2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) | 978-1-6654-0553-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISKE54062.2021.9755411

Local Learning Joint with the Adaptive Graph for Subspace Representation

Yangbo Wang
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, Guangdong 518060, China
2070276074@email.szu.edu.cn

Can Gao
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, Guangdong 518060, China
2005gaocan@163.com

Jie Zhou*
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, Guangdong 518060, China
jie_jpu@163.com

Zhihui Lai
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, Guangdong 518060, China
lai_zhi_hui@163.com

***Abstract*—The technique of local learning that recognizes each sample by its predefined neighbors has been successfully applied in unsupervised learning fields. However, how to determine the proper neighbors of a sample is a challenging problem. Inappropriate selection of neighbors will drastically degrade the performance of local learning methods. In this study, a novel subspace representation method based on local learning joint with the adaptive graph (LLAG) is presented. On the one hand, by exploiting the notion of the adaptive graph and the technique of low-rank constraint, the affinity graph can be constructed iteratively, which prompts the produced subspace to well preserve the local and global structures of the data. On the other hand, the neighbors of samples involved in the local learning are produced adaptively during the optimization process. In this way, more accurate local information that is used to guide the generation of the optimal subspace can be revealed. Extensive experimental results on some benchmark data sets demonstrate the superiority of the proposed method LLAG in comparison with some existing unsupervised learning methods.**

***Index Terms*—adaptive graph, local learning, subspace representation, unsupervised learning**

I. Introduction

Clustering is a commonly used approach in the machine learning fields [1]. It aims to learn the structure information of data and group the samples into clusters, such that the samples within the identical cluster have higher affinity [2]. However, detecting the inherent structure of data, particularly in high-dimensional scenes, is challenging since no structural topologies about data are provided beforehand [3]. In addition, calculating the similarity between two samples in a high-dimensional scene may be disturbed by redundant features.

One way to relieve the above problems in the high-dimensional cases is dimensionality reduction [4], [5], i.e., transforming the original high-dimensional space into a low-dimensional one. Correspondingly, a reasonable dimensionality reduction method can eliminate the negative effects caused by irrelevant variables when calculating distances. Subspace learning [6], [7], as one kind of dimensionality reduction methods, has been widely applied to high-dimensional data clustering. The typical subspace learning methods include Linear Discriminant Analysis(LDA) [8], Locally Preserving Projection (LPP) [9], Principal Component Analysis (PCA) [10], spectrum embedding [11], [12], etc.

Spectral clustering [13]–[15] as a typical application of subspace learning is evolved from graph theory, and it produces clustering results based on low-dimensional representations. Its main idea is to represent the dataset as an undirected weighted graph [16], with the nodes denoting the data points and the weights of the edges to reflect the similarity between the corresponding data points [17]. Subsequently, a low-dimensional representation can be generated by optimizing the cut criteria, such as ratio cut [18], normalized cut [19], min-max cut [20], etc. After generating a low-dimensional representation, k-Means [21] or spectral rotation [5], [22]–[24] can be used in the backend process to yield clustering result. In fact, the affinity matrix plays a critical role in spectral clustering, which is often constructed in advance. However, unsuitable construction may not capture the intrinsic structure information among data. The error may be also accumulated due to the two separate steps involved in spectral clustering.

The local learning technique [25] is another kind of subspace learning method, which constructs a coefficient matrix for representing each data point based on its neighbors. Compared with traditional spectral clustering, the local learning technique has higher discriminability due to the use of local structural information [25], [26]. Recently, local learning methods have been used in supervised learning [25], [27], semi-supervised learning and unsupervised learning areas [26], [28]. Wu and Schölkopf [27] proposed a transductive clas-

*Corresponding Author. The work was supported in part by the National Natural Science Foundation of China (Nos.61806127, 62076164), and in part by the Bureau of Education of Foshan (No.2019XJZZ05) and Guangdong Basic and Applied Basic Research Foundation (No.2021A1515011861).

sification method via local learning regularization (LL-Reg), which constructs a local learning regularizer for each data according to its neighborhood information. Zhang et al. [29] presented a multi-view local learning regularization method for semi-supervised learning. Wu and Schölkopf [26] also introduced a Kernel Ridge Regression (KRR) method for local learning clustering. Zeng and Cheung [30], [31] utilized the feature selection method to improve the accuracy of the local learning clustering. Wang et al. [28] presented a clustering method with local and global regularization (CLGR) which obtains both local and global discriminative information for clustering. However, in these studies, the obtained coefficient matrix is sensitive to the predefined neighbors and the issue that how to determine the reasonable neighbors for each sample is still not solved.

The emergence of the notion of the adaptive graph brings a potential solution for determining the proper neighbors in local learning approaches [32], [33]. By adopting the adaptive graph, the affinity graph can be adaptively and precisely formed rather than predefined. The adaptive graph methods have achieved a wide range of applications [1], [32]–[34]. Nie et al. [32], [33] proposed the clustering and projected clustering with adaptive neighbors (CAN and PCAN) and the constrained Laplacian rank algorithm for graph-based clustering (CLR), in which the data similarity matrix is learned by assigning adaptive and optimal neighbors. Li et al. [34] introduced rank-constrained spectral clustering with flexible embedding, in which the notion of the adaptive graph is utilized to restore the block-diagonal affinity matrix of the ideal graph. Wang et al. [1] presented the spectral embedded adaptive neighbors clustering (SEANC), which utilizes the adaptive graph as a post-process for two-step clustering. All of these methods aim to construct an affinity matrix based on the adaptive graph. Although they provide a reference of adaptive strategy to adjust the neighbors dynamically in local learning approaches, the local structure information is not fully exploited in these methods.

In this study, we put forward a novel subspace representation method based on local learning joint with the adaptive graph (LLAG), in which an adaptive affinity graph can be generated and the neighbors in local learning can be adaptively determined. Due to the adaptive mechanism and local learning regularization technique are exploited, the local and global structures among data can be well detected and more precise local information for guiding the optimal subspace can be revealed. The main contributions of this study are listed as follows:

(1) By exploiting the notion of the adaptive graph and the technique of low-rank constraint, the affinity graph can be constructed iteratively rather than formed beforehand, which prompts the produced subspace to preserve the local and global structures among data. The limitation of error accumulation encountered in the two-step learning framework involving the affinity graph is alleviated.

(2) The neighbors of samples involved in the local learning are produced adaptively during the optimization process rather than predefined. In this way, the subjectivity of neighbor determination for local learning can be alleviated, and more precisely local information guiding the optimal subspace generation can be revealed.

(3) A two-stage iterative method is formed to solve the proposed optimization problem. The convergence and computational complexity are analyzed in detail. Experimental results on some benchmark data sets demonstrate the superiority of LLAG in comparison with the related clustering methods.

The rest of this paper is arranged as follows. Section 2 introduces some notations and reviews the related concepts on the adaptive graph and local learning regularization. Section 3 elaborates the local learning joint with the adaptive graph for subspace representation method and presents a two-stage iterative optimization algorithm. Experimental results are reported in Section 4. Finally, Section 5 concludes the paper and raises several issues for future work.

II. Preliminaries

In this section, we first give some notations and definitions, and then briefly review the notion of adaptive graph and local learning techniques.

A. Notations

All matrices and vectors are written in bold uppercase and lowercase letters respectively, while scalars are written in lowercase letters. For example, $\mathbf{A}$ expresses a matrix, $\mathbf{a}$ represents a vector, and a represents a scalar. $\|\cdot\|_2$ and $\|\cdot\|_F$ indicate the L_2 norm of a vector and the Frobenius norm of a matrix respectively. $Tr(\cdot)$ indicates the trace operation.

B. Adaptive graph

Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \Re^{d\times n}$, where d and n are the numbers of features and samples respectively, and $\mathbf{x}_i \in \Re^{d\times 1}$ is the ith sample.

Nie et al. [32] presented an adaptive graph method based on the assumption that the data points should have higher similarity to their neighbors, and the following objective function needs to be minimized:

$$\min_{\mathbf{S}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \, s_{ij} + \varphi s_{ij}^2 \right), \tag{1}$$
$$s.t \;\; \mathbf{s}_i^T \mathbf{1} = 1, \;\; s_{ij} \geq 0, \;\; rank(\mathbf{L}_s) = n - c,$$

where s_{ij} measures the affinity between data points $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathbf{S} \in \Re^{n\times n}$ denotes the symmetric affinity matrix, $\mathbf{s}_i \in \Re^{n\times 1}$ is the ith column of $\mathbf{S}$, c is the number of dataset $\mathbf{X}$, $\mathbf{L}_s = \mathbf{D} - \mathbf{S} \in \Re^{n\times n}$ stands for the Laplacian matrix of the symmetric affinity matrix $\mathbf{S}$, the degree matrix $\mathbf{D} \in \Re^{n\times n}$ is a diagonal matrix where the ith diagonal element is $\mathbf{D}_{ii} = \sum_{j=1}^{n} s_{ij}$.

The larger the distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ is, the smaller the value of s_{ij} is expected. The second term in (1) is used to avoid a trivial solution, where φ is a regularization parameter. The larger the value of φ, the greater the number of non-zero elements in $\mathbf{s}_i$.

The Laplacian matrix of the affinity matrix $\mathbf{S}$ has an important property [32], [33], [35]: if the affinity matrix

$\mathbf{S}$ is symmetric and nonnegative, the multiplicity c of the eigenvalue 0 in the Laplacian matrix $\mathbf{L}_s$ is equal to the number of connected components in the graph with the affinity matrix $\mathbf{S}$. For that reason, (1) can be converted to the following problem [32]:

$$\min_{\mathbf{S}} \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\|\mathbf{x}_i-\mathbf{x}_j\|_2^2 s_{ij}+\varphi s_{ij}^2+\lambda\|\mathbf{f}_i-\mathbf{f}_j\|_2^2 s_{ij}\Big), \quad (2)$$
$$s.t\ \mathbf{s}_i^T\mathbf{1}=1,\ \ s_{ij}\geq 0,\ \ \mathbf{F}^T\mathbf{F}=\mathbf{I},$$

where $\mathbf{f}_i \in \Re^{c\times 1}$ can be considered as a low-dimensional representation vector of $\mathbf{x}_i$. $\mathbf{F}=[\mathbf{f}_1,\mathbf{f}_2,...,\mathbf{f}_n]^T \in \Re^{c\times n}$ is the low-dimensional representation matrix of the original data matrix $\mathbf{X}$. As gradually increasing the weight of the cut-graph penalty term in (2), i.e., the value of λ, the number of connected components in $\mathbf{S}$ tends to be the number of clusters. Different from the previous graph-based clustering methods, the affinity matrix $\mathbf{S}$ is adaptively learned rather than being constructed in advance.

C. Local learning regularization

Wang et al. [28] proposed a local regularized predictor which is based on the assumption that the class label of each data point can be well predicted based on its neighbors and their labels [27]. For each data point $\mathbf{x}_i$, the proposed objective function is formed as:

$$\min_{\mathbf{W}_i,\mathbf{b}_i} \sum_{\mathbf{x}_j\in\Psi(\mathbf{x}_i)} \left\|\mathbf{W}_i^T(\mathbf{x}_j-\mathbf{x}_i)+\mathbf{b}_i-\mathbf{y}_j\right\|_2^2+\eta Tr(\mathbf{W}_i^T\mathbf{W}_i), \quad (3)$$

where $\mathbf{W}_i \in \Re^{d\times c}$ is the coefficient matrix of the $\mathbf{x}_i$-centered linear model, $\mathbf{b}_i \in \Re^{c\times 1}$ denotes a bias term. $\mathbf{y}_j \in \Re^{c\times 1}$ is the label assignment vector of data point $\mathbf{x}_j$. η is the regularization parameter controlling the capacity of the linear model. $\Psi(\mathbf{x}_i)=\{\mathbf{z}_1^i,\mathbf{z}_2^i,...,\mathbf{z}_k^i\}$ is a set of the k nearest neighbors of $\mathbf{x}_i$ and $\mathbf{z}_h^i \in \Re^{d\times 1}$ is the hth neighbor of $\mathbf{x}_i$. In general, the data point $\mathbf{x}_i$ tends to get a better prediction when its k closest neighbors belong to the same cluster. Conversely, when such nearest samples belong to different clusters, data point $\mathbf{x}_i$ may have a poor prediction result.

III. The proposed method

A. Motivation

Compared with the traditional spectral clustering, better clustering results are often achieved by integrating with the local learning method [27]. However, local learning methods are sensitive to the neighbors of sample, and the unsuitable selected neighbors will make the performance of local learning methods drastically degenerate. On the other hand, the affinity matrix used in the spectral clustering is often established in advance which may not be the best initialization for performing the spectral clustering.

To address the above issues, in this study, we put forward the local learning joint with the adaptive graph for aubspace representation (LLAG) mtehod, which inherit the merits of the adaptive graph and local learning regularization. More specifically, the affinity matrix is iteratively constructed by learning the neighbors of each data point, which not only preserves the local relationships but also the global geometry among data [28]. Secondly, the neighbors of samples involved in the local learning are generated adaptively during the optimization process. Therefore, the subjectivity of neighbor predefinitions for local learning can be alleviated, and more precisely local information guiding the optimal subspace generation can be revealed. By exploiting the local learning technique joint with the adaptive graph, they are beneficial for each other. The optimal subspace optimized through an adaptive graph will help to determine the precise neighbors of each sample in local learning. On the other side, the low representations approximated in local learning, which captures more local information among data, will help to optimize the affinity graph.

B. Objective function

The objective function of the proposed LLAG method is formed as:

$$\begin{aligned} minJ_{LLAG}(\mathbf{P},\mathbf{F},\mathbf{S}) =& (\sum_{i=1}^{n}\sum_{j=1}^{n}(\left\|\mathbf{P}^T\mathbf{x}_i-\mathbf{P}^T\mathbf{x}_j\right\|_2^2 s_{ij} \\ &+\varphi s_{ij}^2+\lambda\|\mathbf{f}_i-\mathbf{f}_j\|_2^2 s_{ij}) \\ &+\mu\sum_{i=1}^{n}\left\|\mathbf{W}_i^T\mathbf{x}_i-\mathbf{f}_i\right\|_2^2), \quad (4) \end{aligned}$$
$$s.t.\ \mathbf{s}_i^T\mathbf{1}=1,\ \ s_{ij}\geq 0,\ \ \mathbf{P}^T\mathbf{S}_t\mathbf{P}=\mathbf{I},\ \ \mathbf{F}^T\mathbf{F}=\mathbf{I},$$

where $\mathbf{P} \in \Re^{d\times r}$ is a projection matrix, $\mathbf{S} \in \Re^{n\times n}$ is the affinity matrix, $\mathbf{f}_i \in \Re^{c\times 1}$ is the low-dimensional representation vector of $\mathbf{x}_i$ and $\mathbf{W}_i \in \Re^{d\times c}$ is a local regular projection operator of the data point $\mathbf{x}_i$, φ is the regularization parameter, λ and μ are two balance factors. $\mathbf{S}_t=\mathbf{X}\mathbf{H}\mathbf{X}^T \in \Re^{d\times d}$ is the total scatter matrix, $\mathbf{H}=\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T \in \Re^{n\times n}$ is the centering matrix. The constraint term $\mathbf{P}^T\mathbf{S}_t\mathbf{P}=\mathbf{I}$ guarantees that the features on the projected space are statistically uncorrelated [32].

There are four parts in (4). The first part is the subspace projection term, which projects the original high-dimensional data onto a low-dimensional subspace. After projection, the higher the affinity, the smaller the distance of the data points in the subspace. The second term in (4) is used to avoid a trivial solution, and the larger the value of φ, the greater the number of non-zero elements in $\mathbf{s}_i$.

The third term is transformed from the low-rank constraint on the affinity matrix. The low-rank constraint ensures the connected components in the adaptive affinity graph approximate to the number of clusters [32], [35]. Since $\mathbf{f}_i$ can be considered as the representation of $\mathbf{x}_i$ in the low-dimensional space, this term also preserves the manifold structure of data [36].

In the fourth term, the local learning method is utilized, which aims to use local neighborhood information to improve the effectiveness of the produced subspace. By solving the following formula [11], $\mathbf{W}_i$ is constructed.

$$\min_{\mathbf{W}_i} \sum_{\mathbf{x}_j\in\Psi(\mathbf{x}_i)} \left\|\mathbf{W}_i^T\mathbf{x}_j-\mathbf{f}_j\right\|_2^2+\eta Tr(\mathbf{W}_i^T\mathbf{W}_i), \quad (5)$$

where η is a regular term coefficient to control the capacity of this linear model. By solving (5), $\mathbf{W}_i$ can be formed as an expression containing $\mathbf{F}$ and $\mathbf{Z}^i$ (The neighbors matrix with respect to $\mathbf{x}_i$). After $\mathbf{W}_i$ being optimized, its solution can be transformed back to (4) when optimizing other variables. It is worth mentioning that the neighbor set of $\mathbf{x}_i$ is not predefined. They are determined iteratively according to the subspace obtained by the projection matrix $\mathbf{P}$ in (4). In this way, when the projection matrix $\mathbf{P}$ is optimized gradually, the neighbors of each sample can be determined precisely due to the irrelevant features are removed from the original space.

C. Optimization and algorithm

Since the objective function (4) is convex with respect to $\mathbf{P}$, $\mathbf{S}$ and $\mathbf{F}$ respectively, it can be minimized by exploiting an effective two-stage iterative method [3].

Update P: When fixing all the variables except $\mathbf{P}$, (4) is simplified as:

$$\min_{\mathbf{P}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\left\| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \right\|_2^2 s_{ij} \right), \qquad (6)$$
$$s.t.\ \ \mathbf{P}^T\mathbf{S}_t\mathbf{P} = \mathbf{I}.$$

And the following equation holds:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \left\| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \right\|_2^2 s_{ij} = 2Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_s\mathbf{X}^T\mathbf{P}). \qquad (7)$$

Thus, (6) can be rewritten as the following problem:

$$\min_{\mathbf{P}^T\mathbf{S}_t\mathbf{P}=\mathbf{I}} Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_s\mathbf{X}^T\mathbf{P}). \qquad (8)$$

By using the generalized eigenvalue decomposition method, the optimal solution $\mathbf{P}$ can be composed of the r eigenvectors of ${\mathbf{S}_t}^{-1}\mathbf{X}\mathbf{L}_s\mathbf{X}^T$ corresponding to the r smallest eigenvalues.

Update S: When fixing $\mathbf{P}$ and $\mathbf{F}$, (4) is simplified as:

$$\min_{\mathbf{S}} \sum_{i=1}^{n}\sum_{j=1}^{n} (\left\| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \right\|_2^2 s_{ij} + \varphi s_{ij}^2 + \lambda \left\| \mathbf{f}_i - \mathbf{f}_j \right\|_2^2 s_{ij}), \qquad (9)$$
$$s.t.\ \ \mathbf{s}_i^T\mathbf{1} = 1, s_{ij} \geq 0.$$

Due to the problem (9) is independent for each $\mathbf{s}_i$, the corresponding optimization problem is expressed as:

$$\min_{\mathbf{s}_i} \sum_{j=1}^{n} (\left\| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \right\|_2^2 s_{ij} + \varphi s_{ij}^2 + \lambda \left\| \mathbf{f}_i - \mathbf{f}_j \right\|_2^2 s_{ij}), \qquad (10)$$
$$s.t.\ \ \mathbf{s}_i^T\mathbf{1} = 1, s_{ij} \geq 0.$$

Denote $d_{ij}^x = \left\| \mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j \right\|_2^2$ and $d_{ij}^f = \left\| \mathbf{f}_i - \mathbf{f}_j \right\|_2^2$, and $\mathbf{d}_i \in \Re^{1\times n}$ stands for the ith row vector of $\mathbf{d} = \left[d_{ij}^x + \lambda d_{ij}^f \right]_{n\times n}$, then (10) can be rewritten in vector form as [32]:

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, s_{ij}\geq 0} \left\| \mathbf{s}_i^T + \frac{1}{2\varphi}\mathbf{d}_i \right\|_2^2. \qquad (11)$$

The minimization of (11) can be solved by utilizing the KKT conditions and the Newton method [37].

Update F: When fixing $\mathbf{P}$ and $\mathbf{S}$, (4) is simplified as:

$$\min_{\mathbf{F}} \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} (\left\| \mathbf{f}_i - \mathbf{f}_j \right\|_2^2 s_{ij}) + \mu \sum_{i=1}^{n} \left\| \mathbf{W}_i^T\mathbf{x}_i - \mathbf{f}_i \right\|_2^2, \qquad (12)$$
$$s.t.\ \ \mathbf{F}^T\mathbf{F} = \mathbf{I}.$$

The closed-form solution of $\mathbf{W}_i$ can be obtained by solving (5) [11], and then $\mathbf{W}_i$ is used back to (12), the following expression holds:

$$\sum_{i=1}^{n} \left\| \mathbf{W}_i^T\mathbf{x}_i - \mathbf{f}_i \right\|_2^2 = Tr(\mathbf{F}^T\mathbf{L}_w\mathbf{F}), \qquad (13)$$

where $\mathbf{L}_w = (\mathbf{N} - \mathbf{I})^T(\mathbf{N} - \mathbf{I})$.

Let $\mathbf{Z}^i = [\mathbf{z}_1^i, \mathbf{z}_2^i, ..., \mathbf{z}_k^i] \in \Re^{d\times k}$ be the neighbors matrix of $\mathbf{x}_i$, $\mathbf{a}_i = \mathbf{x}_i^T(\mathbf{Z}^i\mathbf{Z}^{i^T} + \eta\mathbf{I})^{-1}\mathbf{Z}^i \in \Re^{1\times k}$ is an auxiliary vector, then the j-th entry of ith row of $\mathbf{N}$ as:

$$\mathbf{N}_{i\cdot}^{(j)} = \begin{cases} \mathbf{a}_i^{(h)}, & \text{if } \mathbf{x}_j \in \Psi(\mathbf{x}_i) \text{ and } \mathbf{x}_j = \mathbf{z}_h^i \\ 0, & \text{otherwise} \end{cases}. \qquad (14)$$

The vector form of the first term in (12) is

$$\lambda \sum_{i=1}^{n}\sum_{j=1}^{n} (\left\| \mathbf{f}_i - \mathbf{f}_j \right\|_2^2 s_{ij}) = 2\lambda Tr(\mathbf{F}^T\mathbf{L}_s\mathbf{F}). \qquad (15)$$

Combining (13) and (15), the problem (12) is transformed as follows:

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} Tr(\mathbf{F}^T(\mathbf{L}_w + \tau\mathbf{L}_s)\mathbf{F}), \qquad (16)$$

where $\tau = \frac{\mu}{2\lambda}$. The optimal solution $\mathbf{F}$ to the problem (16) can be composed of the c eigenvectors of $\mathbf{L}_w + \tau\mathbf{L}_s$ corresponding to the c smallest eigenvalues.

According to the obtained $\mathbf{F}$, the final clustering results can be produced by using k-Means or spectral rotation. For the sake of clarity, in Algorithm 1, we present some details of the proposed two-stage iterative algorithm.

D. Convergence and complexity analysis

In Algorithm 1, by solving the problem (8), (11) and (16), the optimal solutions for $\mathbf{P}$, $\mathbf{S}$ and $\mathbf{F}$ can be obtained. Then,

$$\begin{aligned} & J_{LLAG}(\mathbf{P}^{(t)}, \mathbf{S}^{(t)}, \mathbf{F}^{(t)}) \\ & \leq J_{LLAG}(\mathbf{P}^{(t-1)}, \mathbf{S}^{(t)}, \mathbf{F}^{(t)}) \\ & \leq J_{LLAG}(\mathbf{P}^{(t-1)}, \mathbf{S}^{(t-1)}, \mathbf{F}^{(t)}) \\ & \leq J_{LLAG}(\mathbf{P}^{(t-1)}, \mathbf{S}^{(t-1)}, \mathbf{F}^{(t-1)}), \end{aligned} \qquad (17)$$

where $t-1$ and t stand for the $(t-1)$th and tth iteration steps respectively. Inequality (17) shows that the presented method LLAG gradually converges during the iteration procedures.

The complexity of updating $\mathbf{P}$ is $O(T\left(dn^2 + d^3\right))$. The complexity of computing $\mathbf{F}$ is $O(Tn^3)$. The complexity of computing $\mathbf{S}$ is also $O(Tn^3)$. Consequently, the computation complexity of Algorithm 1 is approximate to $O(T(n^3 + dn^2 + d^3))$.

Algorithm 1 Local Learning Joint with the Adaptive Graph for Subspace Representation (LLAG) Algorithm

```
Input:
    Data matrix $\mathbf{X} \in \Re^{d \times n}$, the number of clusters $c$, reduced dimension $r$, parameters $\lambda$, $\mu$ and $\eta$, the size of the neighborhood used to construct the local regular projection operator $k$;
Output:
    Low-dimensional representation matrix $\mathbf{F} \in \Re^{n \times c}$;
Initialize $\mathbf{S}$ and $\varphi$ as the same way in solving (2);
Initialize $\mathbf{F}$ by eigenvalue decomposition of $\mathbf{L}_s = \mathbf{D} - \mathbf{S}$;
Initialize $\mathbf{P}$ by the $r$ eigenvectors of $\mathbf{X}\mathbf{L}_s\mathbf{X}^T$ corresponding to the $r$ smallest eigenvalues;
while not converge do
    For each $i$, update the $i$th row of $\mathbf{S}$ by solving (11);
    Solve $\mathbf{P}$ by (8);
    For each data point $\mathbf{x}_i$, update $\mathbf{Z}^i$ by its nearest $k$ neighbors based on P;
    Solve $\mathbf{F}$ by (16);
end while
return $\mathbf{F}$.
```

IV. Experiments

In this part, we show the performance of the proposed method (LLAG) and visualize the generated subspace representation to analyze its discriminability.

A. Datasets and comparison methods

Nine datasets from UCI repository[1], i.e., Wine, Iris, Glass, Seeds, Balance, Thyroid, Breast, Banknote and Yeast, five image datasets, COIL20[2], ORL[3], YALE[4], PIE[5] and AR[6], and two gene databases[7] Lung Cancer and Ovarian are selected for experiments. The details of these data sets are shown in Table 2.

TABLE I
Selected data sets

Datasets	Samples	Dimensions	Clusters	Data sets	Samples	Dimensions	Clusters
Wine	178	13	3	Yeast	1484	8	5
Iris	150	4	3	COIL20	1440	1024	20
Glass	214	9	2	ORL	400	1024	40
Seeds	210	7	3	YALE	165	1024	15
Balance	625	4	3	PIE	1632	1024	68
Thyroid	215	5	3	AR	2400	2000	120
Breast	683	9	2	LungCancr	181	12533	2
Banknote	1372	4	2	Ovarian	253	15154	2

Some clustering methods are selected for comparison, including k-Means [21], Spectral Clustering (SC) [15], Clustering and Projected Clustering with Adaptive Neighbors (CAN & PCAN) [32], Clustering with Local and Global Regularization (CLGR) [28] and Spectral Embedded Adaptive Neighbors Clustering (SEANC) [1].

B. Parameter setting

For SC, PCAN and LLAG, the range of low-dimensional space is tunning from 2 to $min(n, d-1)$ and for SC, CLGR and SEANC, the affinity matrix is constructed with the self-tune Gaussian method [38]. Considering that the k-means method is sensitive to the initialization, for k-means, SC, CLGR and LLAG, each method executes 10 times and the best result is chosen to represent the performance of the method.

With regard to the proposed method LLAG, the balance factors μ and λ, and the regularization parameter η, need to be given in advance. These parameters are determined by grid search. The ranges of parameters λ, μ and η are $[0.001, 0.1, 10, 100, 1000]$, $[0.002, 0.2, 2, 20, 200, 2000]$ and $[0.01, 0.1, 1, 10, 100]$, respectively. The parameter values of other compared methods are used the ones reported in their original paper.

C. Experimental results

1) Validity indices: Tables 2 and 3 show the ACC and NMI values obtained by the different methods on the selected benchmark datasets respectively. The best results are shown in bold and the second-best results are marked in brackets. The last row of Tables 2 and 3 shows the average ACC and NMI values for each method on selected datasets.

TABLE II
ACC of different methods on benchmark datasets

Datasets	k-Means	SC	CAN	PCAN	SEANC	CLGR	LLAG
Wine	70.22	73.03	72.47	73.03	(87.64)	73.03	**100.00**
Iris	89.33	90.67	90.67	(98.00)	89.33	**99.33**	**99.33**
Glass	89.72	(92.52)	87.38	90.65	(92.52)	89.72	**94.39**
Seeds	89.52	88.57	89.52	**94.76**	90.95	91.43	(94.29)
Balance	57.44	55.36	62.88	(66.88)	57.73	57.12	**74.56**
Thyroid	86.51	67.44	87.91	87.44	84.19	(90.70)	**93.95**
Breast	96.05	97.07	96.63	95.61	**97.51**	97.07	(97.36)
Banknote	61.22	71.79	89.65	(98.32)	57.73	85.28	**98.54**
Yeast	(48.38)	47.10	43.06	43.67	44.14	47.57	**53.57**
COIL20	64.17	86.39	(88.06)	87.36	83.12	84.44	**88.26**
ORL	56.00	65.25	56.75	63.25	61.75	(68.50)	**85.25**
YALE	43.64	43.64	50.91	43.03	41.21	(52.12)	**54.55**
PIE	18.50	29.53	17.28	53.19	(62.75)	32.60	**78.43**
AR	34.79	47.88	41.25	(67.58)	47.58	48.96	**74.21**
LungCancer	63.05	66.01	79.80	(90.64)	78.33	(90.64)	**91.63**
Ovarian	56.92	77.87	70.36	(80.24)	64.43	63.24	**100.00**
Avg.	64.09	68.76	70.29	(77.10)	71.31	73.23	**86.15**

From Tables 2 and 3, k-Means clustering performs poorly when dealing with high-dimensional data sets. Such as the image datasets COIL20, PIE and AR. The main reason is the redundant features have a negative impact when computing the distances between data points and cluster centers. In SC, since the construction of affinity matrix and the acquisition of low-dimensional representations are done in two stages, the constructed affinity matrix may not be optimal, which makes SC perform poorly on Thyroid, Yale and PIE. The performance of CAN is generally poorer than PCAN when dealing with high-dimensional situations, especially for the image data sets

[1] http://www.ics.uci.edu/ml

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html

[3] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[4] http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.htm

[5] http://www.flintbox.com/public/project/4742/

[6] http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html

[7] Microarray Datasets (szu.edu.cn)

ORL, PIE and AR. The reason can be attributed to the lack of dimensionality reduction by using a projection mechanism. CLGR is sensitive to the neighbors of data points, so its performance on different datasets is not stable.

In LLAG, the local learning mechanism and the adaptive graph method are involved to improve the discriminability of generated subspace representation. The experimental results demonstrate the superiority of the proposed method LLAG. Particularly, LLAG obtains the best ACC values on all selected datasets except Breast and Seeds. Regarding the data set Wine and Ovarian, LLAG achieves impressive ACC results which are equal to 1.

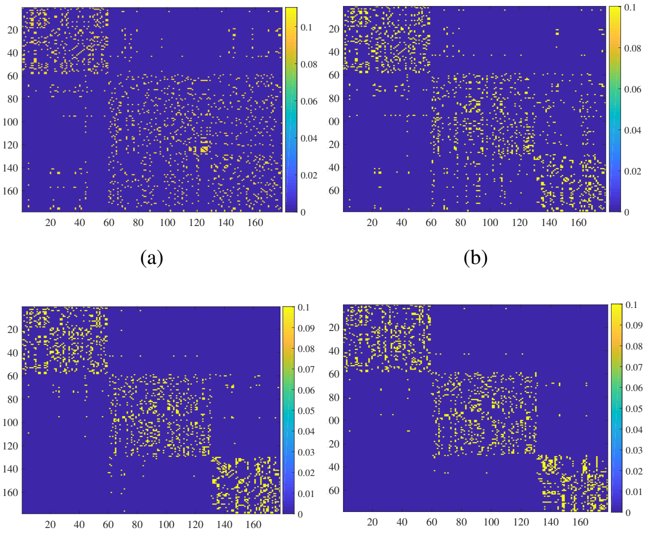
2) Visualization of affinity matrix and subspace: The affinity matrix used in LLAG is adaptively learned rather than being constructed in advance, which can capture the intrinsic structures among data during the optimization process. To observe the variation of the affinity matrix in different iterations, their results on Wine data set are visualized in Fig. 1. In Figs. 1 (a) and 1 (b), some data points belonging to different clusters have high-affinity values. With the increase of iteration steps, these data points are separated well according to their affinity values, as shown in Figs. 1 (c) and 1 (d). The main reason is that the optimal subspace can be gradually achieved, by which the difference between any two data points can be well measured.

Projection technique is utilized in LLAG, and a projection subspace is generated to determine the neighbors in local learning. To analyze the discriminability of the obtained projected space by projection matrix P, the original input space and obtained projected subspace on Wine and Ovarian are visualized in Figs. 2 and 3 by using t-SNE technique [39].

Figs. 2 (a) and 3 (a) show that there are a lot of data points that belong to different clusters close to each other in the original space. Undoubtedly, in CLGR, the neighbors chosen directly based on the distance in the original space will be poor. However, in LLAG, the original space is mapped to a projected subspace, thereby the negative impact caused by the redundant features can be reduced when calculating the similarity based on distance. According to the results displayed in Figs. 2 (b) and 3 (b), more reasonable neighbors can be selected in LLAG when comparing with other methods based on the original space..

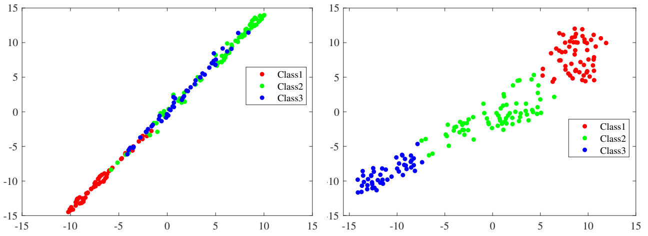
TABLE III
NMI OF DIFFERENT METHODS ON BENCHMARK DATASETS

Datasets	k-Means	SC	CAN	PCAN	SEANC	CLGR	LLAG
Wine	42.88	41.42	39.48	42.30	(66.47)	42.30	**100.00**
Iris	75.82	80.57	80.57	(91.92)	79.07	**97.02**	**97.02**
Glass	43.10	(57.98)	41.37	48.27	(57.98)	43.89	**62.83**
Seeds	69.49	69.95	69.49	(80.44)	70.44	72.42	**81.21**
Balance	(18.79)	15.81	16.01	12.36	0.22	16.49	**32.57**
Thyroid	45.34	23.82	36.93	56.60	41.61	(63.47)	**71.90**
Breast	74.78	81.12	77.52	72.83	**83.14**	80.24	(81.38)
Banknote	3.03	31.67	61.28	(89.31)	4.55	41.03	**90.37**
Yeast	17.04	(23.32)	19.47	14.22	18.16	19.31	**24.07**
COIL20	79.30	93.17	**96.10**	(95.86)	93.26	92.87	92.15
ORL	74.66	79.71	79.21	83.27	81.22	(83.96)	**92.82**
YALE	49.96	(56.18)	48.73	49.63	48.91	**56.26**	56.06
PIE	44.38	58.73	39.06	72.90	(77.81)	60.12	**85.46**
AR	65.33	73.57	66.99	(90.56)	71.63	73.21	**91.72**
LungCancer	48.00	60.49	62.10	(72.90)	60.74	71.74	**74.46**
Ovarian	1.24	25.89	15.42	(35.67)	3.27	3.25	**100.00**
Avg.	47.07	54.59	53.11	(63.07)	53.66	57.35	**77.13**



(a) (b) (c) (d)

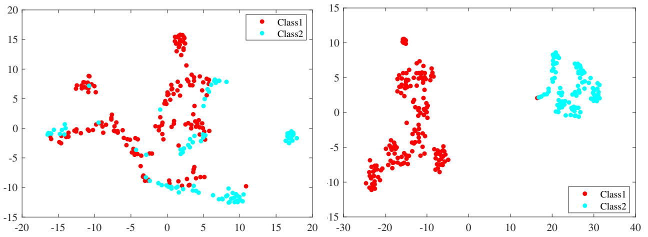
Fig. 1. The affinity matrix of Wine concerning the different number of iterations. (a) iteration step =5. (b) iteration step =10. (c) iteration step =15. (d) iteration step =20.



(a) Original data (b) LLAG

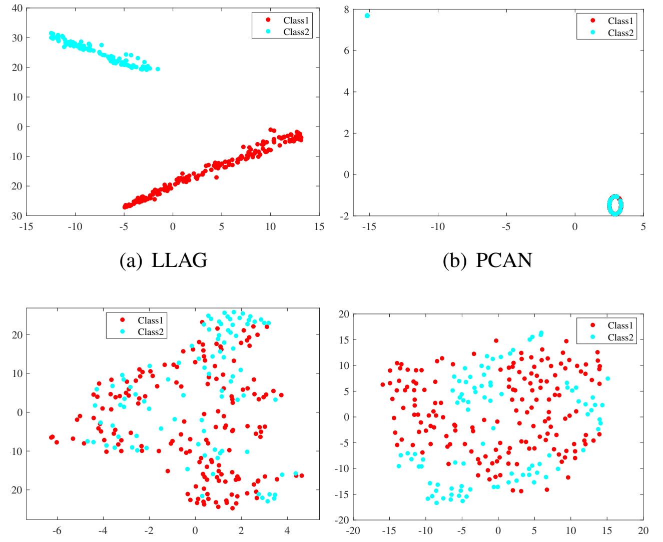
Fig. 2. The visualizations of Wine and projected subspace obtained by LLAG using t-SNE technique.

The visualizations of the final subspace representation for clustering obtained by LLAG, PCAN, CLGR and SC on Ovarian data set are displayed in Fig. 4. Obviously, the samples coming from different clusters are separated well by LLAG. However, this situation is not shown in other three methods, i.e., some samples from different clusters are still mixed together, which makes the final clustering performance poor. In this way, the discriminability of the subspace representation



(a) Original data (b) LLAG

Fig. 3. The visualizations of Ovarian data and projected subspace obtained by LLAG on Ovarian using t-SNE technique.

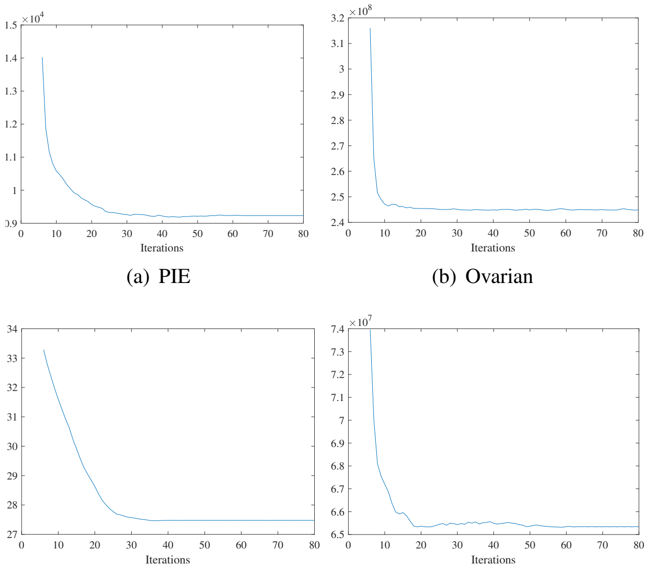


(c) CLGR (d) SC

Fig. 4. The visualizations of subspace representation $\mathbf{F}$ obtained by LLAG, PCAN, CLGR and SC on Ovarian using t-SNE technique.

obtained by LLAG is better than the ones obtained by other compared methods.

3) Convergence : Fig. 5 shows the convergence curves of LLAG on some data sets. This shows that the proposed method LLAG converges approximately after 30 iterations.



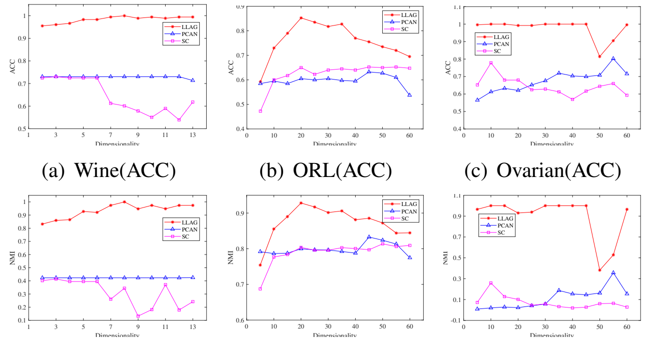
(a) PIE (b) Ovarian

(c) Iris (d) ORL

Fig. 5. The convergence curves of LLAG on different data sets.

4) Parameter sensitivity: Since the subspace learning techniques are involved in SC, PCAN and LLAG, their performances with different values of r on Wine, ORL and Ovarian are shown in Fig. 6.

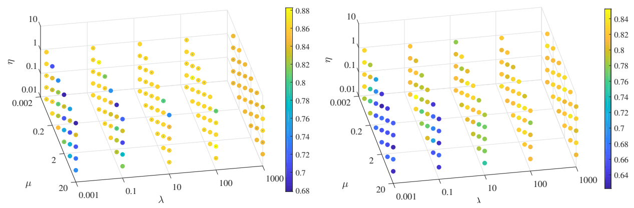
Compared to the other two methods, LLAG almost achieves the best results no matter what the values of r are. It further demonstrats that the discriminability of subspace produced in LLAG is better than the ones obtained by SC and PCAN.



(a) Wine(ACC) (b) ORL(ACC) (c) Ovarian(ACC)

(d) Wine(NMI) (e) ORL(NMI) (f) Ovarian(NMI)

Fig. 6. The ACC and NMI variations concerning the different numbers of subspace dimensions on Wine, ORL and Ovarian.



(a) COIL20 (b) ORL

Fig. 7. The ACC variations concerning different value sets (λ, μ, η). (a): COIL20 when fixing parameter $r = 40$. (b): ORL when fixing parameter $r = 20$.

There are three parameters (i.e., λ, μ and η) in LLAG when r is fixed. λ controls the weight of the manifold regularization term, μ controls the weight of the local projection learning term and η determines the capacity of the linear model. As shown in Figs. 7 (a) and 7 (b), when the parameter $\lambda \geq 10$, the ACC values are usually higher, yet when $\lambda \leq 0.1$, the ACC values are usually lower. The reason is when the parameter λ is enough large, the third term in (4) is approximately equal to 0, which introduces a latent low-rank constraint onto the Laplacian matrix that with respect to the learned affinity matrix. The low-rank constraint ensures the connected components in the adaptive affinity graph are approximately equal to the cluster number, which declines the loss of the post process. It also can be found that when μ is assigned with a relatively large value, such as $\mu = 2000$, the ACC values vary with the parameter λ and they are less affected by η. So LLAG is not sensitive to the parameter η.

V. Conclusions

In this study, a novel subspace representation method based on local learning joint with the adaptive graph (LLAG) is presented, in which an affinity graph is iteratively generated and the neighbors of each sample involved in local learning are also adaptively determined. Due to the adaptive mechanism and local learning regularization technique are exploited, the local and global structures among data can be well detected and more precise local information for guiding the optimal subspace can be revealed. The extensive experiments on sixteen benchmark data sets demonstrate the superiority of the proposed method LLAG when comparing with some state-of-the-art methods. Detecting and integrating the discriminative information between intra and inter clusters in the unsupervised learning areas may further improve the effectiveness of the proposed model which is the future work.

References

[1] Q. Wang and Z. Q. Qin, F. P. Nie, and X. L. Li, "Spectral embedded adaptive neighbors clustering," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 4, pp. 1265-1271, 2019.

[2] M. L. Chen, Q. Wang, and X. L. Li, "Adaptive projected matrix factorization method for data clustering," Neurocomputing, vol. 306, pp. 182-188, 2018.

[3] J. Zhou, W. Pedrycz, X. D. Yue, C. Gao, Z. H. Lai, et al., "Projected fuzzy c-means clustering with locality preservation," Pattern Recognition, vol. 113, 2021.

[4] G. Q. Wen, Y. H. Zhu, and W. Zheng, "Spectral representation learning for one-step spectral rotation clustering," Neurocomputing, vol. 406, pp. 361-370, 2020.

[5] G. Q. Wen, X. X. Li, Y. H. Zhu, L. J. Chen, Q. M. Luo, et al., "One-step spectral rotation clustering for imbalanced high-dimensional data," Information Processing & Management, vol. 58, no. 1, 2021.

[6] F. D. L. Torre and M. J. Black, "A framework for robust subspace learning," International Journal of Computer Vision, vol. 54, no. 1, pp. 117-142, 2003.

[7] C. P. Hou, F. P. Nie, Y. Y. Jiao, C. S. Zhang, and Y. Wu, "Learning a subspace for clustering via pattern shrinking," Information processing & management, vol. 49, no. 4, pp. 871-883, 2013.

[8] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," Institute for Signal and information Processing, vol. 18, no. 1998, pp. 1-8, 1998.

[9] X. F. He and P. Niyogi, "Locality preserving projections," Advances in Neural Information Processing Systems, vol. 16, no. 16, pp. 153-160, 2003.

[10] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1-3, pp. 37-52, 1987.

[11] F. P. Nie, D. Xu, I. W. Tsang, and C. S. Zhang, "Spectral embedded clustering," Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, pp. 1181–1186, 2009.

[12] F. P. Nie, Z. N. Zeng, I. W. Tsang, D. Xu, and C. S. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," IEEE Transactions on Neural Networks, vol. 22, no. 11, pp. 1796-1808, 2011.

[13] T. Semertzidis, D. Rafailidis, M. G. Strintzis, and P. Daras, "Large-scale spectral clustering based on pairwise constraints," Information Processing & Management, vol.51, no. 5, pp. 616-624, 2015.

[14] A. G. Chifu, F. Hristea, J. Mothe, and M. Popescu, "Word sense discrimination in information retrieval: A spectral clustering-based approach," Information Processing & Management, vol.51, no. 2, pp. 16-31, 2015.

[15] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in Neural Information Processing Systems, vol. 2, pp. 849-856, 2002.

[16] F. Wang, C. S. Zhang, and T. Li, "Regularized clustering for documents," Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, pp. 95–102, 2007.

[17] H. X. Zhang and L. L. Cao, "A spectral clustering based ensemble pruning approach," Neurocomputing, vol. 139, pp. 289-297, 2014.

[18] P. K. Chan, M. D. Schlag, and J. Y. Zien, "Spectral k-way ratio-cut partitioning and clustering," IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, vol. 13, no. 9, 1994.

[19] J. B. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, 2000.

[20] F. P. Nie, C. Ding, D. J. Luo, and H. Huang, "Improved minMax cut graph clustering with nonnegative relaxation," Machine Learning and Knowledge Discovery in Databases, European Conference, Barcelona, Spain, pp. 451–466, 2010.

[21] J. Macqueen, "Some methods for classification and analysis of multivariate observations", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.

[22] X. Y. Stella, and J. B. Shi, "Multiclass spectral clustering," 9th IEEE International Conference on Computer Vision (ICCV 2003), Nice, France, pp. 313–319, 2003.

[23] J. Huang, F. P. Nie, and H. Huang, "Spectral rotation versus k-Means in spectral clustering," Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, USA, pp. 431–437, 2013.

[24] Y. W. Pang, J. Xie, F. P. Nie, and X. L. Li, "Spectral clustering by joint spectral embedding and spectral rotation," IEEE transactions on cybernetics, vol. 50, no. 1, pp. 247-258, 2018.

[25] L. Bottou and V. Vapnik, "Local learning algorithms," Neural Computation, vol.4, no. 6, pp. 888–900, 1992.

[26] M. R. Wu and B. Schölkopf, "A local learning approach for clustering," Advances in Neural Information Processing Systems, vol. 19, pp. 1529-1536, 2006.

[27] M. R. Wu and B. Schölkopf, "Transductive Classification via Local Learning Regularization," Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, San Juan, Puerto Rico, pp. 628–635,2007.

[28] F. Wang, C. S. Zhang, and T. Li, "Clustering with local and global regularization," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 12, pp. 1665-1678, 2009.

[29] D. Zhang, F. Wang, C. S. Zhang, and T. Li, "Multi-view local learning," Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, pp. 752–757, 2008.

[30] H. Zeng and Y. M. Cheung, "Feature selection for local learning based clustering," Advances in Knowledge Discovery and Data Mining, Bangkok, Thailand, pp. 414–425, 2009.

[31] H. Zeng and Y. M. Cheung, "Feature selection and kernel learning for local learning-based clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1532-1547, 2011.

[32] F. P. Nie, X. Q. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, pp. 977–986, 2014.

[33] F. P. Nie, X. Q. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, pp. 1969–1976, 2016.

[34] Z. H. Li, F. P. Nie, X. J. Chang, L. Q. Nie, H. X. Zhang, et al., "Rank-constrained spectral clustering with flexible embedding," IEEE transactions on neural networks and learning systems, vol. 29, no. 12, pp. 6073-6082, 2018.

[35] B. Mohar, Y Alavi, G. Chartrand, and O. Oellermann, "The Laplacian spectrum of graphs," Graph theory, combinatorics, and applications, vol.2, no.871-898, pp. 12, 1991.

[36] J. Zhou, W. Pedrycz, C. Gao, Z. H. Lai, J. Wan, et al., "Robust jointly sparse fuzzy clustering with neighborhood structure preservation,", IEEE Transactions on Fuzzy Systems, 2021., in press.

[37] J. Huang, F. P. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, pp. 3569–3575, 2015.

[38] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," Advances in Neural Information Processing Systems 17, Vancouver, British Columbia, Canada, pp. 1601–1608, 2004.

[39] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research, vol. 9, no. 11, pp. 2579-2965, 2008.