Graph Regularized Feature Selection with Data Reconstruction

Zhou Zhao, Xiaofei He, Deng Cai, Lijun Zhang, Wilfred Ng, and Yueting Zhuang

Abstract—Feature selection is a challenging problem for high dimensional data processing, which arises in many real applications such as data mining, information retrieval, and pattern recognition. In this paper, we study the problem of unsupervised feature selection. The problem is challenging due to the lack of label information to guide feature selection. We formulate the problem of unsupervised feature selection from the viewpoint of graph regularized data reconstruction. The underlying idea is that the selected features not only preserve the local structure of the original data space via graph regularization, but also approximately reconstruct each data point via linear combination. Therefore, the graph regularized data reconstruction error becomes a natural criterion for measuring the quality of the selected features. By minimizing the reconstruction error, we are able to select the features that best preserve both the similarity and discriminant information in the original data. We then develop an efficient gradient algorithm to solve the corresponding optimization problem. We evaluate the performance of our proposed algorithm on text clustering. The extensive experiments demonstrate the effectiveness of our proposed approach.

Index Terms—Feature selection, similarity preserving, data reconstruction

1 INTRODUCTION

In many areas such as data mining and information retrieval, one is often confronted with high dimensional data. Given such high dimensional data, the time and space cost for data processing can be significantly huge [19]. Furthermore, the learning algorithm is likely to be over-fitting for the data and the result becomes less interpretable due to the *curse of dimensionality* [37]. To overcome this problem, feature selection [11], [18], [35] and feature extraction [26], [31], [44] techniques are designed to reduce the dimensionality by finding a meaningful feature subset or a set of feature combinations.

Feature selection methods can be classified as supervised feature selection method or unsupervised feature selection method. Supervised feature selection methods [9], [11], [25], [35], [36] utilize the correlation between feature and label information to guide the selection of the important features. However, in this big data era, there is no shortage of data but their labels are still very expensive. Hence, it is of great value to study the the problem of unsupervised feature selection in order to make full use of the data. In this paper, we focus on the problem of unsupervised feature selection which is particularly challenging due to the lack of label information to guide the selection of features.

Currently, the unsupervised feature selection algorithms have been widely used in text clustering [30]. In text clustering, a text or a document is always represented as a bag of words, which gives rise to the high dimensionality of the word space. The unsupervised feature selection methods choose a subset of the words from the original word space according to some criteria. There are two main categories of the unsupervised feature selection algorithms, which are similarity preserving [8], [18], [19], [56] and clustering performance maximization [3], [12], [40], [46], [49]. The similarity preserving approaches select the representative features that best preserve the local structure of the original data space. For example, if the data points are close in intrinsic geometry of the data distribution, then these data points are also close to each other on the selected features. On the other hand, the clustering performance maximization approaches select the discriminant features that can maximize certain clustering criterion. For example, Tang et al. [40] and Yang et al. [49] employ the concept of pseudo labels to select the discriminative features that maximize the clustering performance of the data points.

In this paper, we formulate the problem of unsupervised feature selection from the perspective of graph regularized data reconstruction. Our objective is to select the features that best preserve both the local structure of the original data space and the discriminant information in the original data. The underlying idea of graph regularized data reconstruction for feature selection is that each data point should be approximated by the linear combination of the selected features and the original structure of the data points is also preserved on the selected features. Thus, the graph regularized data reconstruction error becomes a natural criterion

- Z. Zhao and Y. Zhuang are with the College of Computer Science, Zhejiang University, 388 Yu Hang Tang Road, Hangzhou, Zhejiang 310058, China. E-mail: {zhaozhou, yzhuang}@zju.edu.cn.
- L. Zhang is with the Department of Computer Science and Technology, Nanjing University, Xianlin Campus Mailbox 603, 163 Xianlin Avenue, Qixia District, Nanjing, 210023, China. E-mail: zljzju@gmail.com.
- W. Ng is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China. E-mail: wilfred@cse.ust.hk.
- X. He and D. Cai are with the State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, 388 Yu Hang Tang Road, Hangzhou, Zhejiang 310058, China. E-mail: {xiaofeihe, dengcai}@gmail.com.

Manuscript received 14 Jan. 2015; revised 6 Sept. 2015; accepted 18 Sept. 2015. Date of publication 26 Oct. 2015; date of current version 2 Feb. 2016. Recommended for acceptance by J. Ye.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2493537

for measuring the quality of the selected features. We then propose a gradient algorithm to solve the composite objective function of graph regularized data reconstruction for feature selection.

It is worthwhile to highlight several aspects of the proposed approach here:

- We formulate the problem of unsupervised feature selection from the viewpoint of graph regularized data reconstruction. By minimizing the graph regularized reconstruction error, we can select the features that preserve both the structure and discriminant information of the original data space.
- We consider the process of feature selection via sparse learning over the composite objective function. We introduce the sparsity induced norm l_1-norm for the feature selection matrix. The sparsity of the feature selection matrix reduces the redundant or noisy features.
- We propose an iterative gradient algorithm to solve the proposed optimization problem for feature selection. We evaluate the effectiveness of our proposed approach using extensive experiments on text clustering.

The rest of the paper is organized as follows: Section 2 surveys the related work. We present the problem of feature selection from the viewpoint of graph regularized data reconstruction in Section 3. We next propose the composite objective function for discriminant and similarity preserving feature selection in Section 4. We then provide an iterative gradient method for solving the optimization problem in Section 5. A variety of experimental results are presented in Section 6. Finally, we provide the concluding remarks in Section 7.

2 Related Work

The problem of feature selection is to choose a subset of the original features based on a certain criterion. According to the way of utilizing label information, feature selection algorithms can be divided into two categories: supervised feature selection algorithms [23], [29], [39], [45], [48], [52], [59] and unsupervised feature selection algorithms [6], [8], [10], [13], [14], [17], [18], [20], [28], [33], [34], [38], [40], [41], [42], [46], [49], [53]. In this section, we focus more on the unsupervised features selection algorithms, which are mainly based on the criteria of similarity preserving and clustering performance maximization.

2.1 Similarity Preserving Based Feature Selection

The feature selection algorithms in the category of similarity preserving have been widely studied in [8], [18], [34], [53], which select the features that best preserve the local structure of the original data. The similarity preserving criteria for feature selection are unified in [58], which is reformulated as

$$\min_{F_{sub}} \sum_{f \in F_{sub}} \mathbf{f}^T K \mathbf{f},$$

where $\mathbf{f}$ is the feature vector of the data matrix and K is the predefined affinity matrix of data points in the original data space. Thus, the features which are consistent with the manifold structure are considered to be important.

Laplacian score [18] and its extension [47], [54], [55], [56] are the typical similarity preserving methods for feature selection. The idea of Laplacian score is to evaluate the importance of the feature according to its similarity preserving power, which is based on graph model. Some other similarity preserving algorithms are proposed based on different criteria of similarity preserving. Cai et al. [8] propose the multi-cluster structure preserving method for unsupervised feature selection called MCFS. MCFS is based on the spectral analysis of the data and L1-regularized regression model for feature selection. Zhao et al. [53] propose a manifold-based maximum margin method for unsupervised feature selection.

2.2 Clustering Based Feature Selection

The feature selection algorithms in the category of clustering performance maximization have been studied in [10], [12], [15], [28], [34], [38], [40], [41], [42], [43], [46], [49], [57], which aim to select the discriminant features. The clustering based feature selection methods utilize the concept of pseudo-labels, and then jointly perform the feature selection and generate pseudo-labels for data instances. Li et al. [28] perform spectral clustering to learn the pseudo-labels of the data instances, during which the feature selection is performed simultaneously. Yang et al. [49] incorporate discriminative analysis and $l_{2,1}$-norm minimization into a joint framework for unsupervised feature selection, under the assumption that the pseudo-label of input data instances can be predicted by a linear classifier. Dy and Brodley [13] select the features based on the clustering quality. Tang et al. [40] incorporate the discriminant analysis and pesudo-label generation into a joint framework for unsupervised feature selection. Qian and Zhai [38] learn the pseudo cluster labels via local learning regularized robust nonnegative matrix factorization, where $l_{2,1}$ norm minimization is employed on processes of both label learning and feature learning. Tang et al. [41], [42], [43] propose the unsupervised feature selection framework for social media data by considering both instance selection and multi-view of the data. Wolf and Shashua [46] select the features based on least-squares optimization process, which measures the clusterability of the data points on selected features. Masaeli et al. [32] introduce the concept of projection matrix to eliminate the redundancy of selected features and Farahat et al [15] devise an efficient recursive algorithm for that. Du et al. [10] integrate the regression model to detect the cluster structure and perform feature selection. Li et al. [27] leverage both cluster analysis and sparse structure for unsupervised feature selection. Hu et al. [21] develop the unsupervised feature selection method for indexing photo.

Currently, the probabilistic model has been used to tackle the problem of unsupervised feature selection [6], [12], [14], [17]. We consider that the latent variable is the pseudo-label of the data and classify these approach in the category of clustering based feature selection. Dy and Brodley [12] propose a wrapper approach based on expectation maximization. Boutemedjet et al. [6] propose a generative model that clusters the visual features and users into separate classes. Guan et al. [17] propose a unified probabilistic

TABLE 1
Summary of Notation

Notation	Notation Description
$\mathbf{X} \in \mathcal{R}^{n\times m}$	a data matrix of data points
$\mathbf{F} \in \mathcal{R}^{m\times n}$	a feature matrix of data points
$\Lambda \in \mathcal{R}^{n\times n}$	a feature selection matrix of features
$\mathcal{S}$	a selected feature set
$\mathbf{A} \in \mathcal{R}^{n\times n}$	a reconstruction coefficient matrix
$\mathbf{W} \in \mathcal{R}^{n\times n}$	a similarity matrix of data points
$\mathbf{D} \in \mathcal{R}^{n\times n}$	a diagonal matrix of data points
$\mathbf{L} \in \mathcal{R}^{n\times n}$	a Laplacian matrix of data points
$g_1(\cdot), \ldots, g_n(\cdot)$	a set of reconstruction functions
α	a l_1-norm regularization term
β	a Laplacian regularization term

model for global and local unsupervised feature selection. Fan et al. [14] propose a variational inference framework for unsupervised non-Gaussian feature selection.

Unlike the previous studies, we formulate the problem as unsupervised graph regularized feature selection with data reconstruction. We aim to select the features that best preserve both the similarity and discriminant information in the original data.

3 The Problem of Graph Regularized Feature Selection with Data Reconstruction

In this section, we first introduce the notation used in our subsequence discussion, which includes data matrix $\mathbf{X}$, feature matrix $\mathbf{F}$, feature selection matrix $\boldsymbol{\Lambda}$, selected feature set $\mathcal{S}$ and reconstruction coefficient matrix $\mathbf{A}$. We then present the problem of unsupervised feature selection from the perspective of data reconstruction.

We consider that the data matrix is $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ where m is the number of data points and each data point $\mathbf{x}_i$ is represented by an n-dimensional vector (i.e., $\mathbf{x}_i \in \mathcal{R}^n$). That is, there are n features for all the data points in $\mathbf{X}$. Then we denote the row vectors of $\mathbf{X}$ by $\mathbf{f}_i^T \in \mathcal{R}^m$, $(i = 1, \ldots, n)$, each of which corresponds to a feature, and define the feature matrix $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$ (i.e., $\mathbf{F} = \mathbf{X}^T$). We consider that the vector $\mathbf{f}_i = (x_{i1}, \ldots, x_{im})^T$ is the projection of all data points on the ith feature. We consider that the feature selection indicator is defined as $\lambda_i = 1$ if and only if the ith feature is selected. We next denote the selected feature set $\mathcal{S} = \{i | 1 \leq i \leq n, \lambda_i = 1\}$. We let the feature selection vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and then denote the feature selection matrix

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

which is an $n \times n$ diagonal matrix of vector $\boldsymbol{\lambda}$ (i.e., $\Lambda = \mathbf{diag}(\boldsymbol{\lambda})$). Thus, $\mathbf{F}\boldsymbol{\Lambda}$ (i.e., the transpose of $\boldsymbol{\Lambda}\mathbf{X} \in \mathcal{R}^{n\times m}$) is the projection of the data matrix $\mathbf{X}$ on the selected features. The reconstruction coefficient matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n) \in \mathcal{R}^{n\times n}$ is used to reconstruct the original data matrix $\mathbf{X}$ from the projection $\boldsymbol{\Lambda}\mathbf{X}$, where vector $\mathbf{a}_i$ is the coefficient for reconstruct the ith feature of the data points. The notation is summarized in Table 1.

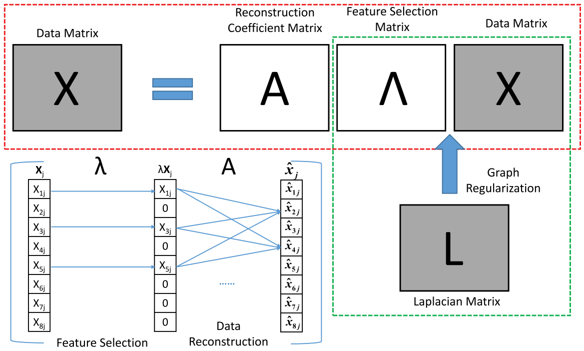


Fig. 1. The framework of graph regularized feature selection with data reconstruction: The red block illustrates the data reconstruction process while the green block illustrates the graph regularized process. The entries in data matrix $\mathbf{X}$ and laplacian matrix $\mathbf{L}$ are known and the entries in reconstruction coefficient matrix $\mathbf{A}$ and feature selection matrix $\boldsymbol{\Lambda}$ are for the optimization. The learning of the feature selection matrix $\boldsymbol{\Lambda}$ preserves both the data reconstruction process and the graph regularized process.

Using the notation above, we define the problem of unsupervised feature selection as follows. Given a data matrix $\mathbf{X}$ and a feature matrix $\mathbf{F}$, we aim to choose the optimal feature set $\mathcal{S}$ of size k from the n features such that the local structure and the discriminant information in the original data space can be best preserved.

4 The Objective Function

The goal of feature selection is to reduce the dimension of the data points on the number of features for high dimensional data processing. Following the principle of dimensional reduction, we want to have a compact representation of the original data points on the selected features. That is, we aim to learn the feature selection matrix Λ for projecting the original data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, where the information loss is minimized and the local structure of the original data space is also preserved. We now present the composite objective function of the graph regularized data reconstruction for feature selection.

The Principle Component Analysis (PCA) [22] is a well-known feature extraction algorithm which extracts the discriminant features based on information loss. Inspired from the idea of the PCA algorithm, we propose a new feature selection criteria based on data reconstruction. That is, we consider that each data point $\mathbf{x}_i$ can be reconstructed via the linear combination from its projected values on the selected features $\Lambda\mathbf{x}_i$ (and the original data $\mathbf{X}$ can be reconstructed from its compact representation $\Lambda\mathbf{X} = (\Lambda\mathbf{x}_1, \ldots, \Lambda\mathbf{x}_m)$). We then learn the feature selection matrix Λ with the minimum reconstruction error.

We first introduce the data reconstruction criteria from the viewpoint of data features and then present the details of the graph regularized data reconstruction for the data points in the original space. We illustrate the framework of graph regularized data reconstruction for feature selection in Fig. 1.

Consider a set of selected features $\mathcal{S} = \{j | 1 \leq j \leq n, \lambda_j = 1\}$ of size k and the corresponding compact representation of the original data $\mathbf{F}\Lambda$. Given the ith feature of all the data points $\mathbf{f}_i = (x_{i1}, \ldots, x_{im})^T$, we present the data reconstruction for the ith feature from $\mathcal{S}$. We denote the data reconstruction function for the ith feature by $g_{\mathbf{a}_i}(\mathcal{S})$ where $\mathbf{a}_i \in \mathcal{R}^n$ is the reconstruction coefficient. Thus, the information loss of the data reconstruction for the ith feature based on $\mathcal{S}$ is given by

$$\mathcal{L}(\mathbf{f}_i, g_{\mathbf{a}_i}(\mathcal{S})) = \|\mathbf{f}_i - g_{\mathbf{a}_i}(\mathcal{S})\|^2, \tag{1}$$

and the total data reconstruction error for all the features is given by

$$\mathcal{L}(\mathbf{F}, g_{\mathbf{A}}(\mathcal{S})) = \sum_{i=1}^{n} \|\mathbf{f}_i - g_{\mathbf{a}_i}(\mathcal{S})\|^2, \tag{2}$$

where $\|\cdot\|$ is the l_2-norm and $\mathbf{A} \in \mathcal{R}^{n\times n}$ is the reconstruction coefficient matrix.

In this paper, we consider that the original data $\mathbf{X}$ is approximately constructed from $\mathcal{S}$ via linear function. For example, the reconstruction function for the data points on the ith feature is given by

$$g_{\mathbf{a}_i}(\mathcal{S}) = \sum_{j\in\mathcal{S}} a_{ij}\mathbf{f}_j = \sum_{j=1}^{n} a_{ij}(\mathbf{f}_j\lambda_j) = \sum_{j=1}^{n} \lambda_j a_{ij}\mathbf{f}_j, \tag{3}$$

which is the linear combination of the selected features in the set $\mathcal{S}$. By doing so, the data points on the ith feature can be approximately reconstructed by $\mathbf{f}_i \approx \sum_{j=1}^{n} \lambda_j a_{ij}\mathbf{f}_j$.

Therefore, the total reconstruction error is given by

$$\begin{aligned}\mathcal{L}(\mathbf{F}, g_{\mathbf{A}}(\mathcal{S})) &= \sum_{i=1}^{n}\left\|\mathbf{f}_i - \sum_{j=1}^{n}\lambda_j a_{ij}\mathbf{f}_j\right\|^2\\ &= \|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2\\ &= \sum_{j=1}^{m}\|\mathbf{x}_j - \mathbf{A}\Lambda\mathbf{x}_j\|^2,\end{aligned} \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm and each data point $\mathbf{x}_j$ is linearly reconstructed from its compact representation on the selected features $\Lambda\mathbf{x}_j$. By minimizing the reconstruction error, we can learn the new features $\mathcal{S}$ such that the original data $\mathbf{X}$ can be well reconstructed from its compact representation $\Lambda\mathbf{X}$.

We illustrate the framework of data reconstruction in Fig. 1. Given the set of selected features $\mathcal{S}$ and reconstruction coefficient matrix $\mathbf{A}$, we give the example of data reconstruction for the jth data point $\mathbf{x}_j = (x_{1j}, x_{2j}, x_{3j}, x_{4j}, x_{5j}, x_{6j}, x_{7j}, x_{8j})^T$. Suppose that the first, third and fifth feature is selected for all the data points in $\mathbf{X}$, then the set of selected features is $\mathcal{S} = \{1, 3, 5\}$ where the feature selection matrix $\Lambda = diag(1, 0, 1, 0, 1, 0, 0, 0)$. The projection of the jth data point is $\Lambda\mathbf{x}_j = (x_{1j}, 0, x_{3j}, 0, x_{5j}, 0, 0, 0)^T$. The data point $\mathbf{x}_j$ can be reconstructed by $\mathbf{x}_j \approx \mathbf{A}\Lambda\mathbf{x}_j$ where its second feature value can be reconstructed by $x_{2j} \approx \mathbf{a}_2\Lambda\mathbf{x}_j = a_{21}x_{1j} + a_{23}x_{3j} + a_{25}x_{5j}$. Thus, the data reconstruction process for the data matrix $\mathbf{X}$ is given by $\mathbf{X} \approx \mathbf{A}\Lambda\mathbf{X}$, which is illustrated by the red block in Fig. 1. The feature selection for data reconstruction is to learn the diagonal matrix Λ such that the reconstruction error $\|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2$ is minimized.

We further want to select the features that also preserve the intrinsic structure in the original data space. A natural assumption is that if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in intrinsic geometry in the original data space, then the projection of two points on the selected features $\Lambda\mathbf{x}_i$ and $\Lambda\mathbf{x}_j$, are also close to each other. This assumption is based on the theory of *local invariance* [4], which has been widely used in various kinds of algorithms including dimensionality reduction [4], semi-supervised learning [5] and matrix factorization [7].

Recent studies in spectral graph theory and manifold learning theory have demonstrated that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scatter of data points [7], [18]. Consider a graph with m vertices where each vertex corresponds to a data point. For each data point $\mathbf{x}_i$, we find its nearest neighbor and put edges between $\mathbf{x}_i$ and its neighbor. We consider the *0-1 weighting* to build the weight matrix $\mathbf{W}$ on the graph where the weight $w_{ij} = 1$ if and only if nodes i and j are connected by an edge.

Let $\mathbf{D}$ be the diagonal matrix with $D_{ii} = \sum_j W_{ij}$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ be the Laplacian matrix. Thus, the local geometrical information of the data on the selected features can be best preserved by minimizing [4]:

$$\begin{aligned}&\frac{1}{2}\sum_{i,j=1}^{m}\|\Lambda\mathbf{x}_i - \Lambda\mathbf{x}_j\|^2 W_{ij}\\ &= \frac{1}{2}\sum_p\left[\sum_{i,j=1}^{n} W_{ij}(\lambda_p x_{ip} - \lambda_p x_{jp})^2\right]\\ &= \sum_p\left(\sum_i \lambda_p x_{ip} D_{ii} x_{ip}\lambda_p - \sum_{i,j}\lambda_p x_{ip} W_{ij} x_{ip}\lambda_p\right)\\ &= \sum_p \lambda_p\left(\sum_i x_{ip} D_{ii} x_{ip} - \sum_{i,j} x_{ip} W_{ij} x_{ip}\right)\lambda_p\\ &= \sum_p \lambda_p\big(\mathbf{X}\mathbf{L}\mathbf{X}^T\big)\lambda_p\\ &= tr\big(\Lambda\mathbf{X}\mathbf{L}\mathbf{X}^T\Lambda\big).\end{aligned} \tag{5}$$

The process of graph regularization is illustrated in the green block in Fig. 1. The feature selection for graph regularized data reconstruction is to learn the diagonal matrix Λ such that both the reconstruction error and graph regularization are minimized.

We then obtain the following optimization problem on the feature selection matrix Λ and the reconstruction coefficient matrix $\mathbf{A}$, given by

$$\begin{aligned}\min\nolimits_{\mathbf{A},\Lambda}\quad & \|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2 + \beta tr\big(\Lambda\mathbf{X}L\mathbf{X}^T\Lambda\big)\\ s.t.\quad & \Lambda = diag(\boldsymbol{\lambda}), Card(\boldsymbol{\lambda}) = k\\ & \lambda_i \in \{0, 1\}, i = 1, \ldots, n,\end{aligned} \tag{6}$$

where trace $tr(\cdot)$ represents the graph regularization for the data matrix $\mathbf{X}$ projected on the selected features, and $\beta \geq 0$ is a tradeoff parameter. The cardinality of the feature selection vector is k (i.e., $Card(\boldsymbol{\lambda}) = k$).

However, we observe that Problem 6 is computationally intractable, since it requires branch-and-bound procedure to optimize integer variables $\boldsymbol{\lambda}$. We thus relax the constraint of the variables in integer vector $\boldsymbol{\lambda}$ to allow them to take real numbers. This relaxation is commonly used in the area of sparse learning [51]. Then, λ_j corresponds to a scaling factor indicating how significantly the jth feature contributes to the minimization of the graph regularized data reconstruction error. We also notice that forcing the diagonal matrix Λ to have more zeros implies that fewer features would be selected. Therefore, we enforce the sparsity of the diagonal matrix Λ by employing the l_1-norm regularization condition on it. Then, the objective function can also be written as

$$\min_{\mathbf{A},\Lambda} \|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2 + \beta tr\left(\Lambda\mathbf{X}\mathbf{L}\mathbf{X}^T\Lambda\right) + \alpha\|\boldsymbol{\lambda}\|_1, \tag{7}$$

where the l_1-norm regularization on the feature selection vector $\|\boldsymbol{\lambda}\|_1$ controls the size of the selected features and also ensures the selection matrix Λ is suitable for feature selection (i.e., by reducing the redundant or even noisy features). The regularization coefficient β is employed to balance the objectives of preserving discriminant information and similarity in the original data space. The similarity is preserved with a large value of β while the discriminant information is preserved on the other hand. With a large value of β, Equation (7) can be considered as similarity preserving based feature selection. The regularization coefficient α is used to control the number of the selected features. With a larger value of α, the vector $\boldsymbol{\lambda}$ becomes more sparse.

5 The Optimization

In this section, we design a gradient method to solve Problem 7. Following the iterative optimization method in [24], we divide the gradient method into two steps: learning the feature selection matrix Λ while fixing the reconstruction coefficient matrix $\mathbf{A}$, and learning reconstruction coefficient matrix $\mathbf{A}$ while fixing the feature selection matrix Λ.

5.1 Learning Feature Selection Matrix Λ

In this section, we discuss how to solve the optimization Problem (7) by fixing the reconstruction coefficients $\mathbf{A}$. Then, the Problem (7) can be reduced as follows:

$$\min_{\Lambda} \|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2 + \beta tr\left(\Lambda\mathbf{X}\mathbf{L}\mathbf{X}^T\Lambda\right) + \alpha\|\boldsymbol{\lambda}\|_1, \tag{8}$$

which is a l_1-norm regularized optimization problem. Unfortunately, we observe that Problem (8) is nondifferentiable when the feature selection vector $\boldsymbol{\lambda}$ contains values of 0. Therefore, the standard unconstrained optimization methods cannot be applied to solve this problem. In this work, we introduce an optimization method based on coordinate descent to solve this problem. It is easy to verify that Problem (8) is convex, thus, the global minimum can be achieved.

We notice that the reconstruction error $\|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2$ can be rewritten as follows:

$$\begin{aligned}\|\mathbf{X} - \mathbf{A}\Lambda\mathbf{X}\|_F^2 &= \sum_{j=1}^{m} \|\mathbf{x}_j - \mathbf{A}\Lambda\mathbf{x}_j\|^2 \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n}\left(x_{kj} - \sum_{i=1}^{n}\lambda_i a_{ik} x_{ij}\right)^2.\end{aligned} \tag{9}$$

We then denote the matrix $\mathbf{Y} = \mathbf{X}\mathbf{L}\mathbf{X}^T$, thus the Laplacian regularizer $tr\left(\Lambda\mathbf{X}\mathbf{L}\mathbf{X}^T\Lambda\right)$ can be rewritten as follows:

$$\begin{aligned}tr\left(\Lambda\mathbf{X}\mathbf{L}\mathbf{X}^T\Lambda\right) &= tr(\Lambda\mathbf{Y}\Lambda) \\ &= tr\left(\sum_{i=1}^{n} Y_{ii}\lambda_i\lambda_i\right) \\ &= \sum_{i=1}^{n} Y_{ii}\lambda_i\lambda_i,\end{aligned} \tag{10}$$

where Λ is a diagonal matrix.

Combing Equations (9) and (10), Problem (8) can be rewritten as:

$$\begin{aligned}\min_{\Lambda} \quad & \sum_{j=1}^{m}\sum_{k=1}^{n}\left(x_{kj} - \sum_{i=1}^{n}\lambda_i a_{ik} x_{ij}\right)^2 + \alpha\sum_{i=1}^{n}|\lambda_i| \\ & + \beta\sum_{i=1}^{n} Y_{ii}\lambda_i\lambda_i.\end{aligned} \tag{11}$$

When we infer the variable λ_p for the pth feature, we keep other variables $\{\lambda_i\}_{i\neq p}$ fixed. Thus, we get the following optimization problem:

$$\begin{aligned}\min_{\lambda_p} f(\lambda_p) = &\sum_{j=1}^{m}\sum_{k=1}^{n}\left(r_{jk}^{-p} - \lambda_p a_{pk} x_{pj}\right)^2 + \alpha|\lambda_p| \\ & + \beta L_{pp}\lambda_p\lambda_p,\end{aligned} \tag{12}$$

where $r_{kj}^{-p} = x_{kj} - \sum_{i\neq p}\lambda_i a_{ik} x_{ij}$. The residue r_{kj}^{-p} is the reconstruction error for kth feature of jth data without considering the pth feature.

However, we notice that the regularizer $\alpha|\lambda_p|$ of Problem (11) is not differentiable at 0. To tackle this non-differential problem, we follow the sub-gradient strategy of the feature-sign search algorithm [24]. We define $h(\lambda_p) = \sum_{j=1}^{m}\sum_{k=1}^{n}(r_{kj}^{-p} - \lambda_p a_{pk} x_{pj})^2 + \beta L_{pp}\lambda_p\lambda_p$, and then let $f(\lambda_p) = h(\lambda_p) + \alpha|\lambda_p|$. We define $\frac{\partial|\lambda_p|}{\partial\lambda_p}$ to be the sub-differential value of $|\lambda_p|$. Based on the definition of sub-gradients [24], $\frac{\partial|\lambda_p|}{\partial\lambda_p}$ is a sub-gradient of $|\lambda_p|$ if and only if

$$|\lambda_p'| \geq |\lambda_p| + \frac{\partial|\lambda_p|}{\partial\lambda_p}(|\lambda_p'| - |\lambda_p|). \tag{13}$$

We can observe that when $|\lambda_p| > 0$, the sub-gradient of the absolute value function $|\lambda_p|$ is given by $\frac{\partial|\lambda_p|}{\partial\lambda_p} = sign(\lambda_p)$ (i.e., $\frac{\partial|\lambda_p|}{\partial\lambda_p} \in \{-1, +1\}$). If $\lambda_p = 0$, then the sub-gradient value $\frac{\partial|\lambda_p|}{\partial\lambda_p}$ is in the set $[-1, 1]$ (i.e., $|\lambda_p'| \geq \frac{\partial|\lambda_p|}{\partial\lambda_p}|\lambda_p'|$). Thus, the conditions for obtaining the optimal value of $f(\lambda_p)$ can be translated into the following expression:

$$\begin{cases}\frac{\partial}{\partial\lambda_p}h(\lambda_p) + \alpha sign(\lambda_p) = 0 & \text{if } |\lambda_p| > 0 \\ |\frac{\partial}{\partial\lambda_p}h(\lambda_p)| \leq \alpha & \text{if } \lambda_p = 0.\end{cases} \tag{14}$$

Then, we consider how to select the optimal feature selection variables λ_p when the conditions for obtaining the optimal value are violated (i.e., $|\frac{\partial}{\partial\lambda_p}h(\lambda_p)| > \alpha$, if $\lambda_p = 0$). Suppose in the case that $\frac{\partial}{\partial\lambda_p}h(\lambda_p) > \alpha \geq 0$, we aim to search the new value of λ_p in order to satisfy the conditions (i.e.,

$\frac{\partial}{\partial \lambda_p} h(\lambda_p) + \alpha sign(\lambda_p) = 0$, if $|\lambda_p| > 0$) such that the value of $f(\lambda_p)$ can be decreased. Thus, we let the sign of λ_p be negative (i.e., $sign(\lambda_p) = -1$) and compute the value of λ_p. Similarly, if $\frac{\partial}{\partial \lambda_p} h(\lambda_p) < -\alpha \le 0$, we let the sign of λ_p be positive. In this way, we can replace the l_1-norm of λ_p by λ_p if $sign(\lambda_p) = 1$, by $-\lambda_p$ if $sign(\lambda_p) = -1$ or by 0. Thus, Problem (11) can be reduced to an unconstrained quadratic programming problem which can be solved by the existing optimization methods.

Algorithm 1. Computing Feature Selection Matrix

Input: A data set of m data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, the coefficient matrix $\mathbf{A}$, the graph Laplacian matrix $\mathbf{L}$, the parameters α and β

1: **Initialize step:**

2: $\boldsymbol{\lambda} = \overrightarrow{0}$, $\theta = \overrightarrow{0}$, and active set $S = \emptyset$, where $\theta_p \in \{-1, 0, 1\}$ denotes sign(λ_p).

3: **Activate step:**

4: From zero coefficient of $\boldsymbol{\lambda}$, select $\lambda_p = \arg\max_{\lambda_p} |\frac{\partial}{\partial \lambda_p} h(\lambda_p)|$. Active λ_p only if it locally improves the objective function, namely:

5: If $\frac{\partial}{\partial \lambda_p} h(\lambda_p) > \alpha$, then set $\theta_p = -1$, $S = \{p\} \cup S$.

6: If $\frac{\partial}{\partial \lambda_p} h(\lambda_p) < -\alpha$, then set $\theta_p = 1$, $S = \{p\} \cup S$.

7: **Feature-sign step:**

8: (a) Compute the solution to the resulting unconstrained QP corresponding to the *active set* S:

$$\begin{aligned} \min\nolimits_{\Lambda} \quad & \sum_{j=1}^{m} \sum_{k=1}^{n} \left(x_{kj} - \sum_{p \in S} \lambda_p a_{pk} x_{pj} \right)^2 \\ + \quad & \alpha \sum_{p \in S} \theta_p \lambda_p + \beta \sum_{p \in S} Y_{pp} \lambda_p \lambda_p, . \end{aligned}$$

9: Let $\frac{\partial}{\partial \lambda_p} f(\lambda_p) = 0$, we can get the optimal value of λ_p under the current *active set* S:

$$\begin{aligned} \lambda_p^{new} = & \left(\sum_{j=1}^{m} \sum_{k=1}^{n} a_{pk}^2 x_{pj}^2 + \beta L_{pp} \right)^{-1} \\ & \times \left[\sum_{j=1}^{m} \sum_{k=1}^{n} r_{jk}^{-p} a_{pk} x_{pj} - \frac{\alpha \theta_p}{2} \right]. \end{aligned}$$

10: (b) Perform a discrete line search on the closed line segment from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}^{new}$: Check the objective value at $\boldsymbol{\lambda}^{new}$ and all points where any coefficient changes sign, and update $\boldsymbol{\lambda}$ to the point with the lowest objective value.

11: (c) Remove zero coefficients of $\boldsymbol{\lambda}$ from the *active set* and update $\boldsymbol{\theta} = sign(\boldsymbol{\lambda})$.

12: **Check the optimality conditions step:**

13: **Condition (a)**: Optimal condition for nonzero coefficients: $\frac{\partial}{\partial \boldsymbol{\lambda}} h(\boldsymbol{\lambda}) + \alpha sign(\boldsymbol{\lambda}) = 0$

14: If condition (a) is not satisfied, go to Feature-sign step; else check condition (b).

15: **Condition (b)**: Optimal condition for zero coefficients: $|\frac{\partial}{\partial \boldsymbol{\lambda}} h(\boldsymbol{\lambda})| \le \alpha$

16: If condition (b) is not satisfied, go to Active step; otherwise return $\boldsymbol{\lambda}$ as the solution.

17: **return** the optimal feature selection matrix $\boldsymbol{\Lambda}$.

We outline the procedure of computing the feature selection matrix $\boldsymbol{\Lambda}$ in Algorithm 1. We keep an active set S for the potentially nonzero feature selection variables $\{p|\lambda_p = 0, |\frac{\partial}{\partial \lambda_p} h(\lambda_p)| > \alpha\}$ and their corresponding signs $\{\theta_1, \ldots, \theta_k\}$. At each iteration, Algorithm 1 selects the variable $\boldsymbol{\lambda}'$ whose violation is the largest (i.e., $\lambda'_p = \arg\max_{\lambda_p} |\frac{\partial}{\partial \lambda_p} h(\lambda_p)|$, $|\frac{\partial}{\partial \lambda_p} h(\lambda_p)| > \alpha$) and adds it to S. The optimization of the variables in S can be done as follows: First, the new analytic solution to unconstrained quadratic programming is computed as $\boldsymbol{\lambda}^{new}$, then, an efficient line search between the current value and the new value $\boldsymbol{\lambda}^{new}$ is invoked. The feature selection matrix $\boldsymbol{\Lambda}$ is returned when all variables satisfy the optimality conditions. The algorithmic procedure of learning feature selection matrix is as follows:

1) For each variable for the pth feature $\lambda_p \in \boldsymbol{\lambda}$, search for its sign $\theta_p \in \boldsymbol{\theta}$;
2) Solve the reduced unconstrained Problem (8) to get the optimal $\boldsymbol{\lambda}^*$ which minimizes the objective function of graph regularized data reconstruction error;
3) Return the optimal feature selection matrix $\boldsymbol{\Lambda}^* = diag(\boldsymbol{\lambda}^*)$

Convergence analysis. We show the convergence of Algorithm 1 by verifying that the solutions to the optimization in feature-sign step will strictly decrease the objective function. Suppose $\boldsymbol{\lambda}^{new}$ is a new solution for the feature selection variable $\boldsymbol{\lambda}$. If the sign of $\boldsymbol{\lambda}^{new}$ is same with the variable $\boldsymbol{\lambda}$ in the active set, then the solution $\boldsymbol{\lambda}^{new}$ is consistent with the objective function and definitely decreases the value of the objective function. On the other hand, if the sign of $\boldsymbol{\lambda}^{new}$ is different with the variable $\boldsymbol{\lambda}$, then we carry out the line search from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}^{new}$ and find a consistent solution $\boldsymbol{\lambda}^c$ whose sign is the same with $\boldsymbol{\lambda}$. Thus, it also makes the value of the objective function decrease.

5.2 Learning Reconstruction Coefficient Matrix A

We now describe the method of learning the coefficient matrix $\mathbf{A}$, while fixing the diagonal feature selection matrix Λ (i.e., $\Lambda = diag(\boldsymbol{\lambda})$). Thus, Problem (8) becomes an unconstrained least squares problem given by

$$\min\nolimits_{\mathbf{A}} \quad \|\mathbf{X} - \mathbf{A}\boldsymbol{\Lambda}\mathbf{X}\|_F^2 \tag{15}$$

and the solution to this problem is as follows

$$\mathbf{A} = \mathbf{X}\mathbf{X}^T \boldsymbol{\Lambda} (\boldsymbol{\Lambda}\mathbf{X}\mathbf{X}^T \Lambda)^{-1}. \tag{16}$$

Note that the l_1-norm of $\boldsymbol{\lambda}$ enforces some elements to be zeros. If the jth diagonal entry of feature selection matrix $\mathbf{L}$ is zero (i.e., $\lambda_j = 0$), then all $a_{1j}, \ldots, a_{nj}$ must be zero which means the jth feature is not selected.

Our graph regularized feature selection with data reconstruction is presented in Algorithm 2. The algorithm selects the features that minimize the graph regularized data reconstruction error, computes the feature selection matrix $\boldsymbol{\Lambda}$ in Step 1 and reconstruction coefficient matrix $\mathbf{A}$ in Step 2 alternatively, and terminates when the feature selection matrix becomes convergent.

6 Experimental Results

In this section, we study the effectiveness of our proposed unsupervised feature selection method. The experiments are conducted by using Matlab, tested on machines with Linux OS Intel(R) Core(TM2) Quad CPU 2.66 Hz, and 32 GB RAM.

Algorithm 2. Graph Regularized Data Reconstruction for Feature Selection (GRFS)

Input: A data matrix of m data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, the graph Laplacian matrix $\mathbf{L}$, the parameters α and β

1: Initialize reconstruction coefficient matrix $\mathbf{A}_0$, and feature selection matrix $\mathbf{\Lambda}_0$, $k = 1$
2: **repeat**
3: **Step 1.** Update $\mathbf{\Lambda}_k \leftarrow$ Algorithm 1
4: **Step 2.** Update $\mathbf{A}_k \leftarrow \mathbf{X}\mathbf{X}^T\mathbf{\Lambda}_k(\mathbf{\Lambda}_k\mathbf{X}\mathbf{X}^T\mathbf{\Lambda}_k)^{-1}$
5: $k \leftarrow k+1$
6: **until** $\|\mathbf{\Lambda}_k - \mathbf{\Lambda}_{k-1}\|_F < \epsilon$
7: **return** feature selection matrix $\mathbf{\Lambda}_k$

6.1 Data Preparation

We evaluate the performance of the algorithms using the TDT2 and the Routers document corpora.

The TDT2 corpus [1] consists of data collected from the first half of 1998 and taken from the following six sources, including two newswires (APW, NYT), two radio programs (VOA, PRI) and two television programs (CNN, ABC). The dataset consists of 11,201 on-topic documents which are classified into 96 semantic categories. In this experiment, those documents appearing in two or more categories are removed, and the categories with more than 10 documents are kept, thus leaving with 10,021 documents in total.

The Reuters corpus [2] contains 21,578 documents which are grouped into 135 clusters. Compared with TDT2 corpus, the Reuters corpus is more difficult for clustering. In TDT2, the content of each cluster is narrowly defined, whereas in Reuters, documents in each cluster have a broader variety of content. Moreover, the Reuters corpus is much more unbalanced, with some large clusters more than 300 times larger than some small ones. In our study, we discard the documents having multiple category labels, and only select the categories with more than 10 documents. This results in 8,213 documents in total.

In both of the two corpora, the stop words are removed and each document is represented as a *tf-idf* vector. We rank the words based on their *tf-idf* score and choose the top 1,000 words as the feature of each document.

In this following sections, we will evaluate our approach in two aspects. First, we compare our method with popular unsupervised feature selection approaches for tackling text clustering by varying the number of selected features. Then, we study the robustness of our algorithm by varying both the graph regularizer β and l_1-norm regularizer α.

6.2 Evaluation Criteria

Following the previous unsupervised feature selection work [18], [50], we evaluate the performance of our method in terms of clustering.

The clustering algorithm generates a cluster label for each data point. The clustering performance is evaluated by comparing the generated class label and the ground truth. In our experiments, the accuracy (AC) and the normalized mutual information metric (NMI) are used to measure the clustering performance. Given a data point x_i, let r_i and l_i be the obtained cluster label and the label provided by the ground truth. The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^{m} \delta(l_i, map(r_i))}{m},$$

where m is the total number of samples and $\delta(x, y)$ is equal 1 if $x = y$ and 0 otherwise. The $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data set. The best mapping can be found by using the Kuhn-Munkres algorithm [16].

Let C denote the set of clusters obtained from the ground truth and C' obtained from our algorithm. Their mutual information metric $MI(C, C')$ is defined as follows:

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)},$$

where $p(c_i)$ and $p(c'_j)$ are the probabilities that a sample point arbitrarily selected from the data set belongs to the cluster c_i and c'_j, respectively. The $p(c_i, c'_j)$ is the joint probability that the arbitrarily selected data point belongs to the cluster c_i as well as c'_j at the same time. In our experiments, we use the normalized mutual information as follows:

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))},$$

where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. It is easy to check that $NMI(C, C')$ ranges from 0 to 1. In particular, we have that $NMI = 1$, if the two sets of clusters are identical, and $NMI = 0$, if the two sets are independent.

To check whether the difference between GRFS and other approaches is significant, we perform the paired t-test on both the accuracy and normalized mutual information evaluation criteria between the result of GRFS and that of other approaches.

6.3 Performance Evaluations and Comparisons

In this section, we demonstrate the effectiveness of our proposed method (*GRFS*) by performing k-means text clustering by using only the selected features. The following four unsupervised feature selection algorithms are used for the comparison:

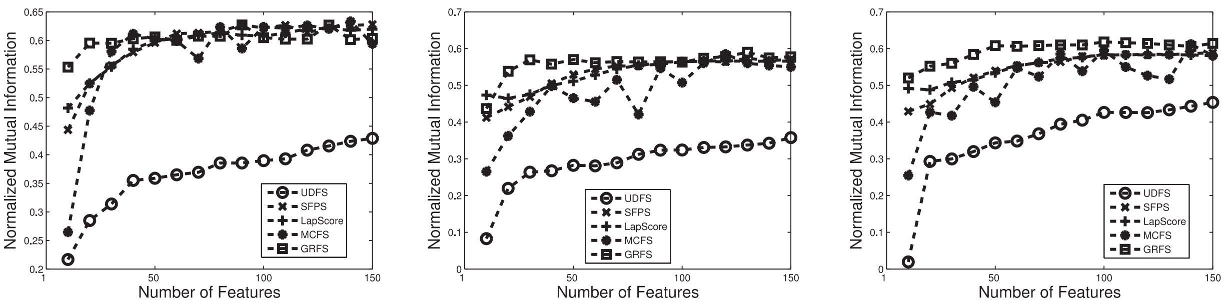
- *AllFea*: All original features are adopted.
- *LapScore*: Laplacian Score [18] selects the features that best preserve the similarity of the original data space.
- *SPFS*: Spectral Feature Selection method [56] selects the features according to the structures of the graph induced from the pairwise instance similarity.
- *UDFS*: Unsupervised Discriminative Feature Selection method [50] selects the features that preserve the discriminative information and feature correlations simultaneously.
- *MCFS*: Multi-cluster feature selection method [8] selects the features that preserve the multi-cluster structure of the data.

TABLE 2
Clustering Performance by Using 50 Features on the TDT2 Corpus

	5 clusters		7 clusters		9 clusters		average	
	AC	NMI	AC	NMI	AC	NMI	AC	NMI
UDFS	0.5907 ± 0.1444	0.4517 ± 0.1296	0.4804 ± 0.1369	0.3522 ± 0.1521	0.3924 ± 0.1444	0.3016 ± 0.1521	0.4878	0.3685
SPFS	0.7508 ± 0.1369	0.6391 ± 0.1444	0.6138 ± 0.1521	0.5609 ± 0.1444	0.5771 ± 0.1521	0.5649 ± 0.1369	0.6472	0.5883
MCFS	0.7133 ± 0.1444	0.6234 ± 0.2025	0.6125 ± 0.1936	0.5603 ± 0.1444	0.5665 ± 0.1521	0.5849 ± 0.1024	0.6308	0.5895
LapScore	0.7282 ± 0.2116	0.628 ± 0.25	0.5866 ± 0.2025	0.5446 ± 0.2401	0.5718 ± 0.1849	0.5649 ± 0.2116	0.6288	0.5792
GRFS	$\mathbf{0.7581 \pm 0.1444}$	$\mathbf{0.6858 \pm 0.1444}$	$\mathbf{0.6312 \pm 0.1369}$	$\mathbf{0.5871 \pm 0.1521}$	$\mathbf{0.582 \pm 0.1444}$	$\mathbf{0.63 \pm 0.1444}$	0.6571	0.6343
AllFea	0.681 ± 0.16	0.6044 ± 0.1521	0.5739 ± 0.1521	0.5448 ± 0.1521	0.5246 ± 0.1444	0.5043 ± 0.1521	0.5932	0.5511

TABLE 3
Clustering Performance by Using 50 Features on the Reuters Corpus

	5 clusters		7 clusters		9 clusters		average	
	AC	NMI	AC	NMI	AC	NMI	AC	NMI
UDFS	0.5155 ± 0.1296	0.3135 ± 0.1444	0.4384 ± 0.1369	0.2656 ± 0.2025	0.3541 ± 0.1369	0.1984 ± 0.1024	0.436	0.2591
SPFS	0.5571 ± 0.1156	0.3965 ± 0.1296	$\mathbf{0.4662 \pm 0.1225}$	0.3628 ± 0.1296	0.3757 ± 0.1225	0.2975 ± 0.1369	0.4663	0.3523
MCFS	0.5542 ± 0.1444	0.3888 ± 0.1764	0.4491 ± 0.1936	0.3448 ± 0.1369	0.3859 ± 0.1156	0.312 ± 0.1521	0.4631	0.3485
LapScore	0.546 ± 0.1849	0.39 ± 0.2116	0.4602 ± 0.1764	0.3573 ± 0.2209	0.3753 ± 0.1369	0.2962 ± 0.2116	0.4605	0.3478
GRFS	$\mathbf{0.6075 \pm 0.1369}$	$\mathbf{0.4741 \pm 0.1369}$	0.465 ± 0.1444	$\mathbf{0.39 \pm 0.1369}$	$\mathbf{0.39 \pm 0.1444}$	$\mathbf{0.3166 \pm 0.1444}$	0.4875	0.3936
AllFea	0.5501 ± 0.0961	0.3803 ± 0.1024	0.4681 ± 0.1156	0.3811 ± 0.1156	0.3696 ± 0.1089	0.303 ± 0.1156	0.4626	0.3548



(a) 6 Class (b) 8 Class (c) 10 Class

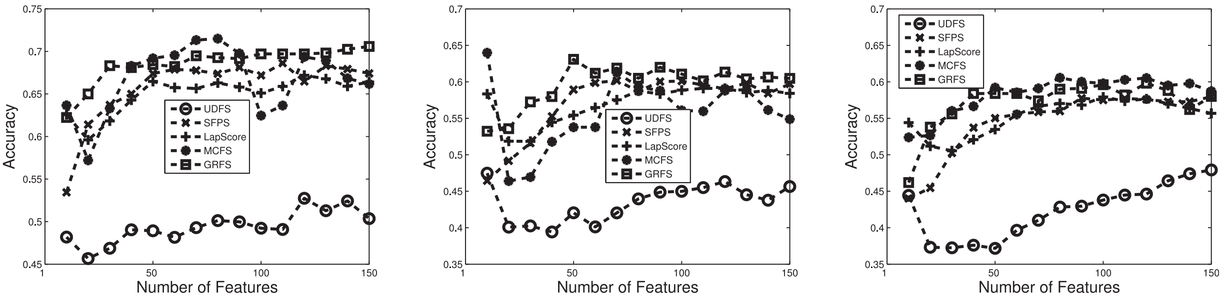
Fig. 2. Effect of the number of features on mutual information on the TDT2 corpus.

The methods *LapScore* and *SPFS* are in the category of similarity preserving based feature selection while the method *UDFS* and *MCFS* are in the category of clustering based feature selection.

For each data set, the evaluations were conducted by using different number of clusters (k). For performing the text clustering on both TDT2 and Reuters corpus, we choose $k = 5, 6, 7, 8, 9, 10$. For each given cluster number k, 10 test runs were conducted on different randomly chosen clusters, and the final performance score were computed by averaging the score from the 10 tests. We also record the standard deviation of the error rate of clustering performance. For each test, we applied the k-means clustering algorithm on different number of selected features. The k-means algorithm was applied 10 times with different starting points and the best results in terms of the objective function were recorded. Table 2 shows the clustering performance, in terms of normalized mutual information and accuracy for the TDT2 corpus. Table 3 shows the clustering performance for the Reuters corpus. The number of clusters is taken to be 5, 7 and 9, and the number of selected features is set to 50. The last two columns of each table record the average clustering performance for different feature selection methods. The last row of each table records the clustering performance by using all the features.

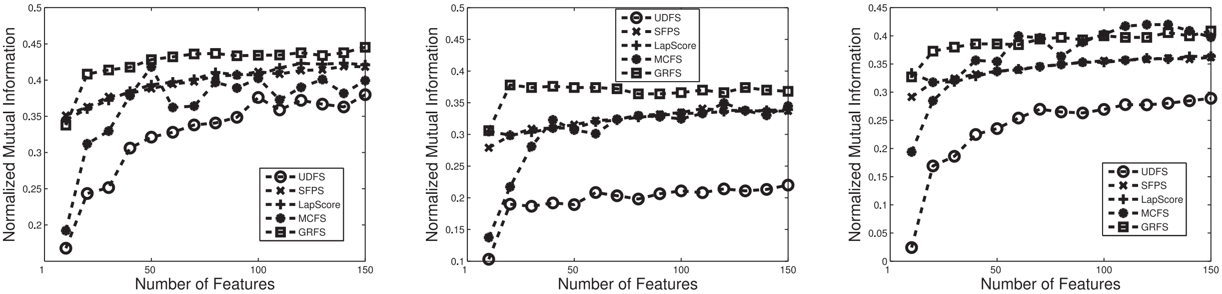
The objective of feature selection is to reduce the dimension of the data on the number of features. We thus illustrate the clustering performance of the feature selection methods versus the number of selected features. The number of clusters is taken to be 6, 8 and 10, and the number of selected features is set from 1 to 150. Figs. 2a, 2b and 2c illustrate the clustering performance of feature selection methods on the number of selected in terms of normalized mutual information for the TDT2 corpus. Figs. 3a, 3b and 3c demonstrate the clustering performance in terms of accuracy for the TDT2 corpus. Figs. 4a, 4b, 4c, 5a, 5b and 5c show the clustering performance on the number of selected features for the Reuters corpus. We observe that our method converges faster than other feature selection methods on the number of features for text clustering.

We show the significant different between GRFS and other approaches using the paired t-test on both the accuracy and normalized mutual information evaluation criteria between the result of GRFS and that of other approaches.



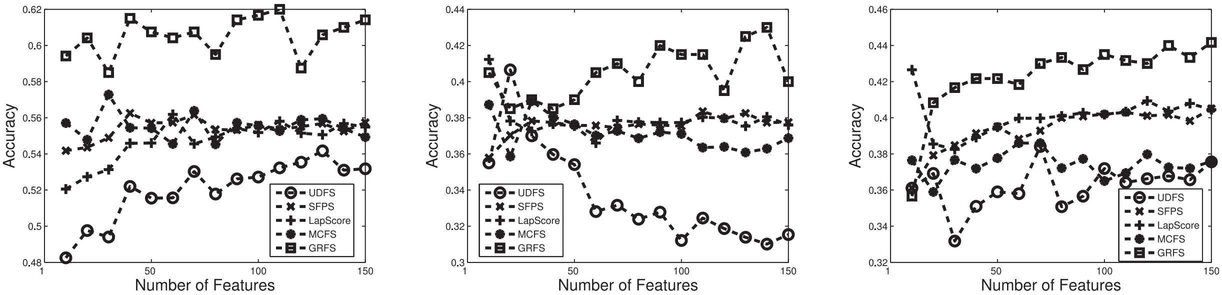
(a) 6 Class (b) 8 Class (c) 10 Class

Fig. 3. Effect of the number of features on accuracy on the TDT2 corpus.



(a) 6 Class (b) 8 Class (c) 10 Class

Fig. 4. Effect of the number of features on mutual information on the Reuters corpus.



(a) 6 Class (b) 8 Class (c) 10 Class

Fig. 5. Effect of the number of features on accuracy on the Reuters corpus.

TABLE 4
The Associated p-Values of the Paired t-Test of GRFS Over other Algorithms on 5 Clusters Using Reuters Corpus

Evaluation criteria	UDFS	SPFS	MCFS	LapScore	AllFea
NMI	1.59×10^{-4}	5.66×10^{-3}	5.61×10^{-3}	3.76×10^{-3}	2.24×10^{-3}
AC	2.56×10^{-6}	5.19×10^{-4}	4.74×10^{-4}	7.39×10^{-4}	9.36×10^{-5}

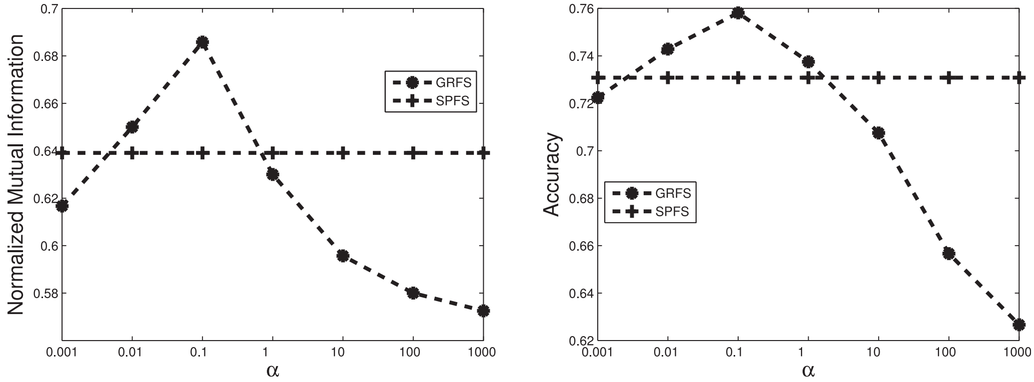
We illustrate the associated p-values on five clusters using Reuters corpus in Table 4. The test at the 99 percent confidence interval demonstrates that our proposed framework can obtain very encouraging and promising results compared to the others.

These experiments reveal a number of interesting points:

- The unsupervised feature selection methods achieve better performance over AllFea method. This suggests that the feature selection improves the performance of text clustering.
- The similarity preserving based methods, both SPFS and LapScore outperform the UDFS method, which shows that the similarity preserving based feature selections are effective on text clustering.
- In all the cases, our GRFS method achieves the best performance. This indicates that the graph regularized feature selection with data reconstruction criteria can further improve the performance of text clustering.

6.4 Parameters Selection

There are two essential parameters in our approach. One is the l_1-norm regularizer α and another is the graph regularizer β. The l_1-norm regularizer α controls the sparsity of



(a) Varying α on Normalized Mutual Information (b) Varying α on Accuracy

Fig. 6. The performance of GRFS versus parameter α.

the feature selection vector $\boldsymbol{\lambda}$. We vary the value of α from 10^{-3} to 10^{3}, and show the evaluation results in Figs. 6a and 6b. We vary the value of β to investigate the benefits of our method from graph regularization and data reconstruction. We vary the value of β from 10^{-3} to 10^{3}, and show the evaluation results in Figs. 7a and 7b. The data set used for this experiment is the TDT2 corpus. We perform the k-means clustering algorithm on the 50 selected features of our method.

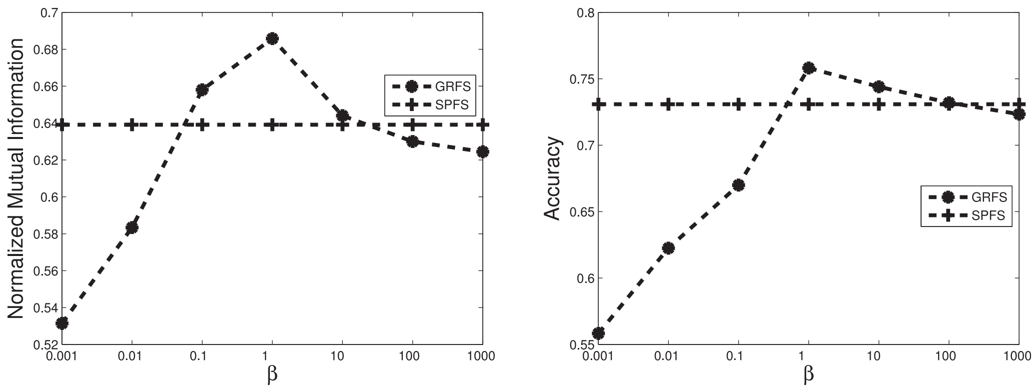
We notice that the SPFS method consistently outperforms other methods in most of the test cases. Thus, we mainly compare our method with SPFS method on the TDT2 corpus by varying the values of α and β. The performance trend of our method by varying these two parameters is similar on the Reuters corpus.

As we can see, the performance of GRFS is very stable with respect to α and β. The GRFS method almost achieve consistently good performance when α varies from 10^{-2} to 10^{0} and β varies from 10^{-1} to 10^{1} in Figs. 6a, 6b, 7a and 7b.

As we have explained in previous sections, the GRFS method selects the features that preserve the similarity and discriminant information in the original data space by minimizing the graph regularized data reconstruction error. The parameter α is used to enforce the sparsity of the feature selection matrix. When the value of α is appropriate, the feature selection matrix is able to reduce the redundant or noisy features. On the other hand, when the value of α becomes extremely large, the features are randomly selected due to the sparsity of the feature selection matrix. The parameter β is the trade-off parameter for feature selection between graph regularization for preserving similarity and data reconstruction for preserving discriminant information of the original data space. When the value of β is large, our method GRFS is considered as in the category of similarity preserving based feature selection method. Thus, its clustering performance is relatively low, which is similar to the performance of SPFS and LapScore (based on similarity preserving). This is the reason why the performance of GRFS method varies according to the values of α and β in Figs. 6a, 6b, 7a and 7b.

7 Conclusion

We formulate the problem of unsupervised feature selection from a new perspective of graph regularized data reconstruction. We consider that the discriminant information can be preserved by selecting the features that minimizes the data reconstruction error. We also preserve the similarity of the original data space by graph regularized feature selection. Our approach integrates both data reconstruction and graph regularization seamlessly into a common framework that tackles the problem of unsupervised feature selection. In this way, our approach selects the features that best preserve the similarity and discriminant information in the original data space via the minimization of the graph regularized data reconstruction error. We devise a novel gradient method to solve the optimization problem. We conduct



(a) Varying β on Normalized Mutual Information (b) Varying β on Accuracy

Fig. 7. The performance of GRFS versus parameter β.

several experiments on the text clustering for TDT2 and Routers corpus. The experimental results demonstrate that our method achieves higher clustering performance compared with three state-of-the-art feature selection algorithms. On the other hand, our method can also be extended to the problem of supervised feature selection. The simplest way is to incorporate the label information for the graph regularization. For example, if two data points have the same label, we can assign a relatively larger weight on the edge connecting them. Thus, our proposed framework of graph regularized feature selection with data reconstruction can be used for both unsupervised and supervised feature selection.

ACKNOWLEDGMENTS

The authors would like to express their thanks to the editor and the reviewers for their careful revisions and insightful suggestions. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB336500, and in part by the National Natural Science Foundation of China under Grant 61233011 and Grant 61125203, and in part by the China Knowledge Centre for Engineering Sciences and Technology (CKCEST), and in part by HKUST FSGRF13EG22 and HKUST FSGRF14EG31.

REFERENCES

[1] (1999). Nist topic detection and tracking corpus [Online]. Available: http://www.nist.gov/speech/tests/tdt/tdt98/index.htm
[2] (2005). Reuters-21578 corpus [Online]. Available: http://www.daviddlewis.com/resources/testcollections/reuters21578/
[3] S. Basu, C. A. Micchelli, and P. Olsen, "Maximum entropy and maximum likelihood criteria for feature selection from multivariate data," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2000, vol. 3, pp. 267–270.
[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 14, pp. 585–591.
[5] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
[6] S. Boutemedjet, D. Ziou, and N. Bouguila, "Unsupervised feature selection for accurate recommendation of high-dimensional image data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007.
[7] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[8] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.
[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2012.
[10] L. Du, Z. Shen, X. Li, P. Zhou, and Y.-D. Shen, "Local and global discriminative learning for unsupervised feature selection," in *Proc. IEEE 13th Int. Conf. Data Mining*, 2013, pp. 131–140.
[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.
[12] J. G. Dy and C. E. Brodley, "Feature subset selection and order identification for unsupervised learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 247–254.
[13] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, 2004.
[14] W. Fan, N. Bouguila, and D. Ziou, "Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1670–1685, Jul. 2013.
[15] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "An efficient greedy method for unsupervised feature selection," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 161–170.
[16] A. Frank, "On kuhn's hungarian methoda tribute from hungary," *Naval Res. Logistics*, vol. 52, no. 1, pp. 2–5, 2005.
[17] Y. Guan, M. I. Jordan, and J. G. Dy, "A unified probabilistic model for global and local unsupervised feature selection," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1073–1080.
[18] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, vol. 186, p. 189.
[19] X. He, M. Ji, C. Zhang, and H. Bao, "A variance minimization criterion to feature selection using Laplacian regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2013–2025, Oct. 2011.
[20] Y. Hou, P. Zhang, T. Yan, W. Li, and D. Song, "Beyond redundancies: A metric-invariant method for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 348–364, Mar. 2010.
[21] J. Hu, J. Pei, and J. Tang, "How can I index my thousands of photos effectively and automatically? An unsupervised feature selection approach," in *Proc. 14th SIAM Int. Conf. Data Mining*, 2014, pp. 136–144.
[22] I. Jolliffe. *Principal Component Analysis*. New York, NY, USA: Wiley, 2005.
[23] X. Kong, W. Fan, and P. S. Yu, "Dual active feature and sample selection for graph classification," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 654–662.
[24] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, vol. 19, no. 801, pp. 801–808.
[25] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The Am. Statist.*, vol. 42, no. 1, pp. 59–66, 1988.
[26] X. Li and Y. Pang, "Deterministic column-based matrix decomposition," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 1, pp. 145–149, Jan. 2010.
[27] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, 2014.
[28] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. AAAI*, 2010.
[29] J. Liang, F. Wang, C. Dang, and Y. Qian, "A group incremental approach to feature selection applying rough set technique," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 294–308, 2014.
[30] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation on feature selection for text clustering," in *Proc. Int. Conf. Mach. Learn.*, 2003, vol. 3, pp. 488–495.
[31] W. Liu, D. Tao, and J. Liu, "Transductive component analysis," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 433–442.
[32] M. Masaeli, Y. Yan, Y. Cui, G. Fung, and J. G. Dy, "Convex principal feature selection," in *Proc. SDM*, 2010, pp. 619–628.
[33] C. Maung and H. Schweitzer, "Pass-efficient unsupervised feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1628–1636.
[34] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.
[35] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, vol. 23, pp. 1813–1821.
[36] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proc. AAAI*, 2008, vol. 2, pp. 671–676.
[37] R. Duda O. Richard, P. E. Hart, and D. G. Stork, "Pattern classification," John Wiley & Sons, 2012.
[38] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1621–1627.
[39] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.
[40] J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection," *SDM*, SIAM, pp. 938–946, 2014.
[41] J. Tang, X. Hu, H. Gao, and H. Liu, "Unsupervised feature selection for multi-view data in social media," in *Proc. SDM*, 2013, pp. 270–278.
[42] J. Tang and H. Liu, "Unsupervised feature selection for linked social media data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 904–912.
[43] J. Tang and H. Liu, "Coselect: Feature selection with instance selection for social media data," in *Proc. SDM*, 2013, pp. 695–703.

[44] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General averaged divergence analysis," in *Proc. 7th IEEE Int. Conf. Data Mining*, 2007, pp. 302–311.

[45] J. Wang, P. Zhao, S. Hoi, and R. Jin, "Online feature selection and its applications," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 698–710, 2014.

[46] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," *J. Mach. Learn. Res.*, vol. 6, pp. 1855–1887, 2005.

[47] 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.

[48] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye, "Feature grouping and selection over an undirected graph," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 922–930.

[49] 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. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.

[50] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l2, 1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, vol. 22, p. 1589.

[51] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 1081–1088.

[52] M. Zhang, C. Ding, Y. Zhang, and F. Nie, "Feature selection at the discrete limit," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1355–1361.

[53] B. Zhao, J. Kwok, F. Wang, and C. Zhang, "Unsupervised maximum margin feature selection with manifold regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 888–895.

[54] J. Zhao, K. Lu, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, vol. 71, no. 10, pp. 1842–1849, 2008.

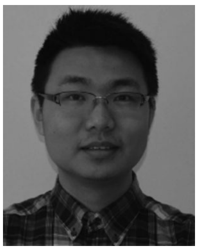
[55] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in *Proc. SDM*, 2007, pp. 641–646.

[56] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.

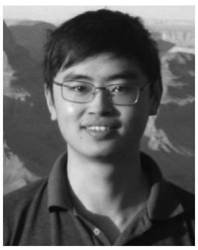
[57] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proc. AAAI*, 2012.

[58] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.

[59] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, and J. Ye, "Feafiner: Biomarker identification from medical data through feature generalization and selection," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1034–1042.



Zhou Zhao received the BS and PhD degrees in computer science from the Hong Kong University of Science and Technology (HKUST), in 2010 and 2015, respectively. He is currently a lecturer with the College of Computer Science, Zhejiang University. His research interests include machine learning and data mining.



Xiaofei He received the BS degree in computer science from Zhejiang University, Hangzhou, China, in 2000, and the PhD degree in computer science from the University of Chicago, Chicago, IL, in 2005. He is currently a professor with the College of Computer Science, Zhejiang University. Prior to joining Zhejiang University, he was a research scientist with Yahoo! Research Labs, Burbank, CA.

Deng Cai received the PhD degree in computer science from the University of Illinois at Urbana Champaign in 2009. He is a professor in the State Key Lab of CAD&CG, College of Computer Science at Zhejiang University, China. His research interests include machine learning, data mining, and information retrieval.

Lijun Zhang received the BS and PhD degrees in software engineering and computer science from Zhejiang University, China, in 2007 and 2012, respectively. He is currently an associate professor in the Department of Computer Science and Technology, Nanjing University, China. Prior to joining Nanjing University, he was a postdoctoral researcher in the Department of Computer Science and Engineering, Michigan State University. His research interests include machine learning, optimization, information retrieval, and data mining.

Wilfred Ng received the MSc (Distinction) and PhD degrees in computer science from the University of London. Currently, he is an associate professor of computer science and engineering at the Hong Kong University of Science and Technology, where he is a member of the database research group. His research interests are in the areas of databases, data mining, and information systems, which include Web data management and social networks. Further information can be found at the following URL: http://www.cs.ust.hk/faculty/wilfred/index.html.

Yueting Zhuang received the BSc, MSc, and PhD degrees in computer science from Zhejiang University, China, in 1986, 1989, and 1998, respectively. From February 1997 to August 1998, he was a visiting scholar in Prof. Thomas Huang's group, the University of Illinois at Urbana-Champaign. Currently, he is a full professor and the dean of the College of Computer Science, Zhejiang University. His research interests mainly include artificial intelligence, multimedia retrieval, computer animation, and digital library.