leukemia	72	7129	2
3	DLBCL	77	5469	2
4	Prostate	136	12600	2
5	SRBCT	83	2308	4
6	Leukemia1	72	5327	3
7	9-Tumor	60	5726	9
8	Prostate1	102	10509	2



(a) (b) (c) (d) (e) (f) (g) (h)

FIGURE 5. The dimension reduction results of the eight gene expression datasets. (a) Colon dataset. (b) Leukemia dataset. (c) DLBCL dataset. (d) Prostate dataset. (e) SRBCT dataset. (f) Leukemia1 dataset. (g) 9-Tumor dataset. (h) Prostate1 dataset.

The first part of this experiment testing the proposed algorithm is to investigate the validity of the dimension reduction results of the LLE algorithm on eight gene expression datasets. The results are shown in Fig. 5. It is easily seen in Fig. 5 that the dimensions of the eight gene expression datasets are greatly reduced.

TABLE 2. Clustering results of the six algorithms for the colon dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	159	1.966	7	0.2643	0.54
LLE+GAP	140	1.763	3	0.5542	0.77
LLE+GKAP	133	1.591	4	0.4925	0.75
LLE+GPAP	100	1.131	3	0.7154	0.86
LLE+GSAP	67	0.747	3	0.6832	0.88
LLE-HKAP	80	1.138	2	0.7338	0.89

The second part of this experiment is to illustrate the efficiency of our proposed hybrid kernel function. The LLE-HKAP algorithm is compared with five combined AP algorithms: (1) the traditional AP algorithm [26]; (2) the Gaussian kernel AP (GAP) algorithm, whose similarity measure is calculated using the Gaussian kernel; (3) the global kernel AP (GKAP) algorithm, whose similarity measure is calculated using the proposed global kernel; (4) the GP kernel AP (GPAP) algorithm, whose similarity measure is calculated by combining the Gaussian kernel and polynomial kernel [38]; and (5) the GS kernel AP (GSAP) algorithm, whose similarity measure is calculated by combining the Gaussian kernel and sigmoid kernel [43]. Here, the selected kernel functions are the Gaussian kernel of GAP, the global kernel of GKAP, the Gaussian kernel and polynomial kernel [38], and the Gaussian kernel and sigmoid kernel [43]. For the eight reduced datasets illustrated in Fig. 5, the corresponding experimental parameters and results are summarized in Tables 2 to 9 and Figs. 6 to 14. Since gene expression data usually consist of thousands of genes and a small number of samples (that is, the data generally have high dimensionality and a small sample size), any attempt to mine knowledge from gene expression data may result in very poor performance without dimension reduction [46]. It follows that for the traditional AP algorithm [26], the AP with the Euclidean distance and without the LLE cannot obtain effective clustering results and is very time consuming to simulate. Thus, the comparison of the AP algorithm with the other six algorithms was ignored in Tables 2 to 9. Following the experimental techniques used in [27], the operation time (in seconds) and the number of iterations are employed to test the performance measures of the six algorithms in Tables 2 to 9, and the operation time is correlated with the number of iterations. It should be noted that, to determine the appropriate parameters of the algorithm and obtain more accurate properties, numerous experiments were performed. Then, the parameters of the algorithm were set as follows: the maximum number of iterations $t = 1000$, damping factor $\lambda = 0.8$, biased parameter p is set initially by the median diagonal value of the similarity matrix [15], and, for the hybrid kernel, the adjustable parameters μ, a, and σ of the eight datasets are shown in Fig. 6. Validation was conducted using 5-fold cross validation, i.e., the dataset was randomly divided into five groups with equal sample sizes, and in each validation process, four groups were used as training sets and one group was used as a test set.

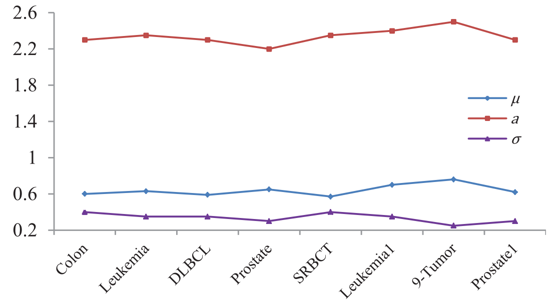


FIGURE 6. The adjustable parameters μ, a, and σ of the eight gene expression datasets.

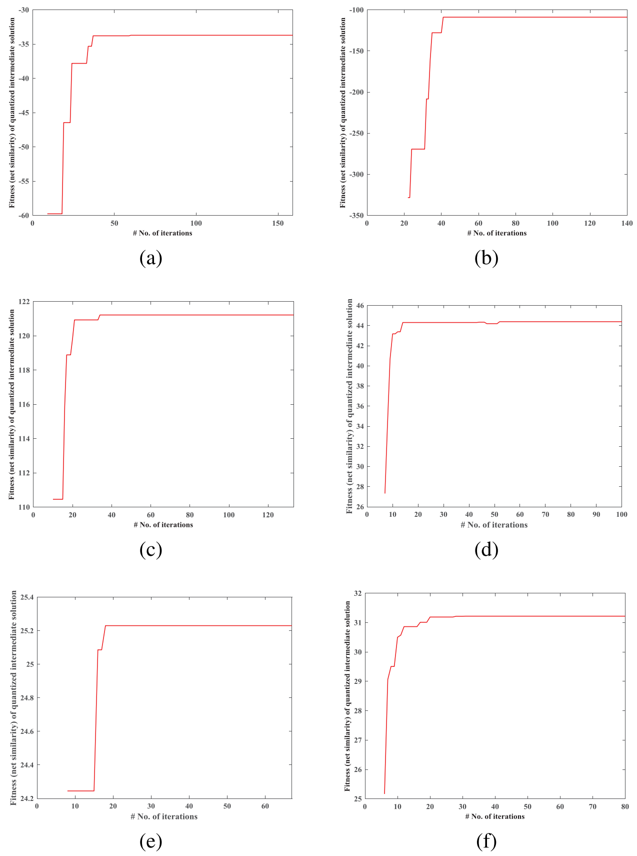


FIGURE 7. The iterations of the six algorithms for the Colon dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

TABLE 3. Clustering results of the six algorithms for the Leukemia dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	157	2.361	6	0.2061	0.57
LLE+GAP	134	1.685	4	0.4826	0.74
LLE+GKAP	132	1.684	3	0.5193	0.81
LLE+GPAP	123	1.384	2	0.6793	0.90
LLE+GSAP	119	1.437	3	0.5638	0.85
LLE-HKAP	82	1.547	2	0.6832	0.91

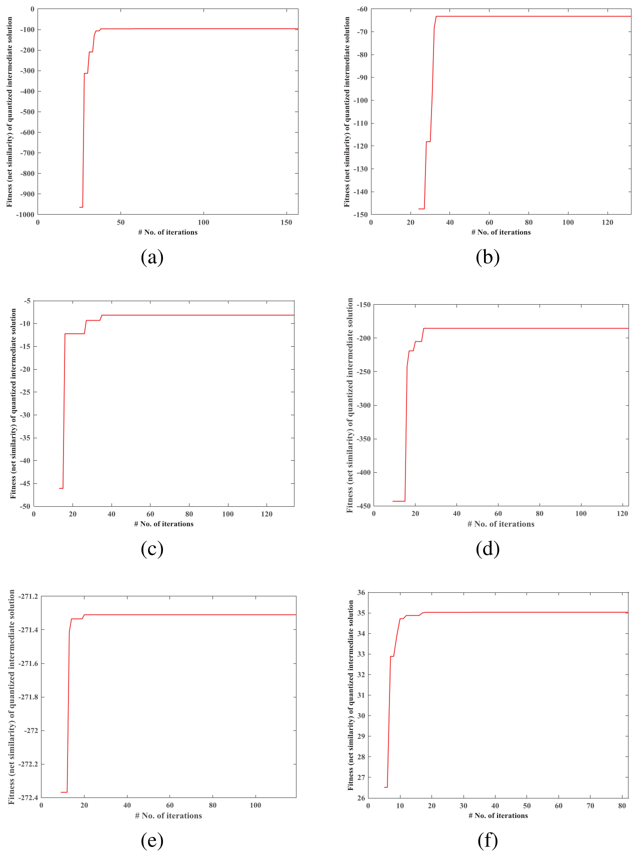


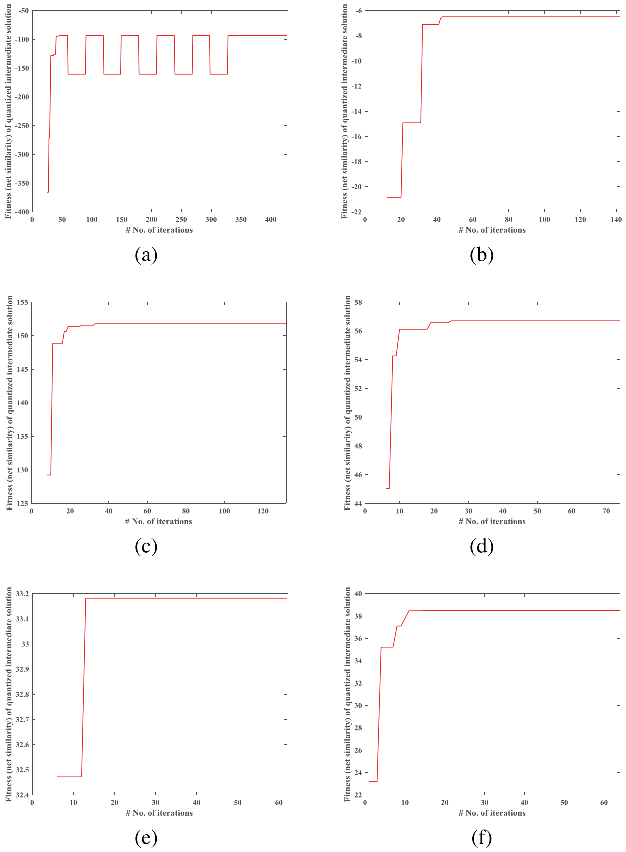
FIGURE 8. The iterations of the six algorithms for the Leukemia dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

From Table 2, we find that, although the operation time of the six algorithms increases in turn, the cluster number obtained by the AP algorithm is much larger than that of the kernel-based AP algorithms. The performance of hybrid-kernel AP algorithms is highlighted in iterations, and the GSAP algorithm performs the best, although it cannot obtain the correct number of clusters. The HKAP algorithm performs well in terms of cluster numbers, and the result is the same as the correct number of clusters. Moreover, the values of two indices indicate the effectiveness of the HKAP algorithm. The experimental results verify that HKAP can provide the most efficient results. Meanwhile, Fig. 7 illustrates the iterations of the six algorithms for the Colon dataset. The ordinate represents the net similarity of the clustering solution. Fig. 7 shows that the number of iterations tends to decrease when using the different similarity measure for the six algorithms; in other words, the kernel-based AP algorithms perform better than the Euclidean distance-based AP algorithm, and the hybrid kernel-based AP algorithms have better performance than the single kernel-based AP algorithms. Compared with the other two hybrid-kernel algorithms (GPAP and GSAP), the results obtained by the HKAP algorithm prove that similarity measure based on the hybrid of the Gaussian kernel and the sigmoid kernel is more effective.

Table 3 denotes that the performances of the GAP and GKAP algorithms are similar, and the GPAP and GSAP algorithms perform better than the GAP and GKAP algorithms in

TABLE 4. Clustering results of the six algorithms for the DLBCL dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	427	4.434	10	0.3156	0.48
LLE+GAP	142	1.51	5	0.5247	0.63
LLE+GKAP	132	1.403	5	0.5763	0.65
LLE+GPAP	74	1.002	4	0.7341	0.76
LLE+GSAP	62	0.865	3	0.8347	0.89
LLE-HKAP	64	1.283	2	0.8596	0.92



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 9. The iterations of the six algorithms for the DLBCL dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

terms of iterations and clusters. All of the algorithms perform markedly better than the AP algorithm, which performs the worst. The reason for failure of AP is that the similarity measure between data points is defined by the Euclidean distance, which considers all data points equally. However, kernel-based algorithms (GKAP, GAP, HKAP, GPAP and GSAP) can retain the original information of the data. It is clear that the HKAP algorithm developed in this paper performs the best for the Leukemia dataset. Fig. 8 intuitively shows the iteration process of the six algorithms for the Leukemia dataset. The ordinate represents the net similarity of the clustering solution. Fig. 8 illustrates that the number of iterations is the largest for the AP algorithm, while that of our proposed algorithm is the smallest. Numerical oscillation occurs in the early iterations of all algorithms, but the LLE-HKAP algorithm tends to converge soonest, that is, our algorithm has fast convergence and few iterations for the test data. The AP algorithm tends to converge slowest and its absolute value of net similarity is too large, demonstrating the poor performance of the traditional AP algorithm. In conclusion, our proposed algorithm is highly robust and efficient.

TABLE 5. Clustering results of the six algorithms for the Prostate dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	179	2.714	4	0.1622	0.63
LLE+GAP	128	1.941	2	0.3558	0.85
LLE+GKAP	149	1.966	2	0.3489	0.86
LLE+GPAP	62	0.834	3	0.6143	0.89
LLE+GSAP	99	1.234	2	0.6559	0.94
LLE-HKAP	69	1.198	2	0.6472	0.94

Fig. 9 shows the iterations of the six tested algorithms for the DLBCL dataset. The ordinate represents the net similarity of the clustering solutions. Fig. 9 denotes that the three hybrid-kernel AP algorithms clearly perform better than the AP and single-kernel AP algorithms in terms of the number of iterations. The number of iterations of the single-kernel AP algorithms is twice that of the hybrid-kernel AP algorithms, which reduces the efficiency of the algorithms. The hybrid-kernel AP algorithms tend to converge sooner than the single-kernel AP algorithms. For the two single-kernel AP algorithms, our proposed global-kernel AP algorithm performs better than the Gaussian-kernel AP algorithm. The AP algorithm exhibits the worst performance in terms of the number of iterations, and its speed of convergence is more than six times that of the LLE-HKAP algorithm. As shown in Table 4, although the HKAP algorithm performs slightly worse than the GPAP and GSAP algorithms in terms of the number of iterations and the operation time, it can obtain the correct cluster number. Because of the large number of iterations, the operation time of single-kernel and traditional AP algorithms are longer than those of the hybrid-kernel algorithms. These results further verify the effectiveness of the kernel-based similarity measure and our proposed global kernel.

Fig. 10 shows the iterations of the six tested algorithms for the Prostate dataset. The ordinate represents the net similarity of the clustering solution. Table 5 and Fig. 10 demonstrate that both the single-kernel AP algorithms and the hybrid-kernel AP algorithms yield the correct number of clusters; however, the single-kernel algorithms perform slightly worse than the hybrid-kernel algorithms in terms of the number of iterations and the operation time. Fig. 10 illustrates that the AP algorithm tends to converge slowest, while the LLE+GPAP algorithm tends to converge quickest. Although the LLE+GPAP algorithm performs slightly better than the LLE-HKAP algorithm in terms of the number of iterations and the operation time, it cannot obtain the correct cluster number, which results in worse indices. The effectiveness of the GKAP and GAP algorithms verifies that the improved similarity measure with a single kernel function can improve the accuracy of the AP algorithm. Meanwhile, the

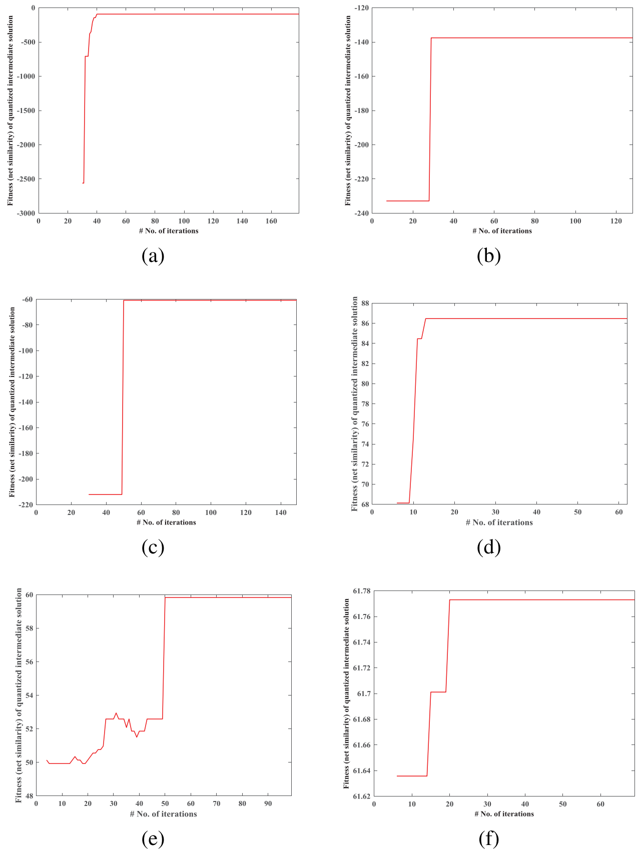


FIGURE 10. The iterations of the six algorithms for the Prostate dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

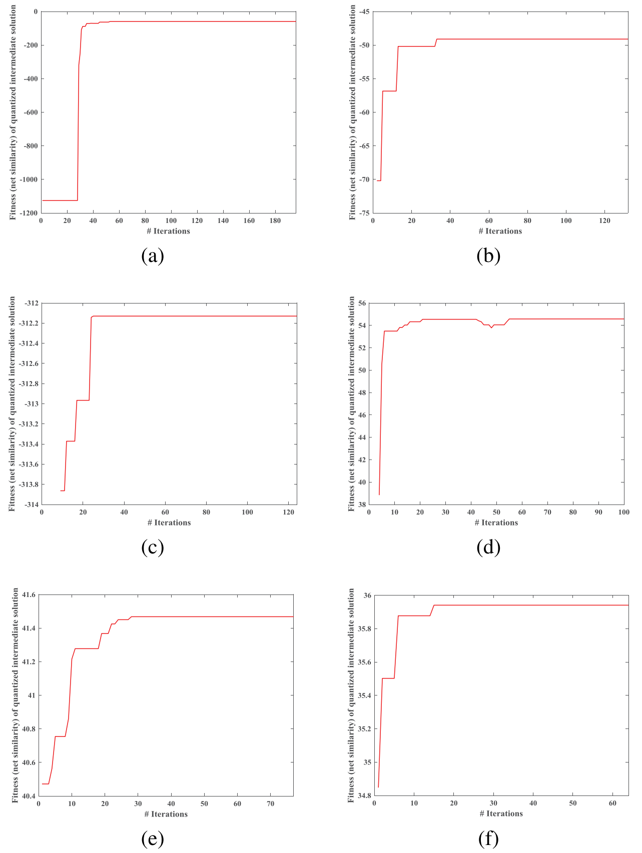


FIGURE 11. The iterations of the six algorithms for the SRBCT dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

TABLE 6. Clustering results of the six algorithms for the SRBCT dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	196	1.893	11	0.2135	0.48
LLE+GAP	132	1.34	4	0.6551	0.73
LLE+GKAP	124	1.251	5	0.6144	0.71
LLE+GPAP	100	1.057	6	0.5827	0.69
LLE+GSAP	77	0.77	4	0.705	0.83
LLE-HKAP	64	0.755	4	0.7362	0.86

TABLE 7. Clustering results of the six algorithms for the Leukemia1 dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	136	1.363	8	1.2285	0.45
LLE+GAP	121	1.209	5	0.4382	0.61
LLE+GKAP	139	1.352	4	0.5517	0.66
LLE+GPAP	100	1.064	3	0.7349	0.84
LLE+GSAP	70	0.727	4	0.6935	0.81
LLE-HKAP	61	0.672	3	0.8154	0.9

single-kernel algorithms, considering only the global or the local information of the data, show poorer performance than the hybrid-kernel algorithms. The experimental results obtained for the Prostate dataset show that the number of iterations and the operation time are greatly reduced when using the HKAP algorithm. Therefore, our proposed algorithm outperforms the other five tested algorithms.

Table 6 shows that the LLE-HKAP algorithm performs best in terms of the *Sil* and *FM* indices; the values of the two indices are the largest. It can be seen that the three algorithms (LLE+GAP, LLE+GSAP, and LLE-HKAP) can obtain the correct number of clusters, while the other three algorithms (LLE+AP, LLE+GKAP, and LLE+GPAP), especially the traditional AP algorithm, have slightly bad performance. Fig. 11 shows the iterations of the six tested algorithms for the SRBCT dataset. From Table 6 and Fig. 11, LLE-HKAP shows the best performance in terms of iterations, and the iterations of LLE+GSAP are a little worse than those of LLE-HKAP. In terms of operation time, the LLE-HKAP and LLE+GSAP algorithms are similar. It can be seen that the number of iterations and the operation times of LLE-HKAP and LLE+GSAP are much smaller than those of the other four algorithms. Table 6 shows that the hybrid-kernel algorithms (LLE+GPAP, LLE+GSAP, and LLE-HKAP) can obtain better performance on the SRBCT dataset.

Table 7 shows the clustering results of the six algorithms for the Leukemia1 dataset. Fig. 12 shows the iterations of the six tested algorithms for the Leukemia1 dataset. As shown in Table 7 and Fig. 12, the three algorithms (LLE+AP, LLE+GAP, and LLE+GKAP) have similar performance in terms of iterations and operation time, but the traditional AP algorithm produces the wrong number of clusters,

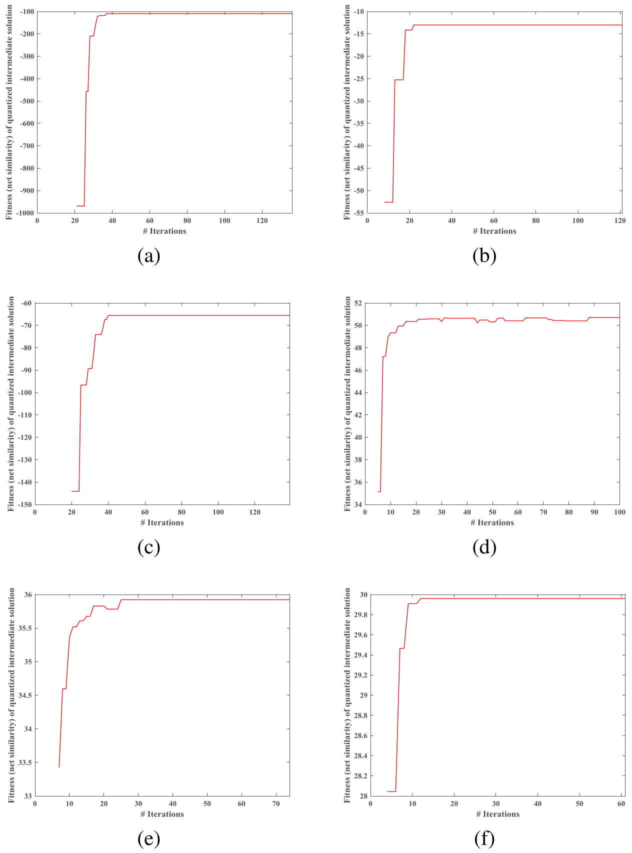


FIGURE 12. The iterations of the six algorithms for the Leukemia1 dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

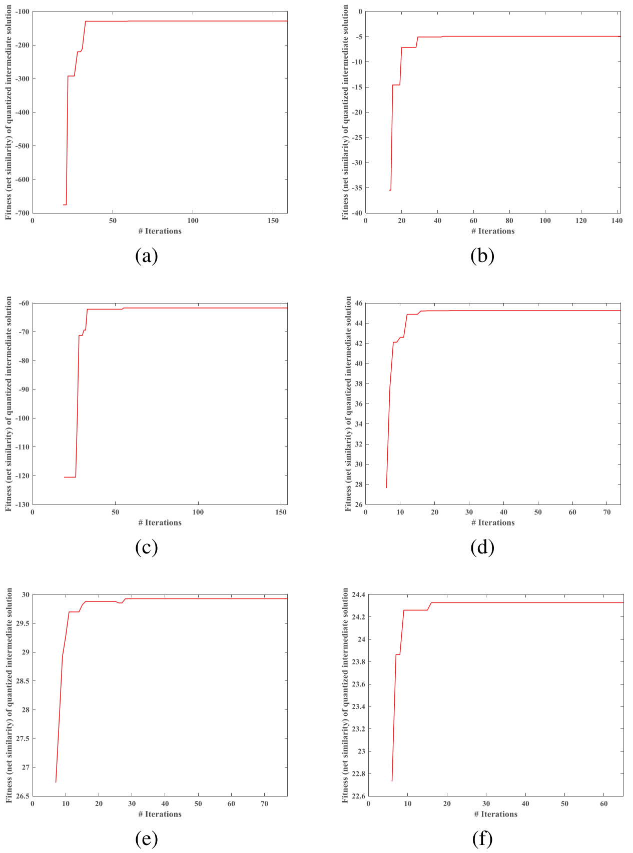


FIGURE 13. The iterations of the six algorithms for the 9-Tumor dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

TABLE 8. Clustering results of the six algorithms for the 9-Tumor dataset.

Algorithm	Iterations	Operation Time (s)	Clusters	Sil	FM
LLE+AP	159	1.514	9	0.1967	0.32
LLE+GAP	142	1.348	5	0.2683	0.39
LLE+GKAP	154	1.442	9	0.3158	0.46
LLE+GPAP	74	0.857	9	0.5167	0.68
LLE+GSAP	77	0.893	4	0.4623	0.62
LLE-HKAP	65	0.745	9	0.6847	0.72

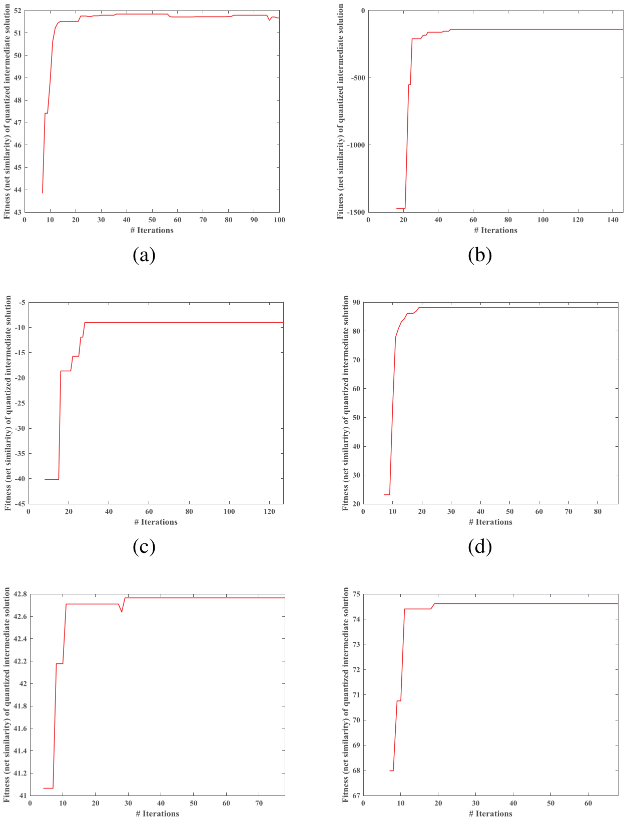
which leads to a lower value of the two evaluation indices. Compared with the single-kernel algorithms (LLE+GAP and LLE+GKAP), the three hybrid-kernel algorithms (LLE-HKAP, LLE+GPAP, and LLE+GSAP) show better performance in terms of iterations and operation times. Among them, the LLE+GPAP algorithm has the slowest iterations and wastes more operation time, and the LLE-HKAP and LLE+GSAP algorithms are similar. The LLE-HKAP and LLE+GPAP algorithms show the correct numbers of clusters, while the other four algorithms perform badly in this regard. In terms of the two evaluation indices, our LLE-HKAP algorithm performs better than the other five algorithms. In summary, the proposed LLE-HKAP algorithm has better clustering results on the Leukemia1 dataset.

Table 8 shows the clustering results of the six compared algorithms for the 9-Tumor dataset. Fig. 13 shows the iterations of the six tested algorithms for the 9-Tumor dataset. It can be observed from Table 8 and Fig. 13 that the LLE-HKAP algorithm has notably better performance than the other five algorithms in terms of iterations and operation time, and the hybrid-kernel algorithms (LLE-HKAP, LLE+GPAP, and LLE+GSAP) are better than the single-kernel algorithms (LLE+GAP and LLE+GKAP) and the traditional AP algorithms. Regarding the number of clusters, the LLE+GAP and LLE+GSAP algorithms perform slightly a little worse than the other four algorithms (LLE+AP, LLE+GKAP, LLE-HKAP, and LLE+GPAP), which obtain correct numbers of clusters. Based on the characteristics of the 9-Tumor dataset, all algorithms have low values for the clustering evaluation indices, and the LLE-HKAP algorithm is far better than the other five algorithms in terms of evaluation indices. These results further demonstrate the effectiveness of the proposed LLE-HKAP algorithm on the 9-Tumor dataset.

Table 9 shows the clustering results of the six contrast algorithms for the Prostate1 dataset. Fig. 14 shows the iterations of the six tested algorithms for the Prostate1 dataset. From Table 9, the LLE+AP algorithm has fewer numbers of iterations and a lower operation time than the LLE+GAP and LLE+GKAP algorithms, but according to

TABLE 9. Clustering results of the six algorithms for the Prostate1 dataset.

Algorithm	Iterations	Operation Time (*s*)	Clusters	*Sil*	*FM*
LLE+AP	100	1.115	13	0.1826	0.25
LLE+GAP	146	1.478	5	0.3749	0.51
LLE+GKAP	127	1.284	5	0.4328	0.55
LLE+GPAP	87	1.05	3	0.6612	0.73
LLE+GSAP	78	0.892	4	0.6278	0.69
LLE-HKAP	68	0.713	2	0.7965	0.86



(a) (b) (c) (d) (e) (f)

FIGURE 14. The iterations of the six algorithms for the Prostate1 dataset. (a) AP algorithm. (b) GAP algorithm. (c) GKAP algorithm. (d) GPAP algorithm. (e) GSAP algorithm. (f) HKAP algorithm.

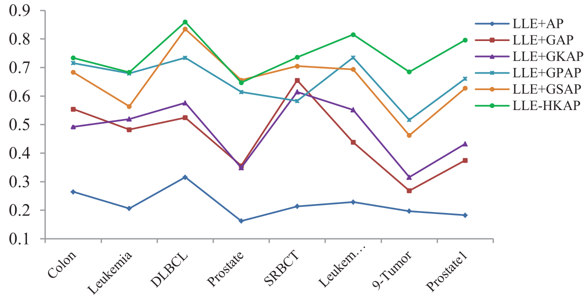


FIGURE 15. The *Sil* index of the six algorithms for the eight gene expression datasets.

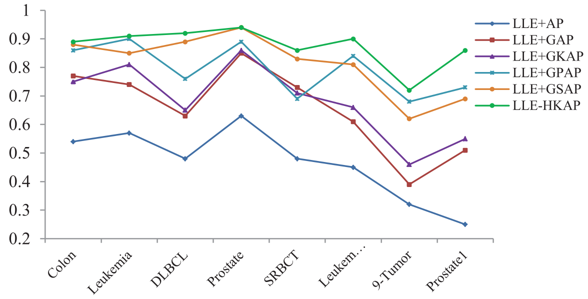


FIGURE 16. The *FM* index of the six algorithms for the eight gene expression datasets.

Fig. 14(a), the LLE+AP algorithm cannot converge. The LLE+AP algorithm shows poor performance in terms of the number of clusters, which leads to the small values of the two evaluation indices. From Table 9 and Fig. 14, the three hybrid-kernel algorithms (LLE-HKAP, LLE+GPAP, and LLE+GSAP) achieved better performance in terms of iterations and operation times, but only the LLE-HKAP algorithm can obtain the correct number of clusters. Meanwhile, the value of the larger evaluation indices indicates that the LLE-HKAP algorithm is superior to the other five algorithms for the Prostate1 dataset.

All comparisons indicate that the LLE-HKAP method yields the correct number of clusters and can provide favorable performance in terms of the number of iterations and the operation time. To further verify the clustering quality of Algorithm 2, two indices (*Sil* and *FM*) are introduced to evaluate the clustering results, which are shown in Figs. 15 and 16. Fig. 15 shows that, for the Colon dataset, the largest value of *Sil* indicates that the LLE-HKAP algorithm shows the best in terms of clustering accuracy, whereas AP shows the worst. The results for the Leukemia, DLBCL and Prostate datasets also demonstrate the good performance of the LLE-HKAP algorithm. Fig. 16 shows, for all datasets, that the values of *FM* of the kernel-based algorithms are clearly larger than those of the AP algorithm; the value of *FM* by the LLE-HKAP algorithm is largest. For the 9-Tumor dataset, based on the characteristics of the dataset, all algorithms have low values for the clustering evaluation indices, which still proves that the proposed LLE-HKAP algorithm is superior to other algorithms. Since *Sil* and *FM* can measure the clustering accuracy of the tested data [44], the largest values of the *Sil* and *FM* indices indicate that our proposed algorithm is effective. In summary, in terms of the number of iterations, the operation time and the clustering accuracy, LLE-HKAP performs better than the other tested methods.

C. COMPARISONS OF TESTING ACCURACY ON GENE EXPRESSION DATASETS

This portion of our experiments, which further concern gene expression data, is conducted to validate the classification accuracy of LLE-HKAP in comparison with other related state-of-the-art classification methods: (1) the hidden Markov model (HMM), which amplifies the probability of the given Information [47]; (2) the information gain and standard

TABLE 10. Comparison of the testing accuracy of the three algorithms on the four gene expression datasets.

Datasets	LLE-HKAP		HMM		IG-SGA	
	R	S	R	S	R	S
Colon	0.9364	0.9522	0.8147	0.9283	0.883	0.9489
Leukemia	0.9786	0.9947	0.9648	0.9913	0.973	0.9978
DLBCL	0.9737	0.9854	0.9697	0.9932	0.9356	0.9624
Prostate	0.9829	0.9617	0.9356	0.8884	1	1

TABLE 11. Comparison of the testing accuracy of the five algorithms on the two gene expression datasets.

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

genetic algorithm (IG-SGA) [45]; (3) the AP-based classifier ensemble selection algorithm (APCES) [15]; (4) the ensemble gene selection algorithm by grouping (EGSG) [48]; (5) the ensemble selection method based on the random subspace method (RSM) [49]; and (6) the random forest (RF)-based feature selection algorithm [50]. Following the experimental techniques designed by Salem *et al.* [45], the four typical gene expression datasets (Colon, Leukemia, DLBCL, and Prostate) are selected from Table 1, and 5-fold cross-validation method is used to evaluate the testing accuracy with recall R and specificity S on the selected four datasets. Table 10 shows the comparison of the testing accuracy of the HMM, IG-SGA with LLE-HKAP algorithms on the four gene expression datasets. Similarly, following the experimental techniques developed by Meng *et al.* [15], the Colon and Leukemia datasets are selected from Table 1 and the 5-fold cross-validation method is applied to test the clustering accuracy with R, S and the accuracy AC on the selected two datasets. The experimental results are shown in Table 11. In the comparative experiments, the large values of the three indices indicate that the algorithm has better performance in terms of clustering accuracy.

According to Table 10, because both index values are 1, the IG-SGA algorithm performs best on the Prostate dataset. The proposed LLE-HKAP algorithm performs slightly worse than the IG-SGA algorithm on the Prostate dataset. However, our algorithm performs better than other two algorithms on the Colon dataset. On the Leukemia dataset, although the LLE-HKAP algorithm shows poorer performance than the IG-SGA algorithm in terms of index S, it performs better than the HMM and IG-SGA algorithms in terms of index R. On the DLBCL dataset, the value of R is the largest for the LLE-HKAP algorithm, which indicates that our algorithm performs best. The value of S for the LLE-HKAP algorithm is smaller than that of the HMM algorithm but larger than that of the IG-SGA algorithm. In general, the experimental results verify the effectiveness of the LLE-HKAP algorithm.

Table 11 shows the three indices (recall R, specificity S and accuracy AC) of the five algorithms on the Colon and Leukemia datasets, where the APCES algorithm uses a *bicor* correlation coefficient as the similarity measure of the AP algorithm [15]. From Table 11, the LLE-HKAP algorithm shows better performance than the other four algorithms (APCES, EGSG, RSM and RF), especially on the Colon dataset. For the Colon dataset, the three evaluation indices of LLE-HKAP are obviously higher than those of the other four algorithms. LLE-HKAP is 8% higher than APCES in terms of the R and AC indices, and up to 13% higher in terms of the S index. For the Leukemia dataset, the RF algorithm performs better than the other four algorithms in terms of the R index but worse in terms of the S and AC indices. Compared with the other four algorithms, the LLE-HKAP algorithm shows better performance in terms of the three indices. The experimental results provide further evidence for the effectiveness of the proposed LLE-HKAP algorithm.

The following section describes the clustering accuracy of the proposed algorithm compared with ten clustering algorithms on five high-dimensional gene expression datasets selected from Table 1. The compared methods include two traditional clustering algorithms (HC [51], K-means [52]), two non-negative matrix factorization algorithms (C-NMF, S-NMF) [53], three subspace segmentation algorithms (LSR, LRR, LatLRR) [54]–[57] and three low rank projection least square regression (LPLSR) subspace segmentation algorithms (LPLSR-1, LPLSR-2, LPLSR) [58]. Following the experimental techniques designed by Chen *et al.* [58], five typical gene expression datasets (DLBCL, SRBCT, Leukemia1, 9-Tumor, and Prostate1) are selected from Table 1 to test the accuracy AC using the above eleven algorithms. To reduce random error, each method is run 10 times, and the results are the mean value of the clustering accuracy of the 10 evaluations. The results are shown in Table 12.

According to Table 12, the traditional clustering and NMF-based algorithms show the worst performance, and the subspace segmentation algorithms perform better than the traditional algorithms in terms of clustering accuracy. Due to the large number of data categories, the clustering accuracies of all algorithms are low on the 9-Tumor dataset, where the HC algorithm performs the worst, and nearly all data points are misclassified. However, compared with K-means and the two NMF-based algorithms, the HC algorithm shows better performance, for which the clustering accuracy is the same as that of the three subspace segmentation algorithms. Although the subspace segmentation algorithms perform better on high-dimensional datasets than the traditional clustering algorithms, it is difficult to determine the size of the subspace. It is known that an inappropriate subspace size can lead to poor results. Compared with all the above algorithms, except LLE-HKAP, the clustering accuracies of the LPLSR-1, LPLSR-2 and LPLSR algorithms were increased by more than 10% on the SRBCT, Leukemia1 and Prostate1 datasets. Nevertheless, all of the compared algorithms show lower clustering accuracy than the LLE-HKAP algorithm. On the 9-Tumor dataset

TABLE 12. The accuracy *AC* of the eleven algorithms on the five gene expression datasets.

Datasets	HC	*K*-means	C-NMF	S-NMF	LSR	LRR	LatLRR	LPLSR-1	LPLSR-2	LPLSR	LLE-HKAP
DLBCL	76.62	68.83	66.62	67.14	76.62	76.62	76.62	72.73	77.92	78.70	85.24
SRBCT	33.73	46.94	49.04	49.27	52.17	61.81	62.53	71.08	74.70	74.22	83.78
Leukemia1	52.78	55.69	54.44	62.64	65.27	62.36	68.47	79.17	80.56	88.89	89.53
9-Tumor	23.33	43.00	41.50	38.50	47.83	41.50	40.67	45.17	43.67	48.83	70.40
Prostate1	51.96	62.94	61.67	58.14	63.72	61.76	62.75	78.43	78.43	78.43	83.59

TABLE 13. Description of the five standard UCI datasets.

No.	Datasets	Samples	Attributes	Clusters
1	Iris	150	4	3
2	Wine	178	13	3
3	Glass	214	10	6
4	Seeds	210	7	3
5	Haberman	306	3	2

in particular, LLE-HKAP performs better than the other ten algorithms. In general, the LLE-HKAP algorithm is efficient in terms of clustering accuracy and outperforms the ten compared approaches. Therefore, our method is concluded to be suitable for high-dimensional data and can achieve better clustering results on gene expression datasets than the compared methods.

D. COMPARISONS OF TESTING ACCURACY ON STANDARD UCI DATASETS

In the previous experiments, it is proved that the LLE-HKAP algorithm is effective for high-dimensional gene expression datasets. To evaluate the feasibility and efficiency of the LLE-HKAP algorithm on large-scale low-dimensional UCI datasets, we conducted experiments on low-dimensional real-world datasets for practical problems that are commonly used to test the performances of clustering algorithms [59]. Several standard UCI datasets can be downloaded from the UCI repository of machine learning databases (http://www.ics.uci.edu). These datasets are described in Table 13. To compare the LLE-HKAP algorithm with the AP algorithm [26], the fireworks explosion optimization semi-supervised affinity propagation (FEO-SAP) algorithm [44] and the adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity (SAAP-SS) [60], our experimental techniques of testing the five selected UCI datasets are the same as those reported in [26], [44], and [60]. The clustering results of the four compared algorithms are shown in Table 14, where *CN* describes the cluster numbers of the four algorithms. To indicate the clustering accuracies of the four algorithms visually, Figs. 17 and 18 display histograms of the *Sil* and *FM* indices in detail, where the best performance for each dataset is highlighted. Similar to Section 4.3, the four methods are executed 10 times to reduce random error, and the results of *CN*, *Sil* and *FM* are the mean values of 10 clustering operations.

As shown in Table 14, the *CN* of the original AP algorithm does not match the actual number of clusters, while those of the other three algorithms match the actual numbers for the five UCI datasets. The evaluation results, when assessed in terms of the *Sil* and *FM* indices, show that the proposed LLE-HKAP algorithm achieves clustering performance that is superior to that of the other three algorithms. This result can be attributed to the defined similarity measure, which can more accurately describe local and global information explicitly. Although the FEO-SAP and SSAP-SS algorithms offer satisfactory performance, they can make only local adjustments to the similarity matrix due to the limited amount of a priori information. Thus, these algorithms can neither comprehensively reflect the similarities among the data points nor discover the global clustering structure of the data. As shown in Figs. 17 and 18, the proposed LLE-HKAP algorithm clearly outperforms the other three algorithms. Therefore, it can be proved that the proposed LLE-HKAP algorithm not only performs well on high-dimensional datasets but also provides favorable performance when using standard UCI datasets.

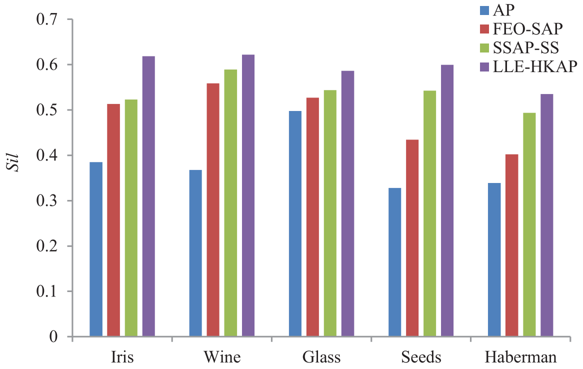


FIGURE 17. The *Sil* index of the four tested algorithms for the five standard datasets.

It is well known that, in many real-world problems, imbalance occurs when a negative class contains many more patterns than dose a positive class [61]. Note that, to date, learning from imbalanced data is still a research focus. To evaluate the classification precision of the LLE-HKAP algorithm, experiments on real-world imbalanced datasets from standard UCI datasets are performed. Information pertaining to three real-world imbalanced datasets is described in Table 15. For the three real-world imbalanced datasets, our LLE-HKAP algorithm is compared with five SVM-based machine learning methods: (1) the adaptive synthetic sampling (ADASYN) algorithm [62]; (2) the different error costs (DEC) algorithm [63]; (3) the random under-sampling

TABLE 14. Clustering results of the four algorithms on the five standard UCI datasets.

Datasets	AP			FEO-SAP			SAAP-SS			LLE-HKAP		
	CN	*Sil*	*FM*	*CN*	*Sil*	*FM*	*CN*	*Sil*	*FM*	*CN*	*Sil*	*FM*
Iris	12	0.3848	0.84	3	0.5135	0.90	3	0.5233	0.94	3	0.6186	0.95
Wine	12	0.3680	0.72	3	0.5590	0.84	3	0.5890	0.89	3	0.6223	0.91
Glass	14	0.4977	0.73	6	0.5271	0.76	6	0.5436	0.86	6	0.5864	0.90
Seeds	17	0.3280	0.79	3	0.4346	0.86	3	0.5424	0.93	3	0.5997	0.94
Haberman	31	0.3390	0.71	2	0.4021	0.76	2	0.4937	0.82	2	0.5753	0.86

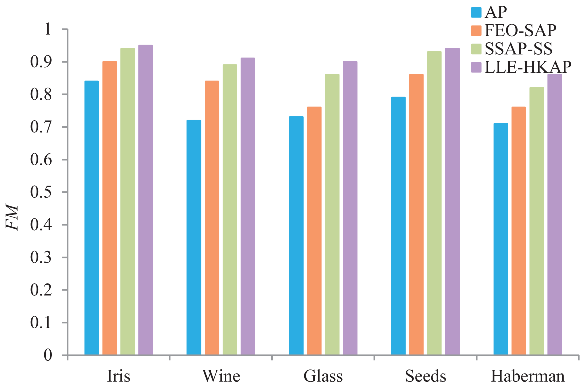


FIGURE 18. The *FM* index of the four tested algorithms for the five standard datasets.

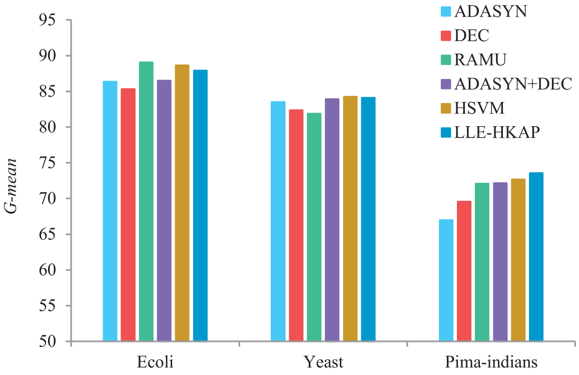


FIGURE 19. The *G-mean* index of the six tested algorithms for the three imbalanced datasets.

TABLE 15. Description of the three imbalanced datasets.

No.	Datasets	Pos.	Neg.	Total	Attributes	Imbalanced ratio
1	Ecoli (im & others)	77	259	336	7	23:77
2	Yeast (ME2 & others)	51	1433	1484	8	3:97
3	Pima-indians(1 & 0)	268	500	768	8	35:65

(RAMU) algorithm [64]; (4) the ADASYN+DEC algorithm [65]; and (5) the hybrid support vector machine (HSVM) algorithm [65]. In this experiment, following the experimental techniques designed by Liu *et al.* [65], 5-fold cross validation is adopted to ensure a fair comparison. The experimental results on real-world imbalanced datasets are presented intuitively by the histogram in Fig. 19, where the index $G\text{-}mean = \sqrt{R \times S} = \sqrt{\frac{TP\times TN}{(TP+FN)\times(TN+FP)}}$ [65] is used to estimate the classification precision of the six algorithms.

Fig. 19 shows that for the Ecoli dataset, the *G-mean* index of our LLE-HKAP algorithm is notably greater than the corresponding indices of ADASYN, DEC and ADASYN+DEC, and slightly less than those of RAMU and HSVM. On the Yeast dataset, the LLE-HKAP algorithm performs markedly better than do DEC and RAMU, and the performances of LLE-HKAP, ADASYN, ADASYN+DEC and HSVM are similar. For the Pima-indians dataset, the LLE-HKAP algorithm exhibits the best performance. It can be concluded from Fig. 19 that LLE-HKAP has the highest classification precision on most of the imbalanced datasets, and the performances of LLE-HKAP and HSVM are very similar. Furthermore, the three algorithms (LLE-HKAP, RAMU and HSVM) perform better than the ADASYN, DEC and ADASYN+DEC algorithms on the three imbalanced datasets, where DEC shows the worst performance. Therefore, the experiments show that our proposed LLE-HKAP algorithm exhibits better classification precision for imbalanced datasets and can efficiently improve the robustness of machine learning models.

Based on the abovementioned experimental results and the comparison of our scheme with other schemes, the contributions of our proposed method can be summarized as follows.

(1) The LLE-based dimension reduction algorithm is introduced to reduce the dimensions of high-dimensional datasets. Because processed low-dimensional data can maintain the original topology, the LLE algorithm can effectively reduce the dimensions of the data, which does not lose potential information.

(2) Compared with existing kernel techniques, a new global kernel function is defined in the proposed HKAP model, and the global kernel function can satisfy the conditions of the SVM kernel function. Meanwhile, our proposed global kernel function has better generalization ability. There is only one parameter in the kernel, and thus the influence of parameter adjustment can be avoided.

(3) The HKAP model is constructed based on the global and Gaussian kernels. Because of both its global and local advantages, the similarity measure calculated by our proposed hybrid kernel can yield better results than the similarity measures of methods based on traditional Euclidean distance and a single kernel. The kernel techniques can also maintain the original structure of the data, which makes the algorithm robust. Finally, the experimental results show that the combination of the LLE-based dimension reduction algorithm and the HKAP algorithm makes the proposed model able to

effectively handle high-dimensional gene expression data and standard UCI data, including real-world imbalanced data.

V. CONCLUSION

Due to their many favorable characteristics, AP clustering methods have received considerable attention in recent years. However, in the face of a growing number of high-dimensional datasets, both AP and extended AP algorithms show poor performance in terms of clustering accuracy. In this paper, an efficient LLE-HKAP algorithm for high-dimensional gene expression datasets and standard UCI datasets is presented, for which the cluster centers and the number of clusters may be unknown in advance. The first part of the algorithm reduces the dimensions of high-dimensional data and retains only the most significant data with the LLE algorithm. The second part investigates a novel HKAP algorithm. A new global kernel is defined and linearly combined with the Gaussian kernel to form a hybrid kernel, which is used to improve the similarity measure for constructing the similarity matrix in the AP algorithm. Then, the HKAP algorithm is proposed. To evaluate the accuracy of the LLE-HKAP method, several evaluation indices were introduced. The results of comparative experiments on several gene expression datasets and UCI datasets indicate that the LLE-HKAP algorithm can provide the correct number of clusters and perform well in terms of the number of iterations, the operation time and the clustering accuracy. Therefore, high-dimensional gene expression data and standard UCI data can be meaningfully analyzed using the presented LLE-HKAP algorithm.

REFERENCES

[1] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[2] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognit.*, vol. 66, pp. 364–374, Jun. 2017.

[3] J. Hu, T. R. Li, C. Luo, H. Fujita, and Y. Yang, "Incremental fuzzy cluster ensemble learning based on rough set theory," *Knowl.-Based Syst.*, vol. 132, pp. 144–155, Sep. 2017.

[4] P. Li, H. Ji, B. Wang, Z. Huang, and H. Li, "Adjustable preference affinity propagation clustering," *Pattern Recognit. Lett.*, vol. 85, pp. 72–78, Jan. 2017.

[5] H. K. Diem, V. D. Trung, N. T. Trung, V. V. Tai, and N. T. Thao, "A differential evolution-based clustering for probability density functions," *IEEE Access*, vol. 6, pp. 41325–41336, 2018.

[6] K. M. Kumar and A. R. M. Reddy, "An efficient k-means clustering filtering algorithm using density based initial cluster centers," *Inf. Sci.*, vols. 418–419, pp. 286–301, Dec. 2017.

[7] L. Xu, S. Ding, X. Xu, and N. Zhang, "Self-adaptive extreme learning machine optimized by rough set theory and affinity propagation clustering," *Cogn. Comput.*, vol. 8, no. 4, pp. 720–728, 2016.

[8] H. Q. Truong, L. T. Ngo, and W. Pedrycz, "Granular fuzzy possibilistic C-means clustering approach to DNA microarray problem," *Knowl.-Based Syst.*, vol. 133, pp. 53–65, Oct. 2017.

[9] I. A. Pagnuco, J. I. Pastore, G. Abras, M. Brun, and V. L. Ballarin, "Analysis of genetic association using hierarchical clustering and cluster validation indices," *Genomics*, vol. 109, nos. 5–6, pp. 438–445, 2017.

[10] J. Xu, G. Y. Wang, and W. H. Deng, "DenPEHC: Density peak based efficient hierarchical clustering," *Inf. Sci.*, vol. 373, pp. 200–218, Dec. 2016.

[11] T. Denoeux, S. Sriboonchitta, and O. Kanjanatarakul, "Evidential clustering of large dissimilarity data," *Knowl.-Based Syst.*, vol. 106, pp. 179–195, Aug. 2016.

[12] 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.

[13] F. Li, Y. Qian, J. Wang, and J. Liang, "Multigranulation information fusion: A Dempster-Shafer evidence theory-based clustering ensemble method," *Inf. Sci.*, vol. 378, pp. 389–409, Feb. 2017.

[14] J. Hu, T. Li, H. Wang, and H. Fujita, "Hierarchical cluster ensemble model based on knowledge granulation," *Knowl.-Based Syst.*, vol. 91, pp. 179–188, Jan. 2016.

[15] J. Meng, H. Hao, and Y. Luan, "Classifier ensemble selection based on affinity propagation clustering," *J. Biomed. Inform.*, vol. 60, pp. 234–242, Apr. 2016.

[16] 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.

[17] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.

[18] R. A. Fisher, "The statistical utilization of multiple measurements," *Ann. Hum. Genet.*, vol. 8, no. 4, pp. 376–386, 2012.

[19] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[20] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, no. 6, 2001, pp. 585–591.

[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[22] Y. Zhang, Y. Fu, Z. Wang, and L. Feng, "Fault detection based on modified kernel semi-supervised locally linear embedding," *IEEE Access*, vol. 6, pp. 479–487, 2017.

[23] R. Hettiarachchi and J. F. Peters, "Multi-manifold LLE learning in pattern recognition," *Pattern Recognit.*, vol. 48, no. 9, pp. 2947–2960, 2015.

[24] F. Castelli, M. Brambilla, A. Gatti, F. Prati, and L. A. Lugiato, "The LLE, pattern formation and a novel coherent source," *Eur. Phys. J. D*, vol. 71, no. 4, p. 84, 2017.

[25] Z. Tang, L. Ruan, C. Qin, X. Zhang, and C. Yu, "Robust image hashing with embedding vector variance of LLE," *Digit. Signal Process.*, vol. 43, pp. 17–27, Aug. 2015.

[26] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[27] G. J. Gan and M. K.-P. Ng, "Subspace clustering using affinity propagation," *Pattern Recognit.*, vol. 48, no. 4, pp. 1455–1464, 2015.

[28] C.-Q. Xia, K. Han, Y. Qi, Y. Zhang, and D.-J. Yu, "A self-training subspace clustering algorithm under low-rank representation for cancer classification on gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1315–1324, Jul. 2018, doi: 10.1109/TCBB.2017.2712607.

[29] J.-J. Li, F. Alzami, Y.-J. Gong, and Z. Yu, "A multi-label learning method using affinity propagation and support vector machine," *IEEE Access*, vol. 5, pp. 2955–2966, 2017.

[30] I. E. Givoni, C. Chung, and B. J. Frey, "Hierarchical affinity propagation," in *Proc. 27th Conf. Uncertainty Artif. Intell.*, 2011, pp. 238–246.

[31] F. Shang, L. C. Jiao, J. Shi, F. Wang, and M. Gong, "Fast affinity propagation clustering: A multilevel approach," *Pattern Recognit.*, vol. 45, no. 1, pp. 474–486, 2012.

[32] Y. Wang and L. Chen, "K-MEAP: Multiple exemplars affinity propagation with specified k clusters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2670–2682, Dec. 2016.

[33] K. Zhang and X. S. Gu, "An affinity propagation clustering algorithm for mixed numeric and categorical datasets," *Math. Problems Eng.*, vol. 2014, Art. no. 486075, Sep. 2014.

[34] W. Hang, F.-L. Chung, and S. Wang, "Transfer affinity propagation-based clustering," *Inf. Sci.*, vol. 348, pp. 337–356, Jun. 2016.

[35] J. Liu, X.-D. Zhao, and Z.-H. Xu, "Identification of rock discontinuity sets based on a modified affinity propagation algorithm," *Int. J. Rock Mech. Mining Sci.*, vol. 94, pp. 32–42, Apr. 2017.

[36] F. Yuan, X. Xia, J. Shi, H. Li, and G. Li, "Non-linear dimensionality reduction and Gaussian process based classification method for smoke detection," *IEEE Access*, vol. 5, pp. 6833–6841, 2018.

[37] G. Wang and J. F. Jiao, "Nonlinear fault detection based on an improved kernel approach," *IEEE Access*, vol. 6, pp. 11017–11023, 2018.

[38] K. Cheng, Z. Lu, Y. Wei, Y. Shi, and Y. Zhou, "Mixed kernel function support vector regression for global sensitivity analysis," *Mech. Syst. Signal Process.*, vol. 96, pp. 201–214, Nov. 2017.

[39] C.-Y. Yeh, C.-W. Huang, and S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2177–2186, 2011.

[40] Y. Wang, X. Liu, Y. Dou, Q. Lv, and Y. Lu, "Multiple kernel learning with hybrid kernel alignment maximization," *Pattern Recognit.*, vol. 70, pp. 104–111, Oct. 2017.

[41] B.-Y. Sun, X.-M. Zhang, J. Li, and X.-M. Mao, "Feature fusion using locally linear embedding for classification," *IEEE Trans. Neural Netw.*, vol. 21, no. 1, pp. 163–168, Jan. 2010.

[42] L. Sun, J. C. Xu, W. Wang, and Y. Ying, "Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification," *Genet. Mol. Res.*, vol. 15, no. 3, p. 15038990, 2016.

[43] N. Cristianini and J.-S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[44] L. Wang, X. Han, and Q. Ji, "Semi-supervised affinity propagation clustering algorithm based on fireworks explosion optimization," in *Proc. Int. Conf. Manage. E-Commerce E-Government*, Oct. 2015, pp. 273–279.

[45] H. Salem, G. Attiya, and N. EI-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput.*, vol. 50, pp. 124–134, Jan. 2017.

[46] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, "Gene selection for tumor classification using neighborhood rough sets and entropy measures," *J. Biomed. Inform.*, vol. 67, pp. 59–68, Mar. 2017.

[47] 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.

[48] H. 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, 2010.

[49] 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.

[50] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[51] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[52] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. Roy. Stat. Soc. C (Appl. Stat.)*, vol. 28, no. 1, pp. 100–108, 1979.

[53] 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.

[54] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[55] 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.

[56] 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.* New York, NY, USA: Springer-Verlag, 2012, pp. 347–360.

[57] G. C. Liu and S. C. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 1615–1622.

[58] X. Y. Chen, B. S. Xiao, and L. Y. Lin, "Low rank projection least square regression subspace segmentation for gene expression data," *Chin. Pattern Recognit. Artif. Intell.*, vol. 30, no. 2, pp. 106–116, 2017.

[59] Y. H. Liu, Z. M. Ma, and F. Yu, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy," *Knowl.-Based Syst.*, vol. 133, pp. 208–220, Oct. 2017.

[60] L. Wang, Q. Ji, and X. Han, "Adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity," *Tech. Gazette*, vol. 23, no. 2, pp. 425–435, 2016.

[61] C. Zhu and Z. Wang, "Entropy-based matrix learning machine for imbalanced data sets," *Pattern Recognit. Lett.*, vol. 88, pp. 72–80, Mar. 2017.

[62] 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.

[63] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. AI*, 1999, pp. 55–60.

[64] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 31–80, 2016.

[65] 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.

LIN SUN received the M.S. degree in computer science and technology from Henan Normal University, China, in 2007, and the Ph.D. degree in pattern recognition and intelligent systems from the Beijing University of Technology in 2015. He became a Post-Doctoral Researcher with the Medical and Biological Engineering Research Group, Henan Normal University, in 2016, where he is currently an Associate Professor with the College of Computer and Information Engineering. He has received funding from 10 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 co-authored over 70 articles. His main 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 main 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 over 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 completed the post-doctoral studies at the School of Computer Science and Technology, Tianjin University, Tianjin, China. He is currently with the College of Computer and Information Engineering, Henan Normal University, China. He has authored more than 10 peer-reviewed papers and has served as a reviewer in several prestigious peer-reviewed international journals. His research interests include knowledge discovery and machine learning.

YUN TIAN received the B.Sc. degree in computer science and technology from Henan Normal University in 2003 and the Ph.D. degree in signal and information processing from Northwestern Polytechnic University in 2007. He is currently an Associate Professor with the College of Information Science and Technology, Beijing Normal University. He has authored or co-authored more than 30 peer-reviewed papers. His research interests include knowledge processing and pattern recognition. He has served as a reviewer in several prestigious peer-reviewed international journals.