Unsupervised Feature Selection via Adaptive Graph Learning and Constraint

Rui Zhang, *Member, IEEE*, Yunxing Zhang, and Xuelong Li, *Fellow, IEEE*

***Abstract*— The performance of graph-based feature selection methods relies heavily on the quality of the construction of the similarity matrix. However, most of the graphs on these methods are initially fixed, where few of them are constrained. Once the graph is determined, it will remain constant in the whole optimization process. In other words, in case that the graph constructed on the raw data is not appropriate, it will drag down the entire algorithm. Aiming to tackle this defect, a novel unsupervised feature selection via adaptive graph learning and constraint (EGCFS) is proposed to select the uncorrelated yet discriminative features by exploiting the embedded graph learning and constraint. The adaptive graph learning method incorporates the structure of the similarity matrix into the optimization process, which not only learns the graph structure adaptively but also obtains the closed-form solution of the graph coefficient. Special graph constraint is embedded with the feature selection process to connect nearer data points with larger probability. The idea of maximizing between-class scatter matrix and the adaptive graph structure is integrated into a uniform framework to obtain excellent structural performance. Moreover, the proposed embedded graph constraint not only performs with manifold structure but also validates the link between graph-based approach and k-means from a unique perspective. Experiments on several benchmark data sets verify the effectiveness and superiority of the proposed method.**

***Index Terms*— Adaptive graph learning, graph constraint, sparsity, unsupervised feature selection.**

I. Introduction

With the development of technology, large amounts of high-dimensional data have become a big problem in many fields, such as computer vision [1], [2], data mining [3]–[5], and pattern recognition [6]. High-dimensional data often contain quite a lot of noise features and redundant information, which is detrimental to data processing. By selecting the most discriminative feature subset from the raw data, feature selection improves the interpretability of the data, which is becoming one of the most important methods to deal with high-dimensional data.

Various methods of feature selection have been proposed in the last decades, and there are three types of feature selection algorithms: filter-based methods [7], wrapper-based methods [8], and embedding-based methods [9]. The filter-based methods score the features with a ranking, and the feature subset is selected based on a well-defined criterion. Wrapper methods are classifier-specific, and the feature subset is selected directly based on the performance of a specific classifier. Embedded methods are often classifier-specific and the classifier training process into one optimization process. In embedded methods, feature search and the learning algorithm are incorporated into a single optimization problem such that a reasonable computational cost can be achieved for good classification performance. Among these methods, the embedded methods are very popular in recent years for the need not to learn a classifier in the process of feature selection.

According to whether or not the availability of labels, feature selection methods can be roughly divided into three categories: supervised methods, semisupervised methods, and unsupervised methods. The supervised feature selection methods determine feature relevance by evaluating the feature's correlation with the class labels. By utilizing the whole label of data, some supervised feature selection methods gain excellent performance. However, the use of whole labeled data makes those methods very expensive and time-consuming. Semisupervised feature selection uses both (small) labeled data and (large) unlabeled data. Unsupervised feature selection exploits the most discriminating features without any class labels. In many real-world applications, the acquisition of data labels is expensive and time-consuming. Benefiting from not requiring label information, the unsupervised feature selection methods are widely used in practical applications [10]–[12].

Among the feature selection methods, graph-based feature selection methods [13], [14] are increasingly popular due to good performance. Conventional graph-based feature selection methods [9], [15], [16] roughly include two steps. First, the data structure is explored by the spectral analysis of graph Laplacian or nonnegative matrix factorization, and so on. Then, the feature selection matrix is learned by virtue of sparsity regularization models. The similarity matrix of these methods is derived from raw data and remains constant for the subsequent process. Nevertheless, real-world data always contain lots of noise samples and features, which makes the similarity matrix unreliable [17]. The unreliable similarity matrix will damage the local manifold structure and ultimately lead to suboptimal results.

The construction and constraint of similarity matrices are critical for graph-based approaches. Though the performance of graph-based feature selection methods relies heavily on the quality of the construction of the similarity matrix, these methods rarely impose constraints on the graph. What is more, the constructed similarity matrices in most of the graph-based methods are based on the raw data and are not embedded in the learning process, which results in that the similarity matrices cannot be effectively updated in the optimization process. To mitigate the impact of the above problems, we propose an unsupervised feature selection via embedded graph learning and constraint method, namely EGCFS. It is worthwhile to highlight the main contributions of this brief as follows.

1) The exquisite adaptive graph learning method embeds the construction of the similarity matrix into the optimization process, which not only makes the graph structure adaptively but also makes the graph coefficient can get the closed-form solution.
2) The special graph constraint method is embedded in the feature selection process to connect nearer data points with greater probability for a clearer graph structure.
3) By applying constraint on the similarity matrix, the idea of maximizing between-class scatter matrix and the adaptive graph structure is integrated into a uniform framework to obtain excellent structural performance.

Manuscript received October 19, 2019; revised February 15, 2020 and August 31, 2020; accepted November 30, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1107400 and in part by the National Natural Science Foundation of China under Grant 61871470, Grant U1801262, and Grant 61761130079. *(Corresponding author: Xuelong Li.)*

The authors are with the School of Computer Science, and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China (e-mail: ruizhang8633@gmail.com; zhangyunxing423@outlook.com; xuelong_li@nwpu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2020.3042330.

Digital Object Identifier 10.1109/TNNLS.2020.3042330

4) The proposed method not only directly embeds the Laplacian graph structure into a low-dimensional subspace to obtain manifold structure but also confirms the link between the graph-based approach and k-means from a unique perspective.

II. Notations

Throughout this brief, $\mathrm{Tr}(\cdot)$ stands for trace operator. Suppose $\mathbf{w}_i$ as the ith row of the matrix $\mathbf{W} = [\mathbf{w}_{ij}] \in \mathbb{R}^{d\times m}$. Then, the Frobenius norm of $\mathbf{W}$ is defined as $\|\mathbf{W}\|_F = (\sum_{i=1}^{d}\sum_{j=1}^{m}\mathbf{w}_{ij}^2)^{1/2} = (\sum_{i=1}^{d}\|\mathbf{w}_i\|_2^2)^{1/2}$, and the $\ell_{2,1}$ norm of $\mathbf{W}$ is defined as $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d}(\sum_{j=1}^{m}\mathbf{w}_{ij}^2)^{1/2} = \sum_{i=1}^{d}\|\mathbf{w}_i\|_2$, where $\|\mathbf{w}_i\|_2$ denotes the ℓ_2-norm of the vector $\mathbf{w}_i$. The $\ell_{2,0}$ norm of $\mathbf{W}$ is defined as $\|\mathbf{W}\|_{2,0} = \sum_{i=1}^{d}\|\sum_{j=1}^{m}\mathbf{w}_{ij}^2\|_0$. $\mathbf{W}^{\perp}$ is defined as the matrix whose columns span orthogonal complement space of the column space of $\mathbf{W}$.

III. Related Work

Unsupervised feature selection has been widely used in practice for its no need for label information. In this section, we review the related work on feature selection and graph learning in recent years.

A. Feature Selection

In general, the purpose of feature selection is to select the most discriminative feature subset from the raw data, thereby removing redundant information in the data and enhancing data processing. Zhang *et al.* [18] have a review of the work related to feature selection. Under the assumption that the class label of input data can be predicted by a linear classifier, Yang *et al.* [19] incorporate discriminative analysis and $\ell_{2,1}$ minimization into a joint framework for unsupervised feature selection. Chen *et al.* [20] propose a novel semisupervised embedded feature selection method that extends the least-squares regression model by rescaling the regression coefficients in the least-squares regression with a set of scale factors. To overly suppress the nonzero rows such that the associated features are insufficient for selection, Zhang *et al.* [21] proposed a self-weighted supervised discriminative feature selection method.

The difference between ours and those methods is that the idea of maximizing between-class scatter matrix and the adaptive graph structure is integrated into a uniform framework to obtain excellent structural performance. Besides that, our method not only directly embeds the Laplacian graph structure into a low-dimensional subspace to obtain manifold structure but also confirms the link between the graph-based approach and k-means from a unique perspective.

B. Graph Learning and Constraint

The construction and constraint of similarity matrices are critical for graph-based approaches. However, most existing graph-based methods are directly composed directly on the raw data matrix, and the similarity matrices are not updated in the optimization. Studying the problem of feature selection in fuzzy-rough sets in the framework of graph theory, Chen *et al.* [22] propose a new mechanism for fuzzy-rough feature selection. Zhang *et al.* [23] proposed an adaptive graph learning unsupervised feature selection method that performs the problem of estimating or learning the data similarity matrix and data regression as simultaneous tasks. To learn/estimate graphs from data, Egilmez *et al.* [24] propose a novel framework that includes the formulation of various graph learning problems, their probabilistic interpretations, and the associated algorithms.

Our method is also a graph method, but there are many differences between ours and those methods. The special graph constraint is embedded in our framework, which makes that the similarity matrix can be adaptively learned. The exquisite adaptive graph learning only makes the graph structure adaptively but also makes that the graph coefficient can get the closed-form solution.

IV. Methodology

In this section, we propose a novel unsupervised feature selection method, EGCFS, that applies constraints on the similarity matrix, and the graph can be adaptively constructed. We not only describe in detail the proposed algorithm but also prove the ingenious connection between the graph-based method and the k-means algorithm. There is a very skillful way to embed similarity matrix construction into the learning process in this brief, which not only can construct the graph adaptively but also avoid adjusting parameters in the graph learning process.

A. Proposed Framework

Given $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ as the centralized data matrix with dimension d and data number n, where $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ is the ith data point. Each data point $\mathbf{x}_i$ is associated with a class label $c_i \in \{1, 2, \ldots, c\}$. Supposed $\mathbf{G} \in \{0, 1\}^{n\times c}$ is the binary index matrix. According to the theory of manifold learning, there always exists a low-dimensional manifold that can express the structure of high-dimensional data. We aim at finding an orthogonal projection matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ that can be linearly combined with the raw data matrix $\mathbf{X}$ to best approximate the low-dimension manifold. By taking advantage of the Laplacian matrix, we embed the low-dimensional manifold data $\mathbf{W}^T\mathbf{X}$ into the laplacian graph to optimize as follows:

$$\min_{\mathbf{W}} \ \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \ \ \text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{1}$$

where $\mathbf{L}$ is the Laplacian matrix that is a basic but important equation in spectral analysis.

The Laplacian matrix can be calculated as $\mathbf{L} = \mathbf{P} - \mathbf{S}$, where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is the similarity matrix and the degree matrix $\mathbf{P}$ is an $n \times n$ diagonal matrix whose ith diagonal element is $\sum_j(s_{ij})$. There is an important property in the Laplacian matrix $\mathbf{L}$ as

$$2\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) = \sum_{i,j=1}^{n}\|f_i - f_j\|_2^2 s_{ij} \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{n\times m}$ and f is the column vector of matrix $\mathbf{F}$.

Based on the supposed binary index matrix $\mathbf{G}$, the idea of maximizing the between-class scatter matrix in the subspace can be expressed as

$$\min_{\mathbf{W},\mathbf{G}} \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}^T\mathbf{W}). \tag{3}$$

In addition, to connect nearer data points with greater probability in the subspace, we fixed the similarity matrix to a symmetric matrix $\mathbf{S} = (\mathbf{S}^T + \mathbf{S})/2$ and impose the constraint on the similarity matrix as $\mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}$, where λ is the constraint parameter. In embedding-based feature selection methods, when $\|\mathbf{w}_i\|_2$ is zero, the ith feature is actually not considered in the training of the model. Therefore, a row-sparse matrix $\mathbf{W}$ could serve as the evaluation criterion for selecting features, where $\|\mathbf{w}_i\|_2$ is the score of the ith feature (deemed as the contribution of each feature). Then, $\|\mathbf{W}\|_{2,0}$ is embedded in the feature selection process as a regularization term. The preliminary model can be expressed as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{G}} \ & \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \\ & - \lambda\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}^T\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,0} \\ \text{s.t. } & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \quad \mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}\end{aligned} \tag{4}$$

where α and λ are the parameters.

As we all know, $\|\mathbf{W}\|_{2,0}$ is difficult to solve directly. However, the constraint of the orthogonal matrix not only makes the samples statistically irrelevant in the manifold structure but also brings convenience to the solution of the model (4). Under the orthogonal

constraint, the solution of $\|\mathbf{W}\|_{2,1}$ is equal to the solution $\|\mathbf{W}\|_{2,0}$ according to [25] and Theorem 1. Thence, we replace $\|\mathbf{W}\|_{2,0}$ in model (4) with $\|\mathbf{W}\|_{2,1}$.

Meanwhile, we make a very bold innovation, embedding the construction process of the graph into the learning process. A novel unsupervised feature selection via adaptive graph learning and constraint method (EGCFS) can be obtained as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{G},\mathbf{S},\gamma}\ & \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) + \gamma\|\mathbf{S}\|_F^2 \\ & - \lambda\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}^T\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \\ \text{s.t. }& \mathbf{W}^T\mathbf{W} = \mathbf{I}, \quad \mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}. \end{aligned} \tag{5}$$

It is particularly worth emphasizing that the graph structure in the model (5) can be adaptively learned. In the conventional graph-based methods, the similarity matrix $\mathbf{S}$ is usually determined initially and remains constant during the optimization process. However, in our method, the construction of the similarity matrix $\mathbf{S}$ is embedded in the optimization process, and the similarity matrix can be adaptively obtained in each iteration for a better subspace graph structure. What is more worth emphasizing is that the coefficient γ is different from the parameters α and λ, and it can also be adaptively obtained during the optimization process. In Section V, we will present an exquisite solution to get the closed-form solution of each variable.

Theorem 1: The solution to problem $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\|\mathbf{W}\|_{2,1}$ is equivalent to the solution to problem $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\|\mathbf{W}\|_{2,0}$ under the orthogonal constraint.

Proof: Suppose the set $\mathcal{A} = \{\mathbf{W} \in \mathbb{R}^{d\times m} | \mathbf{W}^T\mathbf{W} = \mathbf{I}$ with m nonzero rows$\}$. Without loss of generality, we assume $\mathbf{W}^* = \begin{bmatrix}\mathbf{W}_1 \\ \mathbf{0}\end{bmatrix}$ for $\forall\ \mathbf{W}^* \in \mathcal{A}$, where square matrix $\mathbf{W}_1 \in \mathbb{R}^{m\times m}$ is orthonormal and $\mathbf{0} \in \mathbb{R}^{(d-m)\times m}$.

Because $\|\mathbf{W}\|_{2,0} \geq \mathrm{rank}(\mathbf{W}) = m$, so the solution to problem $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\|\mathbf{W}\|_{2,0}$ is an arbitrary orthogonal matrix with m nonzero rows. Otherwise, the rank of $\mathbf{W}$ will be smaller than m, which contradicts the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. In sum, $\mathcal{A}$ is the solution set of problem $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\|\mathbf{W}\|_{2,0}$.

On the other hand, for any orthogonal matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$, we can construct an orthonormal matrix

$$\hat{\mathbf{W}} = [\mathbf{W}, \mathbf{W}^\perp] = \begin{bmatrix}\hat{\mathbf{w}}_1 \\ \vdots \\ \hat{\mathbf{w}}_d\end{bmatrix} \in \mathbb{R}^{d\times d}$$

satisfying $\|\hat{\mathbf{w}}_i\|_2 = 1 \forall i$. Therefore, we have $\|\mathbf{w}_i\|_2 \leq 1$ such that $\|\mathbf{w}_i\|_2^2 \leq \|\mathbf{w}_i\|_2$. In sum, we can infer that

$$\|\mathbf{W}\|_{2,1} \geq \|\mathbf{W}\|_F^2 = m = \|\mathbf{W}_1\|_{2,1} = \|\mathbf{W}^*\|_{2,1}.$$

Therefore, $\mathcal{A}$ is also the solution set of problem $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\|\mathbf{W}\|_{2,1}$. Since these two problems share exactly the same solution set $\mathcal{A}$, the solution of $\|\mathbf{W}\|_{2,1}$ is equal to the solution $\|\mathbf{W}\|_{2,0}$ under the orthogonal constraint. □

B. Relation to k-Means

In model (5), since the symmetric similaity matrix $\mathbf{S}$ satisfies the constraint $\mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}$, the diagonal elements of degree matrix satisfy $p_{ii} = \sum_i \mathbf{s}_{ij} = \lambda$. Then, the Laplacian matrix $\mathbf{L} = \mathbf{P} - \mathbf{S} = \lambda\mathbf{I} - \mathbf{S}$. Problem (5) can be expressed as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{G},\mathbf{S},\gamma}\ & \mathrm{Tr}(\mathbf{W}^T\mathbf{X}(\lambda\mathbf{I} - \mathbf{S})\mathbf{X}^T\mathbf{W}) + \gamma\|\mathbf{S}\|_F^2 \\ & - \lambda\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}^T\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \\ \text{s.t. }& \mathbf{W}^T\mathbf{W} = \mathbf{I}, \quad \mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}. \end{aligned} \tag{6}$$

Problem (6) can be rewritten as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{G},\mathbf{S},\gamma}\ & \lambda\mathrm{Tr}(\mathbf{W}^T\mathbf{X}(\mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)\mathbf{X}^T\mathbf{W}) \\ & - \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{S}\mathbf{X}^T\mathbf{W}) + \gamma\|\mathbf{S}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} \\ \text{s.t. }& \mathbf{W}^T\mathbf{W} = \mathbf{I}, \quad \mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}. \end{aligned} \tag{7}$$

It is obvious that the first item in problem (7) $\Psi = (\mathbf{W}^T\mathbf{X}(\mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)\mathbf{X}^T\mathbf{W})$ is just the objective function of k-means. The reason is that $(\mathbf{X}(\mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)\mathbf{X}^T)$ is the within-class scatter matrix, which confirms that the Ψ is the objective function of k-means. Our approach confirms the link between the graph-based method and k-means from a unique perspective under graph constraint.

V. Optimization Procedure

To facilitate the calculation, we convert the regularization terms $\|\mathbf{W}\|_{2,1}$ in problem (5) into a matrix form $Tr(\mathbf{W}^T\mathbf{D}\mathbf{W})$, where $\mathbf{D}$ is a $d \times d$ diagonal matrix whose diagonal element is $d_{ii} = (1/2(\|\mathbf{w}_i\|_2^2 + \varepsilon)^{1/2})(\varepsilon \to 0, i = 1, 2, \ldots, d)$. Then, problem (5) can be rewritten as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{U},\mathbf{S},\gamma}\ & \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) + \gamma\|\mathbf{S}\|_F^2 \\ & - \lambda\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{U}\mathbf{U}^T\mathbf{X}^T\mathbf{W}) + \alpha Tr(\mathbf{W}^T\mathbf{D}\mathbf{W}) \\ \text{s.t. }& \mathbf{W}^T\mathbf{W} = \mathbf{I}, \quad \mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}, \ \mathbf{U}^T\mathbf{U} = \mathbf{I} \end{aligned} \tag{8}$$

where the indicator matrix $\mathbf{U} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1/2}$ is relaxed from discrete to the orthogonal one, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. There are four variables $\mathbf{W}$, $\mathbf{U}$, $\mathbf{S}$, and γ in the optimization of problem (8). Hereinafter, the coordinate blocking method, i.e., the alternating method, is employed.

A. Solve W by Fixing U, S, and γ

In the case of $\mathbf{U}$, $\mathbf{S}$, and γ are fixed, the solution of the variable $\mathbf{W}$ can be written as

$$\mathbf{W} = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}}{\mathrm{argmin}}\, \mathrm{Tr}\big(\mathbf{W}^T\big(\mathbf{X}(\mathbf{L} - \lambda\mathbf{U}\mathbf{U}^T)\mathbf{X}^T + \alpha\mathbf{D}\big)\mathbf{W}\big). \tag{9}$$

The feature vector corresponding to the first m minimum eigenvalues of $(\mathbf{X}(\mathbf{L} - \lambda\mathbf{U}\mathbf{U}^T)\mathbf{X}^T + \alpha\mathbf{D})$ defined in (9) constitutes the projection matrix $\mathbf{W}$. Since the matrix $\mathbf{D}$ is also dependent on $\mathbf{W}$, it is iteratively updated in the process of optimization $\mathbf{W}$.

B. Solve U by Fixing W, S, and γ

When $\mathbf{W}$, $\mathbf{S}$, and γ are fixed, the optimization of variable $\mathbf{U}$ is simplified to

$$\begin{aligned}\mathbf{U} &= \underset{\mathbf{U}^T\mathbf{U}=\mathbf{I}}{\mathrm{argmin}}\ -\mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{U}\mathbf{U}^T\mathbf{X}^T\mathbf{W}) \\ &= \underset{\mathbf{U}^T\mathbf{U}=\mathbf{I}}{\mathrm{argmax}}\ \mathrm{Tr}(\mathbf{U}^T\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{U}). \end{aligned} \tag{10}$$

The optimal solution $\mathbf{U}$ is formed by the c eigenvectors of $(\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X})$ corresponding to c maximum eigenvalues.

C. Adaptive Learning on Similarity Matrix S

There are many methods to construct the similarity matrix $\mathbf{S}$ of graph, such as the Gaussian kernel, LLE [26], LTSA [27], LSE [28], LRGA [29], and CAN [30]. In this brief, we utilization a very exquisite method to solve the similarity matrix $\mathbf{S}$, which not only makes the learning of the similarity matrix embedded in the overall optimization process but also makes the coefficients γ can be adaptively solved. By embedding the similarity matrix into the learning process, the similarity matrix $\mathbf{S}$ is updated in each iteration, resulting in a better subspace data structure.

When $\mathbf{W}$, $\mathbf{U}$, and γ are fixed, problem (8) can be converted to

$$\begin{aligned} &\min_{\mathbf{S}} \ \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) + \gamma \|\mathbf{S}\|_F^2 \\ &\text{s.t. } \mathbf{S}^T\mathbf{1} = \lambda\mathbf{1}. \end{aligned} \tag{11}$$

In order to ensure that $\mathbf{S}$ is a semipositive definite matrix, we impose the constraints $0 \le \mathbf{s}_i \le 1$ on $\mathbf{S}$. At the same time, problem (11) is written in a vector form

$$\begin{aligned} &\min_{\mathbf{S}} \sum_{i,j=1}^{n} \left(\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2 s_{ij} + \gamma s_{ij}^2\right) \\ &\text{s.t. } \forall i, \quad \mathbf{s}_i^T\mathbf{1} = \lambda, \ 0 \le \mathbf{s}_i \le 1. \end{aligned} \tag{12}$$

Algorithm 1 Algorithm to Solve Problem (8)

Input: The coefficient α and λ, cluster number c, select feature number m, centralized data matrix $\mathbf{X}$.

Initialize Unit diagonal matrix $\mathbf{D} = \mathbf{I}$, random binary index matrix $\mathbf{G} \in \{0, 1\}^{n\times c}$, $\mathbf{U} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1/2}$.

Repeat:

1: Update $\mathbf{W}$ via solving problem (9).
2: Updata $\mathbf{D} = \mathrm{diag}(\frac{1}{2\sqrt{\|\mathbf{w}_1\|_2^2+\varepsilon}}, \frac{1}{2\sqrt{\|\mathbf{w}_2\|_2^2+\varepsilon}}, \ldots, \frac{1}{2\sqrt{\|\mathbf{w}_d\|_2^2+\varepsilon}})$.
3: Update $\mathbf{U}$ via Eq. (10).
4: Update $\mathbf{S}$ via Eq. (21).
5: Dermine the value of γ via Eq. (20)

Until **convergence**

Output: Calculate and sort $\|\mathbf{w}_i\|_2$ $(i = 1, 2, \ldots, d)$ in the descending order, then select the top m ranked features as the results of feature selection.

Denote $d_{ij} = \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\|_2^2$, and note that the problem (12) is independent between different i, so we can transform the problem (12) to a vector form as

$$\min_{\mathbf{s}_i^T\mathbf{1}=\lambda, 0\le\mathbf{s}_i\le 1} \left\|\mathbf{s}_i + \frac{1}{2\gamma}\mathbf{d}_i\right\|_2^2. \tag{13}$$

The Lagrangian function of problem (13) is

$$\mathcal{L}(\mathbf{s}_i, \eta, \beta_i) = \frac{1}{2}\left\|\mathbf{s}_i + \frac{\mathbf{d}_i}{2\gamma}\right\|_2^2 - \eta(\mathbf{s}_i^T\mathbf{1} - \lambda) - \beta_i^T\mathbf{s}_i \tag{14}$$

where η and $\beta_i \ge \mathbf{0}$ are the Lagrangian multipliers. According to the KKT conditions and complementary relaxation condition [31], the optimal solution s_{ij} should be

$$s_{ij} = \left(-\frac{d_{ij}}{2\gamma_i} + \eta\right)_+. \tag{15}$$

D. Determine the Value of γ

In practice, it is very troublesome to adjustment the regularization parameter since its value could be from zero to infinite. In this section, we present an effective method to determine the regularization parameter γ. In practical applications, sparse similarity matrices $\mathbf{S}$ tend to bring better application results. Therefore, only the k sample points closest to x_i are taken into consideration. Without loss of generality, suppose that $d_{i1}, d_{i2}, \ldots, d_{in}$ are ordered from small to large. Because s_i satisfies $s_{ik} > 0 \ge s_{i,k+1}$, we have

$$\begin{cases} s_{ik} > 0 \Rightarrow -\dfrac{d_{ik}}{2\gamma_i} + \eta > 0 \\ s_{i,k+1} \le 0 \Rightarrow -\dfrac{d_{i,k+1}}{2\gamma_i} + \eta \le 0. \end{cases} \tag{16}$$

According to (15) and the constraint $s_i^T\mathbf{1} = \lambda$, we have

$$\begin{aligned} &\sum_{j=1}^{k}\left(-\frac{d_{ij}}{2\gamma_i} + \eta\right) = \lambda \\ &\Rightarrow \eta = \frac{\lambda}{k} + \frac{1}{2k\gamma_i}\sum_{j=1}^{k} d_{ij}. \end{aligned} \tag{17}$$

By substituting the value of η in (17) into (16), we have

$$\frac{k}{2\lambda}d_{ik} - \frac{1}{2\lambda}\sum_{j=1}^{k} d_{ij} < \gamma_i \le \frac{k}{2\lambda}d_{i,k+1} - \frac{1}{2\lambda}\sum_{j=1}^{k} d_{ij}. \tag{18}$$

Therefore, in order to obtain an optimal solution of s_i that has exact k nonzero values, we set γ_i to be

$$\gamma_i = \frac{k}{2\lambda}d_{i,k+1} - \frac{1}{2\lambda}\sum_{j=1}^{k} d_{ij} \tag{19}$$

and then the overall γ is set to the mean of γ_i as

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2\lambda}d_{i,k+1} - \frac{1}{2\lambda}\sum_{j=1}^{k} d_{ij}\right). \tag{20}$$

By taking (19) into (15), we can obtain

$$s_{ij} = \left(\frac{\lambda d_{i,k+1} - \lambda d_{ij}}{k d_{i,k+1} - \sum_{j=1}^{k} d_{ij}}\right)_+. \tag{21}$$

The detail of the EGCFS algorithm is described in Algorithm 1.

E. Convergence of Algorithm 1

Algorithm 1 solves problem (8) with iteratively updating of $\mathbf{W}, \mathbf{D}, \mathbf{U}, \mathbf{S}$, and γ. In order to demonstrate this convergence, Lemma 1 is utilized subsequently, which is proposed and proven in [32].

Lemma 1: For any nonzero vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$, it holds that

$$\|\mathbf{u}\|_2 - \frac{\|\mathbf{u}\|_2^2}{2\|\mathbf{v}\|_2} \le \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}. \tag{22}$$

Theorem 2: Algorithm 1 decreases problem (5) by iteratively updating $\mathbf{W}, \mathbf{D}, \mathbf{U}, \mathbf{S}$, and γ with its optimal solution to problem (8) until convergence.

Proof: Denote the objective value in the tth iteration of problem (8) as $\mathcal{J}(\mathbf{W}^{(t)}, \mathbf{D}^{(t)}, \mathbf{U}^{(t)}, \mathbf{S}^{(t)}, \gamma^{(t)})$, i.e.,

$$\begin{aligned} &\mathcal{J}(\mathbf{W}^{(t)}, \mathbf{D}^{(t)}, \mathbf{U}^{(t)}, \mathbf{S}^{(t)}, \gamma^{(t)}) \\ &= \mathrm{Tr}((\mathbf{W}^{(t)})^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^{(t)}) + \alpha Tr((\mathbf{W}^{(t)})^T\mathbf{D}^{(t)}\mathbf{W}^{(t)}) \\ &\quad - \lambda\mathrm{Tr}((\mathbf{W}^{(t)})^T\mathbf{X}\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\mathbf{X}^T\mathbf{W}^{(t)}) + \gamma^{(t)}\|\mathbf{S}^{(t)}\|_F^2. \end{aligned} \tag{23}$$

Since Algorithm 1 updates $\mathbf{W}, \mathbf{D}, \mathbf{U}, \mathbf{S}$, and γ with the optimal solution to problem (8) in each iteration, for the $(t+1)$th iteration, it must hold

$$\begin{aligned} &\mathcal{J}(\mathbf{W}^{(t+1)}, \mathbf{D}^{(t+1)}, \mathbf{U}^{(t+1)}, \mathbf{S}^{(t+1)}, \gamma^{(t+1)}) \\ &\qquad \le \mathcal{J}(\mathbf{W}^{(t)}, \mathbf{D}^{(t)}, \mathbf{U}^{(t)}, \mathbf{S}^{(t)}, \gamma^{(t)}). \end{aligned} \tag{24}$$

TABLE I
DETAIL INTRODUCTION TO DATA SETS

Datasets	Number of samples	Features	Classes
Breast	669	10	2
Dermatology	366	34	6
Control	600	60	6
DIG	1797	64	10
JAFFE	213	256	10
IMM40	240	1024	40
ORL	400	1024	40
ATT40	400	1024	40
COIL20	1440	1024	20
PIE	3329	4096	68

According to Lemma 2, we have

$$\begin{aligned}
&\|\mathbf{w}_i^{(t+1)}\|_2 - \frac{1}{2\|\mathbf{w}_i^{(t+1)}\|_2}\|\mathbf{w}_i^{(t+1)}\|_2^2 \\
&\quad\le \|\mathbf{w}_i^{(t)}\|_2 - \frac{1}{2\|\mathbf{w}_i^{(t)}\|_2}\|\mathbf{w}_i^{(t)}\|_2^2 \\
&\quad\Rightarrow \alpha\sum_{i=1}^{d}\|\mathbf{w}_i^{(t+1)}\|_2 - \alpha\sum_{i=1}^{d}\frac{1}{2\|\mathbf{w}_i^{(t+1)}\|_2}\|\mathbf{w}_i^{(t+1)}\|_2^2 \\
&\quad\le \alpha\sum_{i=1}^{d}\|\mathbf{w}_i^{(t)}\|_2 - \alpha\sum_{i=1}^{d}\frac{1}{2\|\mathbf{w}_i^{(t)}\|_2}\|\mathbf{w}_i^{(t)}\|_2^2 \\
&\quad\Rightarrow \alpha\|\mathbf{W}^{(t+1)}\|_{2,1} - \alpha\mathrm{Tr}((\mathbf{W}^{(t+1)})^T\mathbf{D}^{(t)}\mathbf{W}^{(t+1)}) \\
&\quad\le \alpha\|\mathbf{W}^{(t)}\|_{2,1} - \alpha\mathrm{Tr}((\mathbf{W}^{(t)})^T\mathbf{D}^{(t)}\mathbf{W}^{(t)}). \qquad (25)
\end{aligned}$$

Combining the result of deduction with (24), it can be inferred that

$$\begin{aligned}
&\mathrm{Tr}((\mathbf{W}^{(t+1)})^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^{(t+1)}) + \gamma^{(t+1)}\|\mathbf{S}^{(t+1)}\|_F^2 \\
&\quad - \lambda\mathrm{Tr}((\mathbf{W}^{(t+1)})^T\mathbf{X}\mathbf{U}^{(t+1)}(\mathbf{U}^{(t+1)})^T\mathbf{X}^T\mathbf{W}^{(t+1)}) + \alpha\|\mathbf{W}^{(t+1)}\|_{2,1} \\
&\le \mathrm{Tr}((\mathbf{W}^{(t)})^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^{(t)}) + \gamma^{(t)}\|\mathbf{S}^{(t)}\|_F^2 \\
&\quad - \lambda\mathrm{Tr}((\mathbf{W}^{(t)})^T\mathbf{X}\mathbf{U}^{(t)}(\mathbf{U}^{(t)})^T\mathbf{X}^T\mathbf{W}^{(t)}) + \alpha\|\mathbf{W}^{(t)}\|_{2,1}. \qquad (26)
\end{aligned}$$

Then, it is obvious that the objective value of problem (5) is decreased by Algorithm 1 in each iteration. More information about the convergence analysis of Algorithm 1 is given in the experiment. □

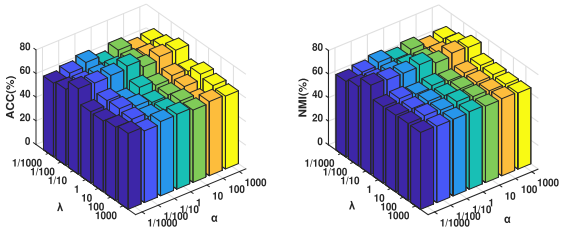
VI. EXPERIMENT

In this section, extensive experiments are performed to demonstrate the superiority and effectiveness of the proposed EGCFS method. We apply our method and the other six competitors on several benchmark data sets with the same experimental settings. Experimental results that are presented in the form of tables and figures verify the effectiveness of the proposed method, and we explain the reasons for the superiority of our method.

A. Data Sets and Compare Methods

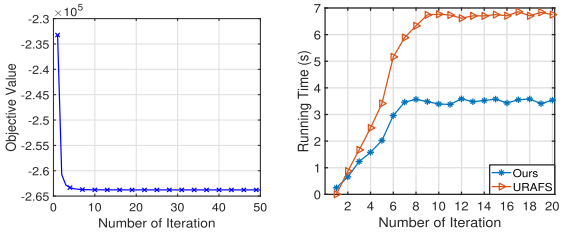
There are a total of ten data sets in the experiments, including Breast, Dermatology, Control, DIG, JAFFE, IMM40, ORL, ATT40, COIL20, and PIE. A detailed introduction to these data sets is shown in Table I. One thing to note is that Breast is a binary class data set, and we make a separate table for it to show each method's performance on this data set.

To validate the advantage of EGCFS, we compare it with several state-of-the-art unsupervised feature selection methods, including the following.



(a) (b)

Fig. 1. Effect of parameters to Algorithm 1. (a) ACC. (b) NMI.



(a) (b)

Fig. 2. Convergence and running time of Algorithm 1. (a) COIL20. (b) ORL.

JELSR is a joint embedding learning and sparse regression, in which the embedding learning and sparse regression are jointly performed [14].

SOGFS performs feature selection and local structure learning simultaneously, and the similarity matrix, thus, can be determined adaptively [33].

UMMFS is an unsupervised maximum margin feature selection algorithm via sparse constraints that combine feature selection and k-means clustering into a coherent framework [34].

URAFS is an unsupervised feature selection method that virtues a generalized uncorrelated constraint to seek the uncorrelated yet discriminative features [35].

CAN is a famous clustering method, which performs better than the spectral clustering, and we make it as a baseline [30].

B. Experimental Settings

To evaluate the performance of feature selection of all methods, we apply the selected features on a typical unsupervised task, i.e., clustering where k-means is adopted. Simultaneously, we perform k-means with all raw features as a baseline to validate the effectiveness of all methods. In order to alleviate the stochastic effect, we run ten times k-means clustering from different starting points and report the average result. When evaluating the performance of each method on the clustering task, we use two classical clustering algorithm evaluation indicators, i.e., accuracy (ACC) and normalized mutual information (NMI).

To evaluate the unbiased performance, the data set is divided randomly into two sets across multiple Monte Carlo runs. The clustering task is only computed on one-fifth of the data set. For each split, the performance on the parameter adjustment is compared for each choice of hyperparameter, and the best hyperparameter is selected under a separate validation set. This process is repeated on each Monte Carlo split, and the average across the Monte Carlos split is then recorded via fourfold validation.

As for the regularization parameters α and λ, they are set via the grid search method in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$. The parameters of other methods are set as the value for best performance. To show

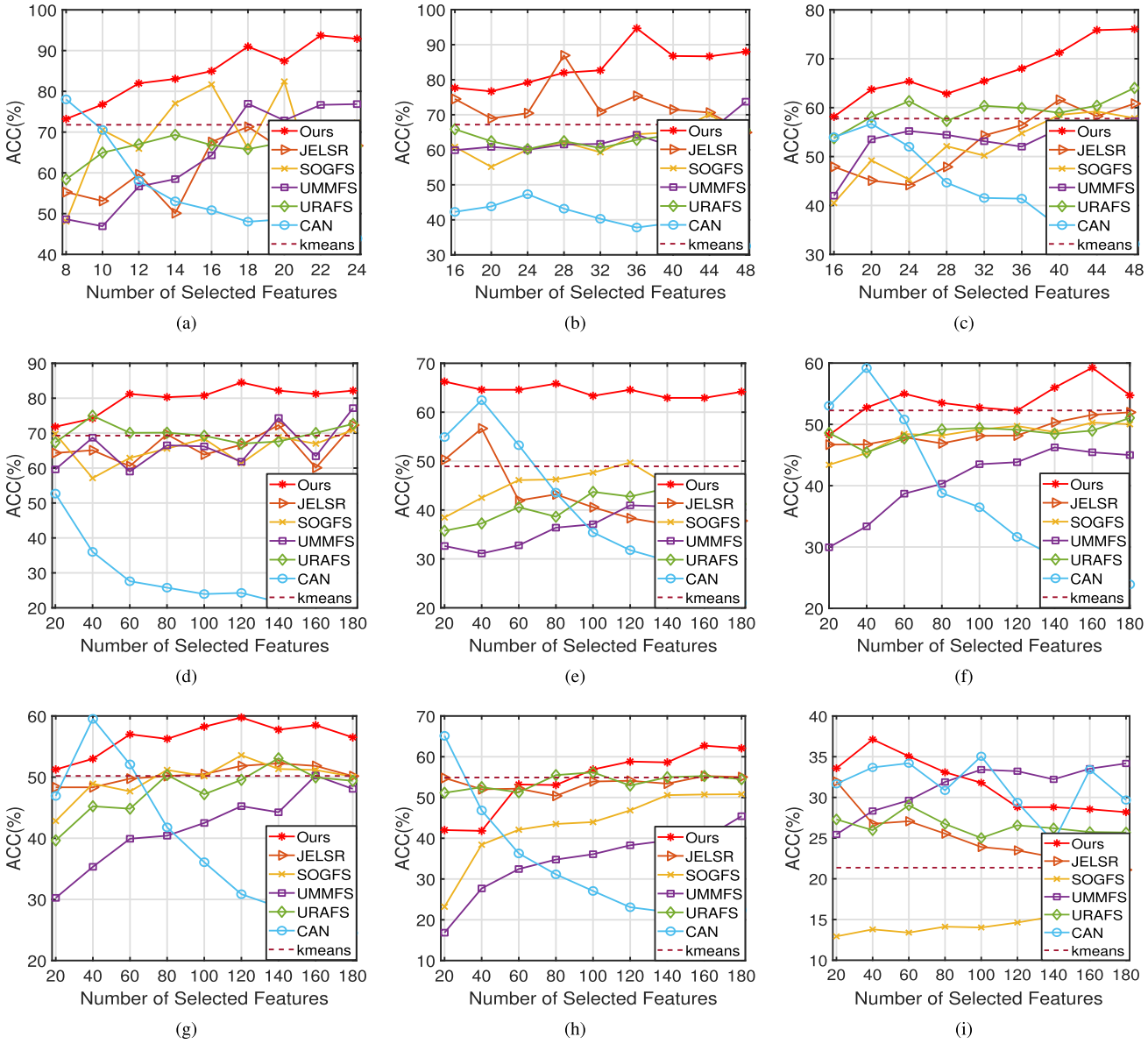


Fig. 3. Performance of all methods with different number of selected features on all data sets. (a) Dermatology. (b) Control. (c) DIG. (d) JAFFE. (e) IMM40. (f) ORL. (g) ATT40. (h) COIL20. (i) PIE.

the parameter sensitivity in our method, Fig. 1 demonstrates the influence of parameters on the Control data set.

C. Convergence Analysis and Running Time

The condition for iteration to stop is set as $|\mathcal{J}_t - \mathcal{J}_{t+1}| \leq \varepsilon$, where $\mathcal{J}$ is the objective value of the proposed EGCFS method in (23), the threshold $\varepsilon = 10^{-3}$, and t represents the number of iterations. The time complexity of the algorithm is $\mathcal{O}((d^3+n^3)t)$. Fig. 2(a) shows the convergence of objective value in (8) with the number of iterations on the COIL20 data set, and the results verify the convergence of Algorithm 1.

In terms of runtime comparison, we compare our method with the algorithm URAFS that is similar to ours and proposed in recent years. Fig. 2(b) is performed on the ORL data set, and the results show ten repeated runs with a varying number of iterations. Once the objective value converges (the decrease in objective values less than 10^{-3}), algorithms would terminate the iterations in Fig. 2(b). From Fig. 2, the proposed EGCFS method can converge fast (less than 15 iterations) and consumes less running time, which demonstrates the efficiency of our method.

D. Experimental Results and Analysis

Fig. 3 shows the performance of each method with the different number of features on nine data sets. Table II shows the best value of each method, which is obtained under the optimal number of selected features in the range of Fig. 3. Table III shows the performance of each method on the binary class data set Breast. The best result on each data set has been highlighted, and the second-best result is underlined in Table II. From the experimental results, the observations are as follows.

1) The proposed EGCFS method performs better than other state-of-the-art unsupervised feature selection approaches both on ACC and NMI in most experiments.
2) The proposed method adopts a special adaptive graph learning and constraint method in the feature selection process. The

TABLE II
PERFORMANCE OF ALL METHODS ON THE OPTIMAL NUMBER OF SELECTED FEATURES

	Methods	Dermatology	Control	DIG	JAFFE	IMM40	ORL	ATT40	COIL20	PIE
ACC(%)	Ours	**93.72**	**94.67**	**76.07**	**84.51**	**66.25**	58.25	**59.75**	62.71	**37.13**
	JELSR	71.31	86.94	61.60	72.30	56.67	51.97	52.25	55.27	31.95
	SOGFS	82.42	70.06	59.23	70.89	49.72	50.30	53.58	50.80	16.29
	UMMFS	76.96	73.72	57.24	77.15	43.33	46.25	50.25	45.41	34.17
	URAFS	69.32	65.87	64.03	74.98	44.46	51.05	53.08	56.17	29.03
	CAN	78.03	58.39	56.70	72.68	62.46	**59.18**	59.53	**65.13**	35.05
	k-means	71.79	67.17	57.76	69.28	48.92	53.60	50.19	54.90	21.37
NMI(%)	Ours	**91.84**	**86.20**	**70.97**	**85.12**	**82.49**	75.16	**78.32**	73.09	**57.73**
	JELSR	78.56	80.43	61.84	77.04	75.88	72.08	73.09	68.43	52.24
	SOGFS	81.81	68.86	59.81	74.90	73.79	71.85	73.77	67.51	38.24
	UMMFS	84.79	73.93	56.73	76.74	65.54	66.38	68.85	57.83	55.96
	URAFS	79.66	70.83	62.23	77.41	67.85	71.78	73.09	71.75	54.26
	CAN	73.94	61.67	60.24	75.81	78.11	**77.11**	77.79	**78.56**	56.16
	k-means	80.46	72.41	57.08	73.85	72.41	72.30	72.72	71.34	47.46

TABLE III
PERFORMANCE OF ALL METHODS ON THE BREAST

Breast	Ours	JELSR	SOGFS	UMMFS	URAFS	CAN
ACC(%)	**95.28**	94.76	94.71	92.68	92.68	94.99
NMI(%)	**70.53**	68.13	67.97	59.49	59.49	69.11

adaptive graph composition method makes the similarity matrix that can better reflect the subspace data structure, thus achieving better performance.

3) Benefiting from the excellent structure and special constraint, the proposed method is 10%–15% higher than other algorithms in some data sets, such as Dermatology, Control, and DIG.
4) It is noteworthy that the clustering accuracy of these methods in Fig. 3 does not always improve as the size of the feature subset increases due to redundant features and noise.
5) We deliberately added a binary class data set Breast, the experiment results in Table III show that our method is still able to perform well in such a binary class data set.

VII. CONCLUSION

This brief proposed a novel unsupervised feature selection via adaptive graph learning and constraint method, namely EGCFS. Different from the existing graph-based feature selection methods, the proposed method directly embeds graph learning into the optimization process. The adaptive graph learning can not only obtain better subspace data structure but also make the coefficient of the similarity matrix get the closed-form solution. By applying constraint on the similarity matrix, the graph-based method and the idea of maximizing the between-class scatter matrix are integrated into a uniform framework to obtain excellent structural performance. Besides, we confirm the link between the graph-based approach and k-means from a spectral perspective. The proposed method not only performs with manifold structure but also is equipped with closed-form solutions. Plenty of experiments are performed on several benchmark data sets to verify the effectiveness and superiority of the proposed method, and we explain the reasons for the superiority of our method.

VIII. FUTURE WORK

This work has achieved good performance via the graph learning and constraint method. This work is carried out under the conventional single-layer model where the deep neural network is not used. In future work, we will extend the adaptive graph learning and graph constraint in this work to the field of graph neural network (GCN). In fact, we have already started this research, and the experiment has achieved good results. If you are interested in our research, please follow our latest research progress.

REFERENCES

[1] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.
[2] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4147–4153.
[3] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *Proc. IJCAI*, 2017, pp. 2201–2207.
[4] R. Zhang and H. Tong, "Robust principal component analysis with adaptive neighbors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6959–6967.
[5] F. Wang, Q. Wang, F. Nie, Z. Li, W. Yu, and R. Wang, "Unsupervised linear discriminant analysis for jointly clustering and subspace learning," *IEEE Trans. Knowl. Data Eng.*, early access, Sep. 4, 2019, doi: 10.1109/TKDE.2019.2939524.
[6] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe, "Optimized graph learning using partial tags and multiple features for image and video annotation," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 4999–5011, Nov. 2016.
[7] M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Trans. Comput.*, vol. 65, no. 10, pp. 2986–2998, Oct. 2016.
[8] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz, "Accelerating wrapper-based feature selection with K-nearest-neighbor," *Knowl.-Based Syst.*, vol. 83, pp. 81–91, Jul. 2015.
[9] X. Li, H. Zhang, R. Zhang, and F. Nie, "Discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning," *IEEE Trans. Image Process.*, vol. 29, pp. 2139–2149, 2020.
[10] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 971–989, Sep. 2016.
[11] R. Hu *et al.*, "Graph self-representation method for unsupervised feature selection," *Neurocomputing*, vol. 220, pp. 130–137, Jan. 2017.

[12] C. Tang *et al.*, "Robust unsupervised feature selection via dual self-representation and manifold regularization," *Knowl.-Based Syst.*, vol. 145, pp. 109–120, Apr. 2018.
[13] W. Zheng, X. Zhu, Y. Zhu, R. Hu, and C. Lei, "Dynamic graph learning for spectral feature selection," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29739–29755, 2018.
[14] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.
[15] Z. Zhang and E. R. Hancock, "A graph-based approach to feature selection," in *Proc. Int. Workshop Graph-Based Represent. Pattern Recognit.* Berlin, Germany: Springer, 2011, pp. 205–214.
[16] X. Bai, L. Zhu, C. Liang, J. Li, X. Nie, and X. Chang, "Multi-view feature selection via nonnegative structured graph learning," *Neurocomputing*, vol. 387, pp. 110–122, Apr. 2020.
[17] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
[18] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Inf. Fusion*, vol. 50, pp. 158–167, 2019.
[19] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "ℓ_2,1-Norm regularized discriminative feature selection for unsupervised learning," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.
[20] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1525–1531.
[21] R. Zhang, F. Nie, and X. Li, "Self-weighted supervised discriminative feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3913–3918, Aug. 2018.
[22] J. Chen, J. Mi, and Y. Lin, "A graph approach for fuzzy-rough feature selection," *Fuzzy Sets Syst.*, vol. 391, pp. 96–116, Jul. 2020.
[23] Z. Zhang, B. Lu, Y. Liang, and E. R. Hancock, "Adaptive graph learning for unsupervised feature selection," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2015, pp. 790–800.
[24] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, Sep. 2017.
[25] D. Wang, F. Nie, and H. Huang, "Unsupervised feature selection via unified trace ratio formulation and K-means clustering (TRACK)," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2014, pp. 306–321.
[26] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[27] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," 2002, *arXiv:cs/0212008*. [Online]. Available: https://arxiv.org/abs/cs/0212008
[28] S. Xiang, F. Nie, C. Zhang, and C. Zhang, "Nonlinear dimensionality reduction with local spline embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1285–1298, Sep. 2009.
[29] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. 17th ACM Int. Conf. Multimedia (MM)*, 2009, pp. 175–184.
[30] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 977–986.
[31] S. Boyd, L. Vandenberghe, and L. Faybusovich, "Convex optimization," *IEEE Trans. Autom. Control*, vol. 51, no. 11, p. 1859, Nov. 2006.
[32] F. Nie, H. Huang, C. Xiao, and C. Ding, "Efficient and robust feature selection via joint ℓ_2, 1-norms minimization," in *Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[33] F. Nie, W. Zhu, X. Li, F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1302–1308.
[34] S. Yang, C. Hou, F. Nie, and Y. Wu, "Unsupervised maximum margin feature selection via L_2,1-norm minimization," *Neural Comput. Appl.*, vol. 21, no. 7, pp. 1791–1799, 2012.
[35] X. Li, H. Zhang, R. Zhang, Y. Liu, and F. Nie, "Generalized uncorrelated regression with adaptive graph for unsupervised feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1587–1595, May 2019.