Neural Processing Letters (2021) 53:257–273
https://doi.org/10.1007/s11063-020-10383-9

Joint Spectral Clustering based on Optimal Graph and Feature Selection

Jinting Zhu[1] · **Julian Jang-Jaccard**[1] · **Tong Liu**[1] · **Jukai Zhou**[1]

Accepted: 22 October 2020 / Published online: 18 November 2020

Abstract
Redundant features and outliers (noise) included in the data points for a machine learning clustering model heavily influences the discovery of more distinguished features for clustering. To solve this issue, we propose a spectral new clustering method to consider the feature selection with the $L_{2,1}$-norm regularization as well as simultaneously learns orthogonal representations for each sample to preserve the local structures of data points. Our model also solves the issue of out-of-sample, where the training process does not output an explicit model to predict unseen data points, along with providing an efficient optimization method for the proposed objective function. Experimental results showed that our method on twelve data sets achieves the best performance compared with other similar models.

Keywords Feature selection · Clustering · Graph matrix, dimensionality reduction, subspace learning

1 Introduction

Clustering is one of a number of unsupervised learning techniques, and its major aim is to group similar features from unlabelled data together and to explore the value information of data [1]. The obtained information can be applied in various applications in the real world, i.e., data predication [2], business intelligence [3], pattern recognition [4], medical diagnosis [5]. To cope with different kinds of clustering tasks, existing methods are usually classified into these categories, non-graph-based methods [6], graph-based methods [7], and deep learning based methods [8]. In recent years, we have witnessed a number of deep learning techniques

✉ Jinting Zhu
j.zhu3@massey.ac.nz

Julian Jang-Jaccard
j.jang-jaccard@massey.ac.nz

Tong Liu
t.liu@massey.ac.nz

Jukai Zhou
z.jukai@massey.ac.nz

[1] Comp Sci/Info Tech, Massey University, Auckland, New Zealand

successfully applied to various applications. These models are often incorporated with a large number of labeled samples to avoid overfitting often associated with the use of a small number of samples. However, the collection of thousands and millions of labelled samples is often labor intensive and not realistic in many application domains (e.g., cybersecurity). In contrast, many traditional learning methods are often better equipped to work well even if an input sample size is small and provide a better interpretable insight concerning the model as well as the interaction with the input sample.

K-means belongs to a class of non-graph-based method. It is relatively simple to implement and produces good clustering performance [9]. It randomly generates cluster centers at initialization so that the results obtained are difficult to reproduce. Also, the selected clusters by the k-means method are dependent on the average values of original data, which may be highly influenced by outliers. Hence, the accuracy of k-means is greatly limited by the distance measurement and the distribution of original data. In order to address this problem, the spectral clustering method, which is a part of a graph-based method, constructs the similar graph representation for the original data points and then learns an algorithm to partition the graph into several reasonable sub-clusters [10]. Specifically, it creates a similarity graph structure between data points, and calculates the first k eigenvectors of the corresponding Laplacian matrix to define a feature vector for each object, so several clusters are obtained by using k-means method. Compared with the k-means method, spectral clustering method can efficiently proceed with the dimensionality reduction on data points by Laplacian eigenmaps. Besides, it is more adaptable to data distribution, and the calculation amount is much smaller while the graph structure is not complex. By these reasons, it has been applied to various tasks, such as classification [6] and segmentation [11].

Much literature [7,12–17] has worked on solving the problem of learning a graph representation for spectral clustering. For example, Yan et al. [12] employed the graph embedding framework to proceed with dimensionality reduction while the penalty graph constructs the relation with marginal points. Peng et al. [14] proposed a subspace learning technique which gives a L_2-graph structure aiming to eliminate the effects of outliers and preserve the information between data points in the same subspace. However, these methods only construct the graph matrix on the original data without considering the robust graph embedding framework so that both outliers and redundant features still affect the clustering performance. Furthermore, many approaches focus on either resolving feature selection [18] or conducting experiment with out-of-sample extension [19]. That is, the generalization of the learnt embedding with feature selection to new samples that are not considered. For example, Zhu et al. [20] proposed a regularized self-representation (RSR) model for unsupervised feature selection, where each feature can be represented by the linear combination of relevant features. Vural et al. [21] developed a semi-supervised method for building an interpolation function that provides an out-of-sample extension [22–25]. Motivated by the above observation, one can integrate feature selection into the graph embedding framework, to yield both the robust performance and interpretation ability.

In this paper, we propose a new joint graph embedding subspace with feature selection (JSCGFS) clustering method to address the above limitations. We first learn the indicator matrix from the low-dimensional space, and employ $L_{2,1}$-norm work [26] on both loss function and projection matrix for improving the effectiveness of indicator learning. We then conduct Laplacian matrix on the indicator matrix for final clustering analysis. Furthermore, we devise an alternative strategy to solve the proposed objective function. Experimental results on twelve real-world benchmark data sets demonstrate that our method outperforms the comparison clustering algorithms regarding three evaluation metrics, such as accuracy (ACC), normalized mutual information (NMI), and Purity.

The contributions of this paper are summarized as follows:

– We develop a new model to consider both clustering and feature selection. Feature selection can select more reliable features for describing samples for clustering analysis. Meanwhile, the results obtained by clustering can be fed back for improvement in feature selection so that the two steps can mutually interact to achieve the best optimal solutions.
– We propose a reasonable constraint to guarantee each sample is extremely relevant to itself, and we provide a new method to optimize the proposed method. The employed experiments on twelve public data sets demonstrated that the proposed clustering method outperforms both spectral clustering method and k-means clustering.

The remainder of this paper is as follows. Section 2 reviews the studies of graph-based methods and spectral feature selection methods. Section 3 introduces the details of the proposed method and optimization process. Section 4 illustrates the results of the experiments, followed by the conclusion.

2 Related Work

In this section, we introduce existing studies related to our method, notably in the area of graph-based methods and spectral feature selection method. Table 1 provides a summary of the related methods included in this section.

2.1 Graph Based Methods

Graph-based methods [27] usually build a similarity matrix on training data to represent the high-order relationship among samples or data points. The details of the inner structure of the data set can be weighted by the graph. The new graph representation can be obtained by the optimal solution of graph cutting [28]. For example, spectral clustering is a classical algorithm of graph-based clustering and is adaptable to data distribution. Furthermore, many researchers have worked on learning a graph structure, for example, Nie et al. [29] have proven that a graph can be adjusted during the clustering procedure. The graph contains the k clusters and generate two new graph-based clustering objectives based upon the L_1-norm [30]

Table 1 Overview of the existing studies related to our method

Graph based		Clustering method
Ref.	Application	
[12]	Image search	Performing Markov random walk in an image graph
[29]	Bioinformatics and image	Constrained Laplacian Rank
[28]	Jinsight trace	Partitioning Heuristics
Spectral feature selection		**Clustering method**
Ref.	Application	
[32]	Multivariate data	Diagonally rescaled gradient descent
[33]	Image clustering	A lowest-rank representation
[34]	Face, object, handwritten	Nonnegative sparse graph

and L_2-norm [31]. Yan et al. [12] defined a correlation graph to construct the relationships between word and image, and using the complex graph clustering and spectral clustering to cluster images into topics. However, current graph partitioning methods based on the original data sets have not yet provided any generally satisfactory results.

2.2 Spectral Feature Selection Methods

The goal of spectral feature selection is to explore the potential structure of features from the original data set and remove both redundancy and noise. Currently, methods tend to conduct feature selection that combines with graph embedding, such as PCA [35–38], Sparse coding [34,39–42] and Low-rank [24,33,43–45]. These approaches are applied in different areas and specific tasks including image retrieve, face recognition, and data mining. Specifically, PCA-based methods based on the graph embedding usually learn an efficient PCA method against the effect generated by the outliers and noise. For example, Feng et al. [37] designed a model which integrating the PCA using the l_p-norm for reducing outliers and noise. The low-rank method based on the graph embedding [32] generally applies NMF on high-dimensional space for computing the effective representation of the original data points. For example, Yin et al. [33] proposed a hyper graph-Laplacian regularization which can be integrated with a non-negative sparse hyper-Laplacian regularized model. Furthermore, the sparse coding methods based on the graph embedding takes the data as a dictionary, and conducts sparse coding to preserve feature characteristics. For example, Fang et al. [34] assumed that non-negative sparse graph learning method which integrate the label prediction with projection into the dictionary.

3 Approach

We now define the notations and discuss the details of our proposed algorithm JSCGFS. We also describe the theory of spectral clustering combing feature selection on the new sparse coding. We denote matrices, vectors, and scalars, respectively, as boldface uppercase letters, boldface lowercase letters, and normal italic letters. We summarize other notations used in this paper in Table 2.

Table 2 The used notations in this paper

$\mathbf{X}$	The feature matrix of the training data	$\mathbf{x}$	A vector of $\mathbf{X}$
$\mathbf{x}^i$	The i-th row of $\mathbf{X}$	$\mathbf{x}^j$	The j-th column of $\mathbf{X}$
$\|\|\mathbf{X}\|\|_\mathbf{F}$	The Frobenius norm of $\mathbf{X}$, $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} x_{i,j}^2}$	$\mathbf{x}^i$	The i-th row of $\mathbf{X}$
$\|\|\mathbf{X}\|\|_{\mathbf{2,1}}$	The $\ell_{2,1}$-norm of $\mathbf{X}$, $\|\mathbf{X}\|_{2,1} = \sum_i \sqrt{\sum_j x_{i,j}^2}$	$tr(\mathbf{X})$	The trace of $\mathbf{X}$
$x_{i,j}$	The element in i-th row and the j-th column of $\mathbf{X}$	$\mathbf{X}^T$	The transpose of $\mathbf{X}$

3.1 Joint Spectral Clustering Preliminaries

3.1.1 K-means

Given the data $\mathbf{X} = \{x_i\}$, the set of n dimensional points is clustered into a set of k clusters, denoted as $c_k, k = \{1, \ldots, K\}$. The algorithm measures the squared error and uses a partition to minimize the distance. For example, let u_k be the mean of cluster c_k. The distance in squared error between u_k and the points of c_k is as follows:

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - u_k\|^2 \tag{1}$$

To minimize the objective function, our model sums the squared error across all k clusters as follows,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - u_k\|^2 \tag{2}$$

Intuitively through observing the equation, we know that the u_k is influenced by the x_i, especially if there are outliers existed in the $\mathbf{X}$.

3.1.2 Spectral Clustering

Spectral clustering connects the local relationship and constructs the similarity graph by performing the dimensionality reduction before clustering. Specifically, given x_n as the feature matrix could be extracted by arbitrary objects, the relationship among data points could be denoted by a similarity matrix, $\mathbf{S} = (s_{ij})\{i, j = 1, \ldots, n\}$. The s_{ij} denoted as the similarity between the corresponding data points x_i and x_j, their edge is weighted by the w_{ij}. We assume that $\mathbf{S}$ is undirected, and w_{ij} is a weight, where $w_{ij} = w_{ji} \geq 0$. Thus, the most important object is the graph Laplacian matrix, which is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{S} \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{n \times n}$ is the symmetric and positive semi-definite. $\mathbf{D}$ is the dignonal matrix, where the elements are defined by $D_{ii} = \sum_j A_{ji}$, the $\mathbf{S} \in \mathbb{R}^{n \times n}$ is the adjacency matrix of the graph.

Assuming that $\mathbf{Y}$ is $\mathbf{Y} = y_i$, where $y_i \in \{0, 1\}^{n \times 1}$ is the cluster indicator vector for x_i. Following the discriminative regularization [46], the scaled cluster indicator matrix $\mathbf{Z}$ as:

$$\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_n}] = \mathbf{Y}(\mathbf{Y^T Y})^{-\frac{1}{2}} \tag{4}$$

where z_i is the scaled cluster indicator of x_j. It can be derived to

$$\mathbf{Z^T Z} = (\mathbf{Y^T Y})^{-\frac{1}{2}} \mathbf{Y^T Y} (\mathbf{Y^T Y})^{-\frac{1}{2}} \tag{5}$$

otherwise, since the $\mathbf{L}$ satisifies the following condition [47]:

$$\mathbf{Z^T Z} = \mathbf{I_c} \tag{6}$$

where $\mathbf{I_n} \in \mathbb{R}^{c \times c}$ is the identity matrix. Then we observe that the equation, $\mathbf{Z^T Z} = \mathbf{I_n}$, where the columns are orthonormal to each other. This standard form now can be generalized to other kinds of spectral clustering.

3.2 JSCGFS

3.2.1 Clustering with Linear Structure

We first introduce the clustering method based on the manifold features. Several methods aim to seek a locally linear representation in Euclidean space. For example, a random projection matrix is widely used on projecting data point x into a random k-dimensional subspace which can be formulated as follows:

$$\left\|\mathbf{X}^{\mathbf{T}}\mathbf{W}-\mathbf{Z}\right\|_F^2 \tag{7}$$

where $\mathbf{X} \in \mathbb{R}^{d\times n}$ is n dimensional data points and $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the projection matrix which is used to evaluate the correlation between data points x_i and manifold features z_i. When w_j, where w_j denotes the j-th row of $\mathbf{W}$, which shrinks to zero meaning that the j-th feature is less correlated to the new representation of z_i. Moreover, some characteristics of data points will be discarded when projection matrix $\mathbf{W}$ projects manifold features $\mathbf{Z}$ on the low dimensional space with other regularization on themselves, such as the L_1-norm [48] or the L_2-norm [31]. This kind of clustering model is sensitive to the noise and outliers since they only consider the linear representation [49]. Obviously, this kind of model is not able to select more valuable features without considering the intrinsic data structures [50].

3.2.2 Spectral Clustering Based on the Optimal Graph

Different from the locally linear representation (LLR) [51] method, another well-known relevant approach is to build the graph with Laplacian matrix. This method demonstrated the ability to detect the cluster structure of data from non-linear manifolds and has attracted lots of attention from researchers [46,52,53]. Optimizing the graph matrix using Laplacian matrix can be defined as following,

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{W},\mathbf{b}}\ & tr(\mathbf{Z}^{\mathbf{T}}\mathbf{L}\mathbf{Z})+\alpha\left\|\mathbf{X}^{\mathbf{T}}\mathbf{W}-\mathbf{Z}\right\|_F^2+\lambda f(\mathbf{W})\\ s.t.\ & \mathbf{Z}^{\mathbf{T}}\mathbf{Z}=\mathbf{I_c}\end{aligned} \tag{8}$$

where $f(W)$ is the regularization term, and the orthogonal constraint, $\mathbf{Z}^{\mathbf{T}}\mathbf{Z}$,is adopted to avoid trivial solutions. The optimal solution is obtained from the eigenvectors of $\mathbf{L}$ corresponding to the smallest eigenvalues.

Generally, the regularization terms $f(W)$ is added into an objective function to keep the sparsity or fit the data. Researchers employ the regularization terms, such as the L_2-norm [31] or the L_1-norm [30], and these norms are widely used as regularization term of a cost function in machine learning. However, the L_2-norm in the subspace learning suffers from the effect of outliers [54] which may increase training error. Hence, the model has to be adjusted appropriately to minimize the errors for outliers than a model using the L_1-norm.

3.2.3 The Objective Function

For the purpose of finding a new robust representation that can capture the characteristics of locality and sparsity, we introduce the regularization term $L_{2,1}$-norm to the clustering

Algorithm 1 JSCGFS(Joint Spectral Clustering based on Optimal Graph and Feature Selection)

Input: Data set $\mathbf{X} \in R^{d\times n}$
Parameter: α and β
Output: Clustering result $\mathbf{C}$

1: Build similarity graph and set up the n nearest neighbor
2: Construct the unnormalized Laplacian matrix $\mathbf{L} \in R^{d\times d}$
3: Initialize $\mathbf{W}$,$\mathbf{b}$,$\mathbf{Z}$
4: **while** iteration **do**
5: Update $\mathbf{b}$ via Eq.(12);
6: Update $\mathbf{W}$ via Eq.(19);
7: Update $\mathbf{Z}$ with GPI [55];
8: **end while**
9: **return** learned feature $\mathbf{Z}$=$\{z_1,\ldots,z_k\}$
10: Cluster the data points $\mathbf{Z}$ with k-means algorithm
Output: A partition of the data points into k disjoint clusters.

model based on the graph, which can be formulated as follows:

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{W},\mathbf{b}} & Tr(\mathbf{Z^T L Z}) + \alpha \left\|\mathbf{X^T W} - \mathbf{Z}\right\|_{2,1} + \beta \left\|\mathbf{W}\right\|_{2,1} \\ & s.t.\ \mathbf{Z^T Z} = \mathbf{I_c}\end{aligned} \tag{9}$$

where $\mathbf{I_c}$ denotes the identity matrix of size c. The first term learns the pseudo class labels with an orthogonal basis while the second term and third term try to learn a latent space with the $L_{2,1}$ norm through the regression model. It is worth noting that the second term penalizes all regression coefficients corresponding to a single feature as a whole. Therefore, we obtain a higher quality intrinsic space for manifold feature $\mathbf{Z}$. The $L_{2,1}$-norm has the effect on the importance of the sample as the k eigenvectors are selected as the columns of $\mathbf{Z}$. Finally, we add the bias term $\mathbf{e^T b}$ into the objective function to obtain the final function as follows:

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{W},\mathbf{b}} & Tr(\mathbf{Z^T L Z}) + \alpha \left\|\mathbf{X^T W} + \mathbf{e^T b} - \mathbf{Z}\right\|_{2,1} \\ & + \beta \left\|\mathbf{W}\right\|_{2,1} \quad s.t.\ \mathbf{Z^T Z} = \mathbf{I_c}\end{aligned} \tag{10}$$

where $\mathbf{e} \in \mathbb{R}^{1\times d}$ and $\mathbf{b} \in \mathbb{R}^{1\times n}$, α and β are hyper-parameters. This objective function considers the possible correlations among all data points and all the features while we employ the $L_{2,1}$-norm as the regularization term, not use the L_1-norm because the L_1-norm is easily to derivative [56] and the solution L_2-norm is generally non-sparse [57]. To directly solve the $L_{2,1}$-norm problem is difficult, thus we develop the optimization method described in the next section. After obtaining a manifold feature $\mathbf{Z}$, traditional clustering methods, such as k-means, are implemented to obtain discrete cluster labels. The Fig. 1 demonstrates the difference between these methods and our algorithm. Previous clustering methods often suffer from the issue of out-of-sample where these proposed methods only partition all data points into clusters but they do not output a model for predicting unseen data points. On the contrary, our proposed method in Eq. (10) outputs the prediction model, i.e., the second term in Eq. (10) to solve the issue of out-of-sample.

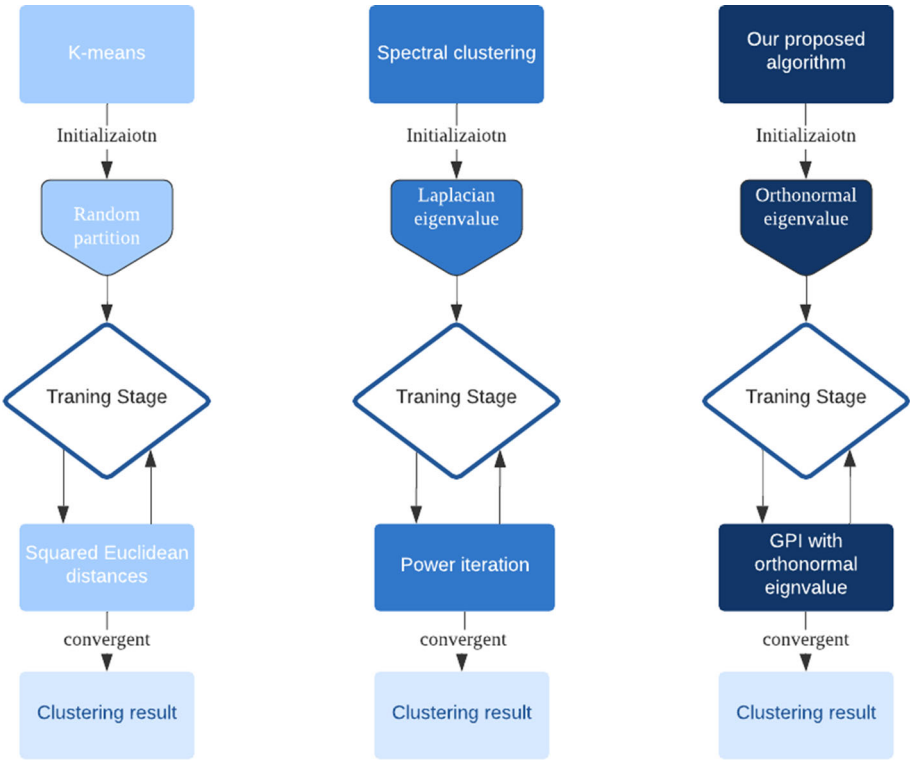

Fig. 1 The structure chart of K-means, Spectral clustering and our algorithm

3.3 Optimization

Since the proposed objection function is not smooth but convex, we can optimize our model to calculate the local minimal solution of the objective function.

3.3.1 Update b with Fixed W, Z

When fixing variables $\mathbf{W}$, $\mathbf{Z}$, which can be considered as constants, the rewritten equation is equivalent to:

$$\min_{\mathbf{b}} \left\| \mathbf{X}^{\mathbf{T}}\mathbf{W} + \mathbf{e}^{\mathbf{T}}\mathbf{b} - \mathbf{Z} \right\|_{\mathbf{2,1}} \tag{11}$$

where $\mathbf{b}$ can be solved iteratively [26] via the following problem:

$$\min_{\mathbf{b}} \; tr(\mathbf{W}^{\mathbf{T}}\mathbf{X} + \mathbf{b}^{\mathbf{T}}\mathbf{e} - \mathbf{Z}^{\mathbf{T}})\mathbf{G}(\mathbf{X}^{\mathbf{T}}\mathbf{W} + \mathbf{e}^{\mathbf{T}}\mathbf{b} - \mathbf{Z}) \tag{12}$$

where $\mathbf{G}$ is a diagonal matrix, and the elements are denoted as

$$G_{ii} = \frac{1}{2\left\| (\mathbf{X}^{\mathbf{T}}\mathbf{W} + \mathbf{e}^{\mathbf{T}}\mathbf{b} - \mathbf{Z})^{i} \right\|_2}$$

and then we make the differential with respect to **b** and have the result:

$$\mathbf{b} = \frac{\mathbf{1}}{\mathbf{n}}((\mathbf{eGe^T})^{-1}(\mathbf{eG}(\mathbf{Z} - \mathbf{X^T W}))) \tag{13}$$

We then substituting **b** into the Eq. (9), and obtain the new equation given by $\mathbf{Z^T Z} = \mathbf{I_c}$, as follows:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{W}} Tr(\mathbf{Z^T LY}) + \beta \left\|\mathbf{W}\right\|_{2,1} \\ &+ \alpha \left\|\mathbf{X^T W} + \mathbf{e^T}(\mathbf{eGe^T})^{-1}\mathbf{eG}(\mathbf{Z} - \mathbf{X^T W}) - \mathbf{Z}\right\|_{2,1} \end{aligned} \tag{14}$$

To simplify (14), we assume that

$$\mathbf{H} = \mathbf{I_c} - (\mathbf{eGe})^{-1}\mathbf{e^T eG} \tag{15}$$

and $(\mathbf{eGe})^{-1}$ is constant, to obtain the following equation:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{W}} Tr(\mathbf{Z^T LZ}) + \alpha \left\|\mathbf{HX^T W} - \mathbf{HY})\right\|_{2,1} + \\ &\beta \left\|\mathbf{W}\right\|_{2,1} \quad s.t. \mathbf{Z^T Z} = \mathbf{I_c} \end{aligned} \tag{16}$$

We further convert the Eq. (9) into the following form [26],

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{W}} Tr(\mathbf{Z^T LZ}) + \alpha \left\|\mathbf{HX^T W} - \mathbf{HZ}\right\|_{2,1} \\ &+\beta tr(\mathbf{W^T DW}) \quad s.t. \ \mathbf{Z^T Z} = \mathbf{I_c} \end{aligned} \tag{17}$$

where **D** is a diagonal matrix and ith element [26] denoted as $D_{ii} = \frac{1}{2\|\mathbf{W}^i\|_2}$.

3.3.2 Update W with fixed Z and b

Fixing **Z**, **b**, which can be regarded as the constant, we have the following problem:

$$\min_{\mathbf{W}} \alpha \left\|\mathbf{HX^T W} - \mathbf{HY}\right\|_{2,1} + \beta(\mathbf{W^T DW}) \tag{18}$$

To solve the problem, we transform it into the following problem:

$$\begin{aligned} &\min_{\mathbf{W}} Tr[\alpha(\mathbf{W^T XH^T} - \mathbf{H^T Z^T})\mathbf{Q}(\mathbf{HX^T W} - \mathbf{HZ}) \\ &+ \beta(\mathbf{W^T DW})] \end{aligned} \tag{19}$$

After taking the derivative with respect to W and then setting the corresponding result as zero, we have:

$$\mathbf{W} = (\mathbf{XH^T QHX^T} + \frac{\beta}{\alpha}\mathbf{D})^{-1}\mathbf{XH^T QZ} \tag{20}$$

where **Q** is diagonal matrix, and elements are denoted as $Q_{ii} = \frac{1}{2\|\mathbf{HX^T W - HZ})^i\|_2}$.

3.3.3 Update Z with Fixed W and b

With the other coefficients updated, and we have the following equation with respect to Z:

$$\begin{aligned} &\min_{\mathbf{Z}} Tr(\mathbf{Z}^{\mathbf{T}}\mathbf{L}\mathbf{Z}) + \alpha \left\| \mathbf{H}\mathbf{X}^{\mathbf{T}}\mathbf{W} - \mathbf{H}\mathbf{Z} \right\|_{2,1} \\ &\quad s.t.\ \ \mathbf{Z}^{\mathbf{T}}\mathbf{Z} = \mathbf{I_c} \end{aligned} \tag{21}$$

which is converted into the following problem:

$$\begin{aligned} &\min_{\mathbf{Y}} Tr[\alpha(\mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{H}^{\mathbf{T}} - \mathbf{H}^{\mathbf{T}}\mathbf{Z}^{\mathbf{T}})\mathbf{Q}(\mathbf{H}\mathbf{X}^{\mathbf{T}}\mathbf{W} - \mathbf{H}\mathbf{Z}) \\ &\quad + \beta(\mathbf{Z}^{\mathbf{T}}\mathbf{L}\mathbf{Z})] \ \ s.t.\ \ \mathbf{Z}^{\mathbf{T}}\mathbf{Z} = \mathbf{I_c} \end{aligned} \tag{22}$$

By solving the quadratic problem on the Stiefel manifold (QPSM) [55] applied to solve the formula and we can obtain:

$$\begin{aligned} &\min_{\mathbf{Z}} Tr[(\mathbf{Z}^{\mathbf{T}}\mathbf{L}\mathbf{Z}) - 2\alpha\mathbf{Z}^{\mathbf{T}}(\mathbf{H}^{\mathbf{T}}\mathbf{Q}\mathbf{H}\mathbf{X}^{\mathbf{T}}\mathbf{W} - \mathbf{H}^{\mathbf{T}}\mathbf{Q}\mathbf{H}\mathbf{Z})] \\ &\quad s.t.\ \ \mathbf{Z}^{\mathbf{T}}\mathbf{Z} = \mathbf{I_c} \end{aligned} \tag{23}$$

This quadratic problem on the manifold can be solved by QPSM until it achieves convergence.

3.3.4 Convergence Analysis

In this section, we use Theorem 1 to prove the convergence of our proposed method.

Theorem 1 *By denoting* $\mathbf{b^{(t)}}$, $\mathbf{W}^{(t)}$, *and* $\mathbf{Z}^{(t)}$, *respectively, as the* t*-th iterations of* $\mathbf{b}$,$\mathbf{W}$ *and* $\mathbf{Z}$, *we denote the objective function value of Eq.*(10) *in the* t*-th iteration as* $\mathbf{L}(\mathbf{b}^{(t)}, \mathbf{W}^{(t)}$, *and* $\mathbf{Z}^{(t)})$. *According to Eq.*(12) *in Sect.* 3.3.1, $b^{(t)}$ *has a closed-form solution, thus we have the following inequality:*

$$\mathbf{L}(\mathbf{W}^{(t)}, \mathbf{Z}^{(t)}, \mathbf{b}^{(t+1)}) \leq \mathbf{L}(\mathbf{W}^{(t)}, \mathbf{Z}^{(t)}, \mathbf{b}^{(t)}) \tag{24}$$

According to Eq.(19) *in Sect.* 3.3.2, $W^{(t)}$ *has a closed-form solution, thus we have the following inequality*

$$\mathbf{L}(\mathbf{W}^{(t+1)}, \mathbf{Z}^{(t)}, \mathbf{b}^{(t+1)}) \leq \mathbf{L}(\mathbf{W}^{(t)}, \mathbf{Z}^{(t)}, \mathbf{b}^{(t+1)}) \tag{25}$$

According to Eq.(23) *in Sect.* 3.3.3 , $Z^{(t)}$ *has a closed-form solution, thus we have the following inequality:*

$$\mathbf{L}(\mathbf{W}^{(t+1)}, \mathbf{Z}^{(t+1)}, \mathbf{b}^{(t+1)}) \leq \mathbf{L}(\mathbf{W}^{(t+1)}, \mathbf{Z}^{(t)}, \mathbf{b}^{(t+1)}) \tag{26}$$

Finally, we have the inequality by integrating the above three inequalities,

$$\mathbf{L}(\mathbf{W}^{(t+1)}, \mathbf{Z}^{(t+1)}, \mathbf{b}^{(t+1)}) \leq \mathbf{L}(\mathbf{W}^{(t)}, \mathbf{Z}^{(t)}, \mathbf{b}^{(t)}) \tag{27}$$

Hence, the objective function value in Eq.(10) *is monotonously decreased via iterating* ***Algorithm 1***.

Table 3 The details of the used data sets

Data sets	Samples	Dimensions	Classes	Data sets	Samples	Dimensions	Classes
Bina	1404	320	36	Crx	690	15	2
Card	2126	41	3	Diab	1151	19	2
Wine	1599	118	6	Park	1040	2	2
Uspst	2007	256	10	Segment	2310	19	7
Wf-21	2746	54	3	Wf	5000	21	3
Bank	1372	4	2	Yeast	1484	8	10

Table 4 The ACC performance of all algorithms on twelve benchmark datasets

ACC					
Data sets	JSCGFS	K-means	SSC	LRR	ULGE
Bina	**0.48**	0.41	0.21	0.41	0.42
Card	0.64	0.45	0.46	**0.72**	0.50
Wine	**0.38**	0.27	0.31	0.30	0.29
Uspst	**0.78**	0.64	0.57	0.54	0.68
Wf-21	**0.62**	0.51	0.50	0.47	0.52
Bank	0.68	0.61	0.72	0.65	**0.89**
Crx	**0.60**	0.55	0.53	0.56	0.57
Diab	**0.55**	0.51	0.54	0.51	0.52
Park	**0.57**	0.52	0.53	0.51	0.54
Segment	0.62	0.51	**0.71**	0.56	0.55
Wf	**0.54**	0.50	0.51	0.49	0.51
Yeast	0.34	**0.36**	0.31	0.31	0.36

4 Experiment

4.1 Data Sets

We used twelve data sets in our experiments from the UCI website. [1] We summarized them in Table 2. We operate all algorithm 50 times and averaged the results. The dimension of feature s is from 8 to 320, and the number of samples varies from 690 to 5000. The number of features of all data sets is less than the number of samples (Table 3).

4.2 Comparison Methods

We employ four comparison methods and the details were given as follows:

K-means [9] selects the centroids randomly, which are treated as the initialization point for each cluster, and calculate the positions of the centroids iteratively according to the means of data points.

SSC [58] classifies data points that scatter in low-dimensional subspace so that the original data points obtain sparse representation in subspace that can deal with data noise and outliers.

[1] http://www.escience.cn/people/chenxiaojun/index.html.

Table 5 The NMI performance of all algorithms on twelve benchmark datasets

NMI					
Data sets	JSCGFS	K-means	SSC	LRR	ULGE
Bina	**0.63**	0.575	0.351	0.581	0.572
Card	**0.045**	0.026	0.044	0.144	0.022
Wine	0.044	0.039	0.034	**0.085**	0.038
Uspst	**0.793**	0.610	0.672	0.675	0.730
Wf-21	0.330	0.363	0.359	0.186	**0.370**
Bank	0.253	0.030	0.151	0.076	**0.573**
Crx	**0.043**	0.005	0.009	0.002	0.042
Diab	0.008	0.003	0.004	**0.021**	0.003
Park	**0.092**	0.004	0.043	0.004	0.013
Segment	0.330	0.363	**0.621**	0.507	0.567
Wf	0.097	0.362	0.363	0.260	**0.370**
Yeast	**0.263**	0.234	0.211	0.107	0.263

Table 6 The Purity performance of all algorithms on twelve benchmark datasets

Purity					
Data sets	JSCGFS	K-means	SSC	LRR	ULGE
Bina	**0.510**	0.442	0.231	0.460	0.446
Card	0.778	0.780	0.779	**0.810**	0.779
Wine	0.456	0.483	0.482	**0.530**	0.486
Uspst	**0.840**	0.711	0.705	0.708	0.768
Wf-21	**0.622**	0.534	0.516	0.487	0.518
Bank	0.685	0.612	0.723	0.656	**0.887**
Crx	**0.598**	0.556	0.555	0.558	0.585
Diab	**0.554**	0.531	0.540	0.531	0.531
Park	0.566	**0.711**	0.532	0.510	0.544
Segment	0.622	0.516	**0.711**	0.604	0.565
Wf	**0.540**	0.531	0.506	0.487	0.508
Yeast	**0.547**	0.514	0.455	0.378	0.524

LRR [59] learns a combination of various subspace through finding the lowest rank of candidate features which can be treated as the new representation of linear combinations. It also groups the similar samples into corresponding basket and may separate the outliers as soon as possible.

ULGE [60] constructs similarity matrix and conducts the analysis for the spectral clustering so that it has an obvious improvement on speed.

4.3 Evaluation Measures

To compare our proposed JSCGFS method with other similar methods, we employed three evaluation metrics that include accuracy (ACC), normalized mutual information (NMI), and Purity respectively. ACC measures how many samples are correctly allocated to the corre-

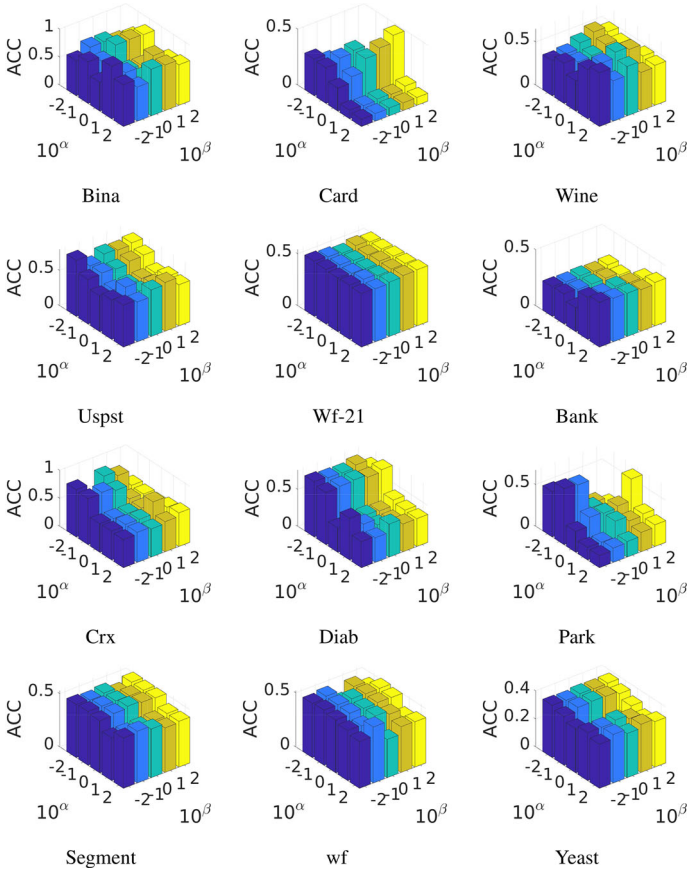

Fig. 2 Accuracy result of all methods on all data sets at different number of selected features

sponding class. NMI is for measuring the similarity between the ground label and predicted label and scales the results between 0 and 1. Purity is used for measuring the rate of samples which are correctly allocated into the corresponding cluster. The definitions of the ACC evaluation metrics are as follows:

$$ACC = N_{correct}/N \tag{28}$$

where $N_{correct}$ denotes how many samples are correctly allocated into the corresponding cluster,and N denotes that the total sample. The definition of NMI is:

$$NMI(A,B) = \frac{\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} log(n_{ij}n/n_i^A n_j^B)}{\sqrt{\sum_{i=1}^{C_A} n_i^A log(n_i^A/n)\sum_{j=1}^{C_B} n_j^B log(n_j^B/n)}} \tag{29}$$

where A,B represents two partitions of n samples into C_A and C_B clusters respectively.

$$Purity = \sum_{i=1}^{k}(S_i/n)P_i \tag{30}$$

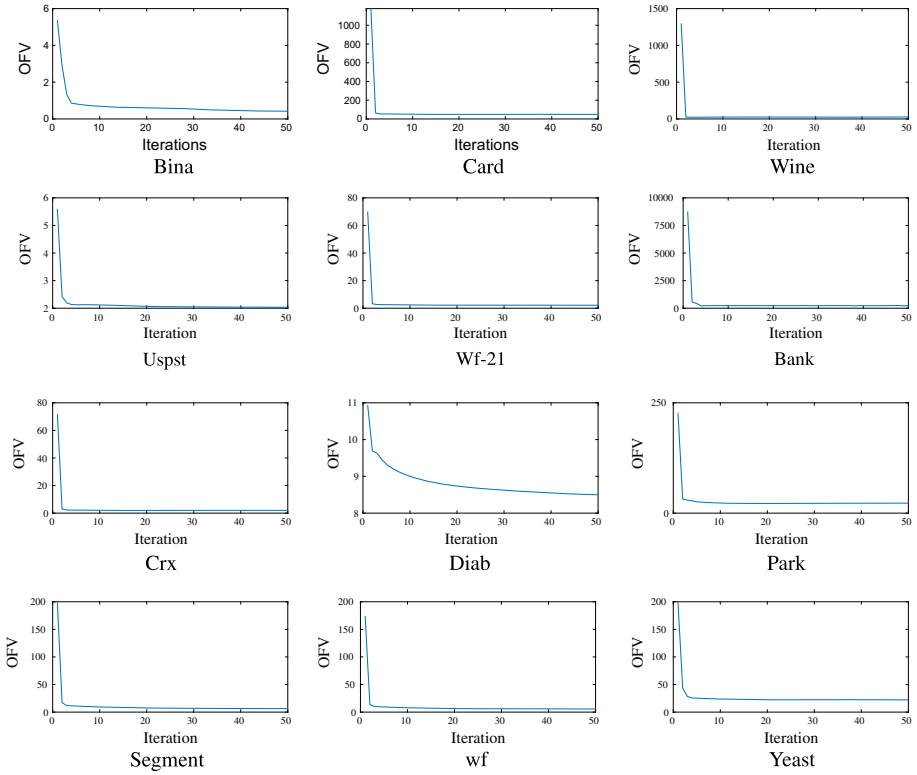

Fig. 3 Objective function value of our proposed method at different number of ranks

where k denotes that the number of clusters and n represents total samples. S_i is the number of samples in the ith cluster. P_i is denoted that the distribution of samples which are allocated correctly.

5 Parameter Sensitivity

From Fig. 2 we see the performance of our method depends on the setting of parameters α and β. In the dataset Wine, our method is sensitive to α and β in this range like the situation in the dataset Uspst. But in the Bank, our method is insensitive to parameter α in the range $[0, 2]$, and it is insensitive to parameter α in the range $[-2, -1]$ in Diab. Our method is neither sensitive to α nor to the β in Wf-21. Therefore, we find our method is partially sensitive to the setting of parameters α and β. This also reveals that our algorithm has efficiently deals with features in different data sets.

6 Convergence

To optimize our proposed objective function Eq. (10) and theoretically prove its convergence, we reported the results on 12 datasets in Tables 4,5 and 6, while setting the stop criteria

of our algorithm as $\frac{\|obj(t+1)-obj(t)\|_2^2}{obj(t)} \leq 10^{-5}$, where $obj(t)$ represents the t-th iteration objective function value of Eq. (10). Figure 3 shows that the trend of objective function values monotonously decreased with respect to the number of iterations and it is obvious that our proposed objective function in Eq. (9) becomes convergent in the first 50 iterations.

7 Conclusion

This paper presents a new spectral feature selection method for clustering analysis. We embed the graph into the subspace and consider joint graph embedding subspace into the united framework. Experimental results demonstrated the proposed method over other methods on three evaluation metrics. Moreover, through the experiment analysis, we recognized that our model could be sensitive to particular characteristics of samples from different application areas. Therefore in future work, we will explore further tuning of our model by taking into considerations of different application settings such as industrial applications and business intelligence.

References

1. Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3(1):1–130
2. Berkhin P (2006) A survey of clustering data mining techniques. In: Kogan J (ed) Grouping multidimensional data. Springer, Berlin, pp 25–71
3. Bodea CN, Dascalu MI, Lipai A (2012) Clustering of the web search results in educational recommender systems. In: Olga C (ed) Educational recommender systems and technologies: practices and challenges. IGI Global, Pennsylvania, pp 154–181
4. Fabrizio C et al (2018) 4.2 Paper V: application of data clustering to railway delay pattern recognition. In: Analytical, big data, and simulation models of railway delays, pp 121
5. Li H, He X, Tao D, Tang Y, Wang R (2018) Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. Pattern Recognit 79:130–146
6. Zhu X, Zhang S, Li Y, Zhang J, Yang L, Fang Y (2018) Low-rank sparse subspace for spectral clustering. IEEE Trans Knowl Data Eng 31:1532–1543
7. Zhu Y, Zhong Z, Cao W, Cheng D (2016) Graph feature selection for dementia diagnosis. Neurocomputing 195:19–22
8. Li X, Li X, Ma H (2020) Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery. Mech Syst Sig Process 143:106825
9. Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recognit Lett 31(8):651–666
10. Chan PK, Schlag MDF, Zien JY (1994) Spectral k-way ratio-cut partitioning and clustering. IEEE Trans Comp-Aided Des Integr Circuits Syst 13(9):1088–1096
11. Li Z, Chen J (2015) Superpixel segmentation using linear spectral clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1356–1363
12. Yan Y, Liu G, Wang S, Zhang J, Zheng K (2017) Graph-based clustering and ranking for diversified image search. Multimed Syst 23(1):41–52
13. Bunke H, Riesen K (2011) Improving vector space embedding of graphs through feature selection algorithms. Pattern Recognit 44(9):1928–1940
14. Peng X, Yu Z, Yi Z, Tang H (2017) Constructing the l2-graph for robust subspace learning and subspace clustering. IEEE Trans Cybern 47(4):1053–1066
15. He W, Zhu X, Cheng D, Hu R, Zhang S (2017) Low-rank unsupervised graph feature selection via feature self-representation. Multimed Tools Appl 76(9):12149–12164
16. Zhao Z, He X, Cai D, Zhang L, Ng W, Zhuang Y (2015) Graph regularized feature selection with data reconstruction. IEEE Trans Knowl Data Eng 28(3):689–700
17. Wang S, Zhu W (2018) Sparse graph embedding unsupervised feature selection. IEEE Trans Syst Man Cybern Syst 48(3):329–341

18. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182
19. Inoue A, Kilian L (2005) In-sample or out-of-sample tests of predictability: Which one should we use? Econom Rev 23(4):371–402
20. Zhu P, Zuo W, Zhang L, Hu Q, Shiu SCK (2015) Unsupervised feature selection by regularized self-representation. Pattern Recognit 48(2):438–446
21. Vural E, Guillemot C (2016) Out-of-sample generalizations for supervised manifold learning for classification. IEEE Trans Image Process 25(3):1410–1424
22. Zhuang L, Gao H, Lin Z, Ma Y, Zhang X, Yu N (2012) Non-negative low rank and sparse graph for semi-supervised learning. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 2328–2335. IEEE
23. Lu X, Wang Y, Yuan Y (2013) Graph-regularized low-rank representation for destriping of hyperspectral images. IEEE Trans Geosci Remote Sens 51(7):4009–4018
24. Li W, Liu J, Du Q (2016) Sparse and low-rank graph for discriminant analysis of hyperspectral imagery. IEEE Trans Geosci Remote Sens 54(7):4094–4105
25. Kuang D, Yun S, Park H (2015) Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering. J Glob Optim 62(3):545–574
26. Nie F, Huang H, Cai X, Ding CH (2010) Efficient and robust feature selection via joint l2, 1-norms minimization. In: Avances in neural information processing systems, pp 1813–1821
27. West DB et al (1996) Introduction to graph theory, vol 2. Prentice hall, Upper Saddle River, NJ
28. Hogstedt K, Kimelman D, Rajan VT, Roth T, Wegman M (2001) Graph cutting algorithms for distributed applications partitioning. ACM SIGMETRICS Perform Evaluat Rev 28(4):27–29
29. Nie F, Wang X, Jordan MI, Huang H (2016) The constrained laplacian rank algorithm for graph-based clustering. In: hirtieth AAAI Conference on Artificial Intelligence
30. Nie F, Wang H, Deng C, Gao X, Li X, Huang H (2016) New l1-norm relaxations and optimizations for graph clustering. In: Thirtieth AAAI Conference on Artificial Intelligence
31. Peng X, Yu Z, Yi Z, Tang H (2016) Constructing the l2-graph for robust subspace learning and subspace clustering. IEEE Trans Cybern 47(4):1053–1066
32. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562
33. Yin M, Gao J, Lin Z (2015) Laplacian regularized low-rank representation and its applications. IEEE Trans Pattern Anal Mach Intell 38(3):504–517
34. Fang X, Xu Y, Li X, Lai Z, Wong WK (2015) Learning a nonnegative sparse graph for linear regression. IEEE Trans Image Process 24(9):2760–2771
35. Zhu X, Li X, Zhang S, Xu Z, Yu L, Wang C (2017) Graph pca hashing for similarity search. IEEE Trans Multimed 19(9):2033–2044
36. Shahid N, Perraudin N, Kalofolias V, Puy G, Vandergheynst P (2016) Fast robust pca on graphs. IEEE J Sel Top Sig Process 10(4):740–756
37. Feng CM, Gao YL, Liu JX, Zheng CH, Yu J (2017) Pca based on graph laplacian regularization and p-norm for gene selection and clustering. IEEE Trans Nanobiosci 16(4):257–265
38. Chen F, Wang B, Kuo CCJ (2019) Deepwalk-assisted graph pca (dgpca) for language networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2957–2961. IEEE
39. Montanari A (2015) Finding one community in a sparse graph. J Statist Phys 161(2):273–299
40. Pedarsani R, Yin D, Lee K, Ramchandran K (2017) Phasecode: fast and efficient compressive phase retrieval based on sparse-graph codes. IEEE Trans Inf Theory 63(6):3663–3691
41. Wang S, Zhu W (2016) Sparse graph embedding unsupervised feature selection. IEEE Trans Syst Man Cybern Syst 48(3):329–341
42. Xue Z, Du P, Li J, Su H (2015) Simultaneous sparse graph embedding for hyperspectral image classification. IEEE Trans Geosci Remote Sens 53(11):6114–6133
43. Li X, Cui G, Dong Y (2017) Graph regularized non-negative low-rank matrix factorization for image clustering. IEEE Trans Cybern 47(11):3840–3853
44. Zhuang L, Gao S, Tang J, Wang J, Lin Z, Ma Y, Yu N (2015) Constructing a nonnegative low-rank and sparse graph with data-adaptive features. IEEE Trans Image Process 24(11):3717–3728
45. Li S, Fu Y (2015) Learning balanced and unbalanced graphs via low-rank coding. IEEE Trans Knowl Data Eng 27(5):1274–1287
46. Yang Y, Shen HT, Nie F, Ji R, Zhou X (2011) Nonnegative spectral clustering with discriminative regularization. In: Twenty-Fifth AAAI Conference on Artificial Intelligence
47. Von Luxburg U (2007) A tutorial on spectral clustering. Statist Comput 17(4):395–416

48. Soltanolkotabi M, Elhamifar E, Candes EJ et al (2014) Robust subspace clustering. Ann Statist 42(2):669–699
49. Vidal R (2011) Subspace clustering. IEEE Sig Process Mag 28(2):52–68
50. Yang Y, Ma Z, Yang Y, Nie F, Shen HT (2014) Multitask spectral clustering by exploring intertask correlation. IEEE Trans Cybern 45(5):1083–1094
51. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
52. Kang Z, Peng C, Cheng Q, Xu Z (2018) Unified spectral clustering with optimal graph. In: Thirty-Second AAAI Conference on Artificial Intelligence
53. Li Z, Yang Y, Liu J, Zhou X, Lu H (2012) Unsupervised feature selection using nonnegative spectral analysis. In: Twenty-Sixth AAAI Conference on Artificial Intelligence
54. Pang Y, Yuan Y (2010) Outlier-resisting graph embedding. Neurocomputing 73(4–6):968–974
55. Nie F, Zhang R, Li X (2017) A generalized power iteration method for solving quadratic problem on the stiefel manifold. Sci China Inf Sci 60(11):112101
56. Dodge Y (2012) Statistical data analysis based on the L1-norm and related methods. Birkhäuser, Basel
57. Kloft M, Brefeld U, Laskov P, Sonnenburg S (2008) Non-sparse multiple kernel learning. In: NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels
58. Elhamifar E, Vidal R (2013) Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 35(11):2765–2781
59. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y (2012) Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell 35(1):171–184
60. Nie F, Zhu W, Li X (2017) Unsupervised large graph embedding. In: Thirty-first AAAI conference on artificial intelligence

Joint Spectral Clustering based on Optimal Graph and Feature Selection

Zhu J

2021-02