2013 IEEE 13th International Conference on Data Mining

# Local and Global Discriminative Learning for Unsupervised Feature Selection

Liang Du*†, Zhiyong Shen‡, Xuan Li‡, Peng Zhou*† Yi-Dong Shen*
*State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
†University of Chinese Academy of Sciences, Beijing 100049, China
‡Baidu Inc. Beijing China {duliang,lixuan,ydshen}@ios.ac.cn, {shenzhiyong,lixuan04}@baidu.com

***Abstract*—In this paper, we consider the problem of feature selection in unsupervised learning scenario. Recently, spectral feature selection methods, which leverage both the graph Laplacian and the learning mechanism, have received considerable attention. However, when there are lots of irrelevant or noisy features, such graphs may not be reliable and then mislead the selection of features. In this paper, we propose the Local and Global Discriminative learning for unsupervised Feature Selection (LGDFS), which integrates a global and a set of locally linear regression model with weighted $\ell_2$-norm regularization into a unified learning framework. By exploring the discriminative and geometrical information in the weighted feature space, which alleviates the effects of the irrelevant features, our approach can find the most representative features to well respect the cluster structure of the data. Experimental results on several benchmark data sets are provided to validate the effectiveness of the proposed approach.**

## I. Introduction

In many applications of machine learning and data mining, one is often confronted with very high dimensional data. Therefore, feature selection has become increasingly important since it can speed up the learning process, alleviate the curse of dimensionality, and even provide significant insights into the nature of the problem [1], [2].

In recent years, a lot of methods have been proposed to address the problem of unsupervised feature selection [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. These methods usually use the graph Laplacian to characterize the structure of high dimensional data. The selection of features is performed according to either some specified criterion or sparse spectral regression model.

Though many spectral feature selection algorithms have been proposed, there are still two problems at least not properly addressed. One problem is on the exploited graph, which is used to characterize the desired structure of the data, e.g., discriminative and geometrical information. It has been pointed out [10] that the performance of these feature selection algorithms is largely determined by the effectiveness of graph construction. However, these methods use all features with the same importance to compute the graph. Since the importance of features are different, it is natural to construct graph using weighted features. Yet it is still difficult to choose appropriate weight for each feature [5]. The other problem is on the exploited spectral regression model, which selects the most representative features through sparsity regularization to preserve the cluster structure detected from the graph Laplacian. Though the nonlinear geometric structure has been characterized by various graphs, they are only used to detect the cluster structure. The subsequent sparsity regularized linear regression models mainly select the relevant features to preserve the most linearly separable clusters and fail to preserve the nonlinear separable clusters.

In this paper, we propose the Local and Global Discriminative Learning for unsupervised Feature Selection (LGDFS) to select the most representative features which consistently respect both the local and global cluster structure of the data. Particularly, we introduce the Discriminative Feature Selection (DFS) cost function, a weighted $\ell_2$ norm regularized linear regression model by attaching a weight to each feature under the discriminative clustering framework [13], [14], [15], to evaluate the relevance for features. Our goal is to select these features that well respect the most linearly separable clusters. Moreover, due to the non-linearly separable nature of many high-dimensional data, we also split the whole data set into multiple overlapped local regions and employ the DFS model on each local region to select informative features for separating local data points. Thus, to select the most discriminative features which consistently respect the cluster structure across all local regions and the whole data set, we estimate the sparse weights for features by aggregating all local and global models into a unified framework. The induced optimization problem can be solved by iterating the learning of graph embedding and the estimation of feature weight until convergence. Experimental results on benchmark data sets demonstrate the effectiveness of the proposed approach.

Compared with previous work such as [10], [9], the main advantages of the proposed algorithm are as follows. On one hand, to detect cluster structure in the data, our method iteratively constructs the graph Laplacian through local learning in the weighted feature space, which alleviates the side effects of irrelevant features. Thus, it is expected to better characterize the intrinsic structure of the data and improve the estimation of cluster structure. On the other hand, to evaluate the relevance of features, we aggregate all local and global regression models with weighted $\ell_2$-norm regularization into a unified framework. In this way, our method selects the most representative features to best

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.23

respect both the linearly and nonlinearly separable clusters.

The reminder of the paper is organized as follows. In Section 2, we provide a brief review of the related work. In Section 3, we present the proposed LGDFS method and the optimization scheme in detail. In Section 4, we discuss the relations between LGDFS and other methods. In Section 5, extensive experiments are conducted to show the effectiveness of the proposed method. Finally, the conclusions and future works are discussed in Section 6.

## II. Related Work

Feature selection has been extensively studied in last decades. Based on whether the label information is available, feature selection algorithms can be roughly classified into two categories, i.e., supervised and unsupervised methods. Based on whether to optimize the performance of the learning algorithm, feature selection methods can also be classified into filter and wrapper methods.

Supervised filter methods usually evaluate the feature importance by the correlation between feature and class label. The typical methods include Fisher score, information gain, and relief [16]. Supervised wrapper methods select the most relevant features by optimizing the performance of the learning algorithm. The typical methods include Lasso [17], LARs [18], SVM-RFE [19], and RFS [20].

Due to the absence of class label, unsupervised feature selection is a much harder problem. Unsupervised filter methods usually select features to best preserve the structure of the data according to certain criteria. The typical algorithms in this category include Maximum Variance, Laplacian Score [3], SPEC [4], EVSC [5]. Laplacian Score selects features which can best preserve the manifold structure of the data. EVSC aims at seeking these features with high impact on graph Laplacian's eigenvalues. A limitation of these approaches is that the correlation among features is neglected [6], [10]. For unsupervised wrapper methods, clustering is a commonly used learning algorithm to measure the quality of features. The typical ones include Q-a [21], MCFS [6], MRSF [7], FSSL [8], UDFS [9], and JELSR [10]. These algorithms apply the following two steps separately [6], [7], [8] or jointly [9], [10]: 1)estimating the cluster structure via spectral embedding of graph Laplacian, and 2)estimating the weights for features along the embedding using sparsity regularization models, i.e., $\ell_1$-norm [6] and $\ell_{2,1}$-norm [7]–[10] regularized spectral regression.

## III. The Proposed Method

The generic problem of feature selection is as follows. Given a data matrix $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^{d\times n}$, whose columns $\boldsymbol{x}_i \in \mathbb{R}^d$ correspond to data instances and rows to features, feature selection aims to find a feature subset with size $m$, which contains the most representative features. As a result, the data points represented by the selected features can well preserve the discriminative and geometrical structure as the data represented by the original $d$-dimensional features.

### A. Discriminative Feature Selection

Since label information is unavailable for unsupervised feature selection, we resort to the discriminative clustering framework [13], which is designed for seeking the most linearly separated clusters $P \in \{0,1\}^{n\times c}$ through a multi-output regularized linear regression model, to detect the cluster structure. In order to select the most discriminative features so that the separability between these clusters is maxmized, we introduce an indicator variable $\boldsymbol{z}$ to represent whether a feature is selected or not, where $\boldsymbol{z} = (z_1, \ldots, z_d)^T$ and $z_i \in \{0,1\}, i = 1, \ldots, d$. The induced regression problem can be formulated as follows:

$$\begin{aligned} \min_{P,W,\boldsymbol{b},\boldsymbol{z}} \quad & ||P - X^T \mathrm{diag}(\sqrt{\boldsymbol{z}})W - \mathbf{1}_n \boldsymbol{b}||^2 + \lambda||W||^2 \qquad (1) \\ \text{s.t.} \quad & P \in \{0,1\}^{n\times c}, P\mathbf{1}_c = \mathbf{1}_n, \boldsymbol{z} \in \{0,1\}^d, \mathbf{1}_d^T \boldsymbol{z} = m, \end{aligned}$$

where $W$ is the transform matrix, $\boldsymbol{b}$ is the bias term, and $\mathrm{diag}(\boldsymbol{z})$ is a diagonal matrix with its diagonal elements being $\boldsymbol{z}$.

Due to the combination nature of the cluster indicator matrix $P$ and the feature indicator variable $\boldsymbol{z}$, the optimization problem in Eq. (1) is NP-hard. As in [22], instead of computing the partition matrix $P$ directly, we first substitute it with a scaled partition matrix $Y = P(P^TP)^{\frac{1}{2}}$. It is easy to verify that $Y^TY = I$. We also relax the integer constraint on $z_j$ and allow $z_j$ to take real nonnegative values $z_j \geq 0, j = 1, \ldots, d$. Then, we substitute $\mathrm{diag}(\sqrt{\boldsymbol{z}})W$ with $W$. Moreover, for feature selection, it is desired that $z_j$ would be sufficiently sparse and many of them should be zeros. Though the sparseness of $\boldsymbol{z}$ can be controlled through minimizing the $\ell_1$ norm of $\boldsymbol{z}$, i.e., $||\boldsymbol{z}||_1$, it will introduce additional regularization parameter. To obtain a sparse solution on $\boldsymbol{z}$ without introducing additional free parameter, we further impose a simplex constraint on $\boldsymbol{z}$, that is, $\sum_j^d z_j = 1$. Finally, we get the following weighted $\ell_2$-norm regularized regression problem:

$$\begin{aligned} \min_{Y,W,\boldsymbol{b},\boldsymbol{z}} \quad & ||Y - X^TW - \mathbf{1}_n\boldsymbol{b}||^2 + \lambda \mathrm{tr}(W^T\mathrm{diag}(\boldsymbol{z}^{-1})W) \\ \text{s.t.} \quad & Y^TY = I, \boldsymbol{z} \geq 0, \sum_j^d z_j = 1. \qquad (2) \end{aligned}$$

The main advantage of the cost function in Eq. (2) with weighted $\ell_2$-norm regularization is that it is suitable for the task of feature selection. In other words, the value of $z_j$ indicates how significantly the $j$-feature contributes to the minimization in Eq. (2). A small value of $z_j$, which is expected to associated with an irrelevant feature, will result in a large penalization on $\{W_{jc'}\}_{c'=1}^c$.

Similar to $\ell_2$-norm regularized regression, the close form solution of $W$ and $\boldsymbol{b}$ can be obtained as follows:

$$W = (XHX^T + \lambda \mathrm{diag}(\boldsymbol{z}^{-1}))^{-1} XHY, \tag{3}$$

$$\boldsymbol{b} = \frac{1}{n}\mathbf{1}_n^T (Y - X^T W). \tag{4}$$

where $H = I - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^T$ is the centering matrix. Substituting the optimal values of $W$ and $\boldsymbol{b}$ into Eq. (2), the *discriminative feature selection* (DFS) can be formulated as the following optimization problem:

$$\begin{aligned} \min_{Y,\boldsymbol{z}} \quad & \mathrm{tr}(Y^T H(I + \frac{1}{\lambda} HX^T \mathrm{diag}(\boldsymbol{z}) XH)^{-1} HY) \\ \text{s.t.} \quad & Y^T Y = I, \boldsymbol{z} \geq 0, \sum_j^d z_j = 1. \end{aligned} \tag{5}$$

As can be seen, the objective function in Eq. (5), parameterized by $Y$ and $\boldsymbol{z}$, jointly evaluates the relevance of features and the separability of clusters. In other words, when the data is represented by these weighted features $\boldsymbol{z}$, we aim to find most linear separable clusters by minimizing the cost function of discriminative clustering. When the cluster structure $Y$ is identified, we are looking for those features that can well preserve the cluster structure.

### *B. Local Discriminative Feature Selection*

One disadvantage of the objective function in Eq. (5) is that, it can only be used to estimate the relevance of features to preserve the cluster structure that are linearly separable. It may not be easy to evaluate the importance of features to separate the data sampled from nonlinear manifold of the ambient Euclidean space into correct clusters [14]. That is to say, the intrinsic manifold structure should be considered while measuring the goodness of the clusters [23]. Recently, in order to characterize the underlying manifold structure, many manifold learning algorithms have been proposed, such as Local Linear Embedding (LLE) [24] and ISOMAP [25]. Many unsupervised feature selection algorithms [3], [6], [10] use various graphs to capture the manifold structure. However, most existing works construct graphs to approximate the manifold structure in the uniform weighted Euclidean space (the ambient space). When there are lots of irrelevant or noisy features, the neighborhood relationship defined in the original space can be completely different from the true relationship. Which means the quality of these graphs to characterize the intrinsic manifold structure cannot be guaranteed and it will mislead the results of clustering and feature selection [5]. Actually, for the task of feature selection, it is assumed that features are weighted differently (either binary or continuous weights), and formed a weighted feature space. Therefore, to evaluate the importance of features and detect the cluster structure of data, it may be more appropriate to construct a graph to capture the manifold structure which is embedded in the weighted ambient Euclidean space. The main problem to construct such graph, however, is that it is still difficult to choose the proper weights and the neighborhood relationship is unknown before learning. To handle this dilemma, we approximate the true feature weights by $\boldsymbol{z}^{\mathrm{old}}$, estimated from previous iteration. More clearly, we aim to evaluate the relevance of features $\boldsymbol{z}$ to respect the intrinsic manifold structure embedded in the ambient space weighted by $\boldsymbol{z}^{\mathrm{old}}$.

To achieve this goal, we consider the linear approximation of the local manifold structure throught local learning [22]. Concretely, we first split the whole data set into $n$ overlapped local regions $\{X_i, Y_i\}_{i=1}^n$, where the local region data matrix $X_i = [\boldsymbol{x}_i, \boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_{k-1}}]$ is consist of $\boldsymbol{x}_i$ and its $k-1$ nearest neighbors whose neighborhood is determined by $\boldsymbol{z}^{\mathrm{old}}$, and $Y_i = [\boldsymbol{y}_i, \boldsymbol{y}_{i_1}, \ldots, \boldsymbol{y}_{i_{k-1}}]^T$ is the local scaled partition matrix for the $i$-th region. Then we employ a linear discriminative feature selection model in Eq. (5) on each local region to evaluate the relevance of features for separating these local data points, and have

$$\begin{aligned} J_i(Y_i, \boldsymbol{z}) &= \mathrm{tr}(Y_i^T H_i (I + \frac{1}{\lambda_i} H_i X_i^T \mathrm{diag}(\boldsymbol{z}) X_i H_i)^{-1} H_i Y_i) \\ &= \mathrm{tr}(Y_i^T L_i Y_i), \end{aligned} \tag{6}$$

where $L_i = H_i(I + \frac{1}{\lambda_i} H_i X_i^T \mathrm{diag}(\boldsymbol{z}) X_i H_i)^{-1} H_i$. Similar to Eq. (5), Eq. (6) also selects the most discriminative features to best separate samples within the local region. Since we expect that the selected features can best preserve the cluster structure in all the local regions, we ensemble all these local models by summing Eq. (6) over all regions, and get:

$$\begin{aligned} J_{\text{local}}(Y, \boldsymbol{z}) &= \sum_{i=1}^n \mathrm{tr}(Y_i^T L_i Y_i) \\ &= \sum_{i=1}^n \mathrm{tr}(Y^T S_i L_i S_i^T Y) = \mathrm{tr}(Y^T L_l Y), \end{aligned} \tag{7}$$

where $L_l = \sum_{i=1}^n S_i L_i S_i^T$ and $S_i \in \{0,1\}^{n \times k}$ is the selection matrix for $i$-th region with its element $(S_i)_{jk'} = 1$ if $\boldsymbol{x}_j$ is selected as the $k'$ neighbor of $\boldsymbol{x}_i$; $(S_i)_{jk'} = 0$, otherwise. It can be shown that the matrix $L_l$ defined in Eq. (6) is a Laplacian matrix [14].

The objective function in Eq. (7) is also parameterized by $Y$ and $\boldsymbol{z}$. In other words, when $Y$ and $\{X_i, Y_i\}_{i=1}^n$, determined by the previous estimated weights, are given, we aim to evaluates the relevance of features to separate the data points for all local regions and respect the intrinsic manifold structure embedded in weighted ambient space. On the other hand, when the feature weights is given, we re-estimate the cluster structure based on the reconstructed graph Laplacian.

### *C. The Objective Function of LGDFS*

Compared with the aggregated local cost function $\mathcal{J}_{\text{local}}$ in Eq. (7), we denote the cost function defined on the whole data set in Eq. (5) as $\mathcal{J}_{\text{global}}$. By incorporating the local and global discriminative feature selection models in Eq. (7) and

Eq. (5) into a joint framework, our proposed Local and Global Discriminative Learning for unsupervised Feature Selection (LGDFS) method then can be formulated as the following optimization problem:

$$\begin{aligned}
&\min_{Y,\boldsymbol{z}} \quad \mathcal{J}_{\text{local}}(Y,\boldsymbol{z}) + \alpha \mathcal{J}_{\text{global}}(Y,\boldsymbol{z}) \qquad (8)\\
&= Y^T \left( \sum_{i=1}^{n} S_i [H_i (I + \frac{1}{\lambda_i} H_i X_i^T \text{diag}(\boldsymbol{z}^{-1}) X_i H_i)^{-1} H_i] S_i \right) Y \\
&+ \alpha Y^T [H(I + \frac{1}{\lambda} H X^T \text{diag}(\boldsymbol{z}^{-1}) X H)^{-1} H] Y \\
&\text{s.t.} \quad Y^T Y = I, \boldsymbol{z} \geq 0, \sum_{j}^{d} z_j = 1,
\end{aligned}$$

where $\alpha$ is a regularization parameter to trade off $\mathcal{J}_{\text{local}}$ and $\mathcal{J}_{\text{global}}$. Clearly, the first term $\mathcal{J}_{\text{local}}$ explores the local discriminative information to select those features that can best separate the data points in local regions, and the second term $\mathcal{J}_{\text{local}}$ selects the most representative features which best repeat the global cluster structure captured by $Y$. In this way, the proposed LGDFS explicitly exploits both local and global discriminative information to perform feature selection and clustering simultaneously.

Similar to discriminative feature selection, LGDFS also has a weighted $\ell_2$-norm regularized least squares regression interpretation.

**Theorem 1.** *local and global discriminative learning for unsupervised feature selection is equivalent to*

$$\begin{aligned}
&\min \quad \mathcal{J}(Y, W, \boldsymbol{b}, \{W_i, \boldsymbol{b}_i\}_{i=1}^n, \boldsymbol{z}) \qquad (9)\\
&= \sum_{i=1}^{n} \left( ||Y_i - X_i^T W_i + \mathbf{1}_{n_i} \boldsymbol{b}_i^T||^2 + \lambda_i \text{tr}(W_i^T \text{diag}(\boldsymbol{z}^{-1}) W_i) \right) \\
&+ \alpha \left( ||Y - X^T W + \mathbf{1}_n \boldsymbol{b}^T||^2 + \lambda \text{tr}(W^T \text{diag}(\boldsymbol{z}^{-1}) W) \right) \\
&\text{s.t.} \quad Y^T Y = I, \sum_{j=1}^{n} z_j = 1, z_j \geq 0, j = 1, \ldots, d.
\end{aligned}$$

The proof is similar to that of discriminative feature selection in Section 3.1 and hence we omit it here. From Theorem 1, we can find that LGDFS selects the most representative features across all local tasks defined on local regions and the global task in the whole input space.

### D. Algorithm to Solve LGDFS

In this subsection, we introduce an alternating iterative algorithm to estimate the feature weight $\boldsymbol{z}$ and the clustering result $Y$.

*1) Solving $Y$ when $\boldsymbol{z}$ is fixed:* For a given $\boldsymbol{z}$, we need to re-compute the nearest neighbors $\mathcal{N}(\boldsymbol{x}_i)$ for each point $\boldsymbol{x}_i$ according to the weighted Euclidean distance:

$$\text{Dist}_{\boldsymbol{z}}(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{\boldsymbol{x}^T \text{diag}(\boldsymbol{z}) \boldsymbol{x}'} = \sqrt{\sum_{j=1}^{d} z_j (x_j - x_j')^2}. \qquad (10)$$

With the fixed feature weight $\boldsymbol{z}$, the local graph Laplacian $L_i$ for each data point and their global integration $L_l$ can be computed using Eq. (6) and Eq. (7). The global graph Laplacian $L_g$ can also be computed using Eq. (5). Therefore, the optimal $Y$ is obtained by solving the following optimization problem:

$$\min_{Y} \quad \text{tr}(Y^T (L_l + \alpha L_g) Y), \quad \text{s.t.} \quad Y^T Y = I. \qquad (11)$$

It is clear that $L_l + \alpha L_g$ is also a Laplacian matrix, and so the problem in Eq. (11) is just one variant of spectral clustering. Thus, the global optimal solution is top eigenvectors of the combined Laplacian matrix.

*2) Solving $\boldsymbol{z}$ when $Y$ is fixed:* With the fixed $Y$, it is still difficult to solve the remained non-convex optimization problem in Eq. (8) due to the involvement of the estimation of neighborhood structure and multiple matrix inversion operations. To obtain a local optimum, we fix the neighborhood structure determined by the estimated $\boldsymbol{z}$ in previous iteration. Due to the multiple matrix inversion operations, the sequential or the convex SDP solver [15], [26] designed for single matrix inversion cannot be easily adapted. In this paper, we resort to the equivalent multi-task regression formulation in Eq. (9).

When $Y$ is given, the optimal value of $\{W_i\}_{i=1}^n$ and $W$ can be computed using Eq. (3), and the computation cost of the matrix inversion is $O(d^3)$. Following the Woodbury-Morrison formula, Eq. (3) for the computation of $W_i$ can be simplified as

$$\begin{aligned}
&W_i = \frac{1}{\lambda_i} \text{diag}(\boldsymbol{z}) X_i H_i \\
&[I_i - (\lambda_i I_i + H_i X_i^T \text{diag}(\boldsymbol{z}) X_i H_i)^{-1} H_i X_i^T \text{diag}(\boldsymbol{z}) X_i H_i] Y_i \\
&= \frac{1}{\lambda_i} \text{diag}(\boldsymbol{z}) X_i H_i [I_i - (\lambda_i I_i + K_i^z)^{-1} K_i^z] Y_i, \qquad (12)
\end{aligned}$$

in which the complexity of the matrix inversion is only $O(n_i^3)$. For high-dimensional data, we often have $n_i \ll d$ and even $n \ll d$; so it is much faster to calculate $\{W_i\}_{i=1}^n$ and $W$ in Eq. (12) in comparison to Eq. (3).

When $\{W_i\}_{i=1}^n$ and $W$ is computed, the estimation of $\boldsymbol{z}$ reduces to minimize the following objective function:

$$\mathcal{O}(\boldsymbol{z}) = \sum_{i=1}^{n} \lambda_i \text{tr}(W_i^T \text{diag}(\boldsymbol{z}^{-1}) W_i) + \alpha \lambda \text{tr}(W^T \text{diag}(\boldsymbol{z}^{-1}) W),$$

with the constraint $\sum_{j=1}^{d} z_j = 1, z_j \geq 0, \forall j$. Let $\eta$ and $\nu_j$ be the Lagrange multiplier for constraints $\sum_{j=1}^{d} z_j = 1$ and $z_j \geq 0$, respectively. Then the Lagrange $\mathcal{L}$ is

$$\mathcal{L}(\boldsymbol{z}, \eta, \nu) = \mathcal{O}(\boldsymbol{z}) + \eta (\sum_{j=1}^{d} z_j - 1) - \sum_{j=1}^{d} \nu_j z_j. \qquad (13)$$

To find the optimal solution of $\boldsymbol{z}$, we set the derivative of $\mathcal{L}$ with respect to $z_j$ to zero, that is:

$$\frac{\partial \mathcal{L}}{\partial z_j} = - \frac{\sum_i^n \lambda_i \sum_{c'=1}^{c} (W_i)_{jc'}^2 + \alpha \lambda \sum_{c'=1}^{c} W_{jc'}^2}{z_j^2} + \eta - \nu_j = 0.$$

Combining the KKT conditions $\nu_j z_j = 0, \forall j$ and the constraint $\sum_j z_j = 1$, the analytical solution of $\boldsymbol{z}$ can be computed as follows:

$$z_j = \frac{\sqrt{\sum_{i=1}^{n} \lambda_i \sum_{c'=1}^{c} (W_i)_{jc'}^2 + \alpha\lambda \sum_{c'=1}^{c} W_{jc'}^2}}{\sum_{j'=1}^{d} \sqrt{\sum_{i=1}^{n} \lambda_i \sum_{c'=1}^{c} (W_i)_{jc'}^2 + \alpha\lambda \sum_{c'=1}^{c} W_{jc'}^2}}. \tag{14}$$

The Complete algorithm is described in Algorithm 1.

**Algorithm 1** Algorithm to Solve LGDFS

**Input:** Data set $\{\boldsymbol{x}_i\}_{i=1}^n$, number of selected features $m$, number of clusters $C$, size of the neighborhood $k$, local regularization parameters $\{\lambda_i\}_{i=1}^n$, global regularization parameter $\lambda$, and trade-off parameter $\alpha$.

**Output:** $m$ selected features

1: Initialize $\boldsymbol{z} = [\frac{1}{d}, \ldots, \frac{1}{d}]^T$
2: **repeat**
3: Construct the $k$-nearest neighborhoods for each point using Eq. (10);
4: Construct the local Laplacian $L_l$ using Eq. (7);
5: Compute the global Laplacian $L_g$ using Eq. (5);
6: Solve the eigenvalue decomposition problem in Eq. (11) to obtain the optimal $Y$;
7: Compute the local and global discriminative models $\{W_i\}_{i=1}^n$ and $W$ using Eq. (12);
8: Update $\boldsymbol{z}$ using Eq. (14);
9: **until** Converges
10: Sort each feature according to the optimal $\boldsymbol{z}$ in descending order and select the top-$m$ ranked ones.

*3) Complexity Analysis:* We now analyze the computational complexity of the proposed algorithm in each iteration. In each iteration, we first need $O(n^2 d)$ to construct the $k$-nearest neighbor. Then we need $O(k^3)$ to compute the matrix $L_i$ for each data point. Thus, the total complexity to compute $L_l$ is $O(nk^3)$. Then we need $O(n^3)$ to compute the matrix $L_g$. The complexity of the eigenvalue decomposition is $O(n^3)$. The complexity of $n$ local and global regression problems is $O(nk^3 + n^3)$, The complexity for updating $\boldsymbol{z}$ is $O(ndC)$. Hence, the overall complexity for each iteration is $O(n^3)$.

## IV. Relationship to Other Methods

Our work is related to several existing approaches in the literature of machine learning and data mining. The first class of them is unsupervised feature selection methods, such as MCFS [6], JELSR [10] and UDFS [9]. These methods estimate cluster structure via spectral embedding of graph Laplacian defined in uniform feature space and perform feature selection using sparse spectral regression. To better capture the intrinsic structure of the data and alleviate the effects of irrelevant features, our method constructs the graph Laplacian through local learning in the weighted ambient space. In this way, the most representative features are naturally defined as those features that can best respect both the local and global cluster structure of the data. Compared with previous spectral feature selection approaches, our method can jointly improve the estimation of cluster structure and the weights for features.

Another set of related approaches are spectral clustering, such as the local learning based clustering, e.g. [14], [22], and clustering with local and global mixed Laplacian, e.g. [27], [28]. Obviously, these methods also construct various Laplacian matrices on uniform feature space. Our approach estimates the cluster structure in a non-uniform weighted feature space, which alleviates the side effects of the irrelevant or noisy features.

Finally, LGDFS draws the connections to multi-task feature learning, e.g., [29], [30]. In fact, The estimation of feature weight in Eq. (9) can be regarded as $n + 1$ (n local and 1 global) regression tasks with weighted $\ell_2$-norm regularization associated with the simplex constraint. In the following, we show that such regularization is the upper bound of the square of the $\ell_1$ norm regularization on the feature level. Thus, it is expected that such regularization will produce at least as sparse as that of the squared $\ell_1$ norm regularization.

**Lemma 1.** *The following inequality always holds in the input feature space:*

$$\sum_{i=1}^{n} \lambda_i \mathrm{tr}(W_i^T \mathrm{diag}(\boldsymbol{z})^{-1} W_i) + \alpha\lambda \mathrm{tr}(W^T \mathrm{diag}(\boldsymbol{z})^{-1} W)$$
$$= \sum_{j=1}^{d} \frac{||\bar{W}_j||^2}{z_j} \geq \left( \sum_{j=1}^{d} ||\bar{W}_j|| \right)^2 \tag{15}$$

*where* $||\bar{W}_j|| = \sqrt{\sum_{i=1}^{n} \lambda_i \sum_{c'=1}^{c} (W_i)_{jc'}^2 + \alpha\lambda \sum_{c'=1}^{c} W_{jc'}^2}$ *and* $\sum_j z_j = 1, z_j \geq 0$.

*Proof:* This is the corollary of Theorem 1 in [31]. ■

## V. Experimental Results

In this section, we evaluate the performance of our proposed algorithm LGDFS. Following [6], [10], we perform K-means clustering by using the selected features and compare the results of different algorithms.

### A. Data Sets

Five data sets are used in our experiments. These data sets include one UCI data set ECOLI, two preprocessed microarray data sets from [1], i.e., CARCINOMAS and LUNG, and two face image diastases, i.e., UMIST and ORL. Table I summarizes the statistics of the data sets.

Table I: Summary of data sets

| Dataset | Size | Dimensions | Classes |
|---|---|---|---|
| ECOLI | 336 | 343 | 8 |
| CARCINOMAS | 174 | 9182 | 11 |
| LUNG | 203 | 3312 | 5 |
| UMIST | 575 | 10304 | 20 |
| ORL | 400 | 10304 | 40 |

### B. Evaluation Metrics

To evaluate their performance, we compare the generated clusters with the ground truth by computing the following two performance measures.

**Clustering accuracy (ACC)**. The first performance measure is the clustering accuracy, which discovers the one-to-one relationship between clusters and classes. Given a point $\boldsymbol{x}_i$, let $p_i$ and $q_i$ be the clustering result and the ground truth label, respectively. The ACC is defined as follows:

$$\text{ACC} = \frac{1}{n}\sum_{i=1}^{n}\delta(q_i, map(p_i)), \tag{16}$$

where $n$ is the total number of samples and $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise, and $map(\cdot)$ is the permutation mapping function that maps each cluster index to a true class label. The best mapping can be found by using the Kuhn-Munkres algorithm [32]. The greater clustering accuracy means the better clustering performance.

**Normalized mutual information (NMI)**. Another evaluation metric that we adopt here is the normalized mutual information, which is widely used for determining the quality of clustering. Let $\mathcal{C}$ be the set of clusters from the ground truth and $\mathcal{C}'$ obtained from a clustering algorithm. Their mutual information $MI(\mathcal{C}, \mathcal{C}')$ is defined as follows:

$$MI(\mathcal{C}, \mathcal{C}') = \sum_{c_i \in \mathcal{C}, c'_j \in \mathcal{C}'} p(c_i, c'_j) \log \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}, \tag{17}$$

where $p(c_i)$ and $p(c'_j)$ are the probabilities that a data point arbitrarily selected from the data set belongs to the cluster $c_i$ and $c'_j$, respectively, and $p(c_i, c'_j)$ 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:

$$NMI(\mathcal{C}, \mathcal{C}') = \frac{MI(\mathcal{C}, \mathcal{C}')}{\max(H(\mathcal{C}), H(\mathcal{C}'))}, \tag{18}$$

where $H(\mathcal{C})$ and $H(\mathcal{C}')$ are the entropies of $\mathcal{C}$ and $\mathcal{C}'$, respectively. Again, a larger NMI indicates a better performance.

### C. Comparisons and Parameter Settings

We evaluate and compare the following five unsupervised feature selection approaches:

- Laplacian Score [3][1], which selects those features that can best preserve the local manifold structure. The neighborhood size is set to 5. The width of the Gaussian kernel is searched from the grid $\{4^{-3}\sigma_0, 4^{-2}\sigma_0, 4^{-1}\sigma_0, \sigma_0, 4^1\sigma_0, 4^2\sigma_0, 4^3\sigma_0\}$, where $\sigma_0$ is the mean distance between any two samples in the data set.
- EVSC [5], which selects those features with high impact on graph Laplacian's eigenvalues. The width of the Gaussian kernel is searched from the grid $\{4^{-3}\sigma_0, 4^{-2}\sigma_0, 4^{-1}\sigma_0, \sigma_0, 4^1\sigma_0, 4^2\sigma_0, 4^3\sigma_0\}$.
- MCFS [6][2], which selects the features using spectral regression with $\ell_1$-norm regularization. The neighborhood size is set to 5. The dimensionality of embedding is set to the number of clusters. The width of the Gaussian kernel is searched from the grid $\{4^{-3}\sigma_0, 4^{-2}\sigma_0, 4^{-1}\sigma_0, \sigma_0, 4^1\sigma_0, 4^2\sigma_0, 4^3\sigma_0\}$.
- JELSR [10], which selects the features using joint embedding learning and sparse regression with $\ell_{2,1}$-norm regularization. The neighborhood size is set to 5. The dimensionality of embedding is set to the number of clusters. Both the trade off parameter and the $\ell_{2,1}$-norm regularization parameter are searched from the grid $\{0.01, 0.1, 1, 10, 100\}$.
- Our proposed LGDFS algorithm. The neighborhood size is set to 5. The dimensionality of embedding is set to the number of clusters. The trade off parameter $\alpha$, the weighted $\ell_2$-norm regularization parameter $\lambda$, and the local regularization parameters ($\{\lambda_i\}$ are set to be the same value) are searched from $\{0.01, 0.1, 1, 10, 100\}$.

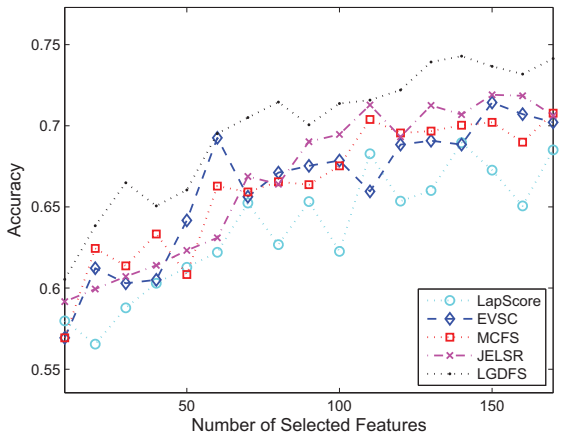
LapScore and EVSC perform feature selection by computing specified evaluation criteria. MCFS, JELSR, and LGDFS are wrapper methods using sparse regression. For these compared unsupervised feature selection algorithms, it is difficult to select the optimal parameters since there is no label information available. Therefore, we search the parameters over the grid and report the best result it can achieve. Though such strategy could be biased, it is still a fair comparison since we did this for all the methods as long as they have parameters to tune. Besides, how to decide the optimal number of selected features is data dependent and still an open problem, here we report the results over a relative large range of features.

### D. Clustering Results

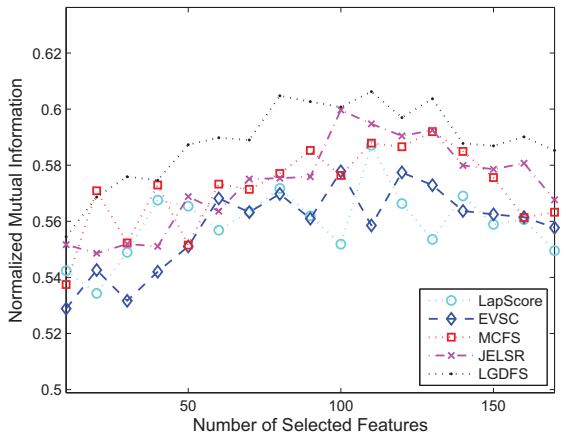
We first evaluate the clustering performance on the entire data sets. In this experiment, the K-means algorithm is applied 10 times with random initialization and the average result is reported. Figure 1 shows the plots of clustering

[1]The MATLAB source code: http://www.zjucadcg.cn/dengcai/Data/code/LaplacianScore.m

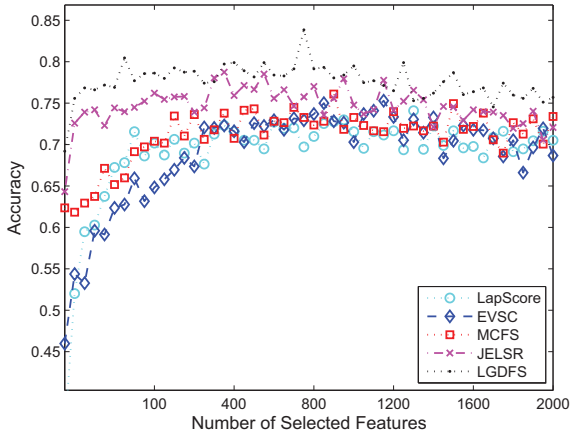
[2]The MATLAB source code: http://www.zjucadcg.cn/dengcai/Data/code/MCFS_p.m



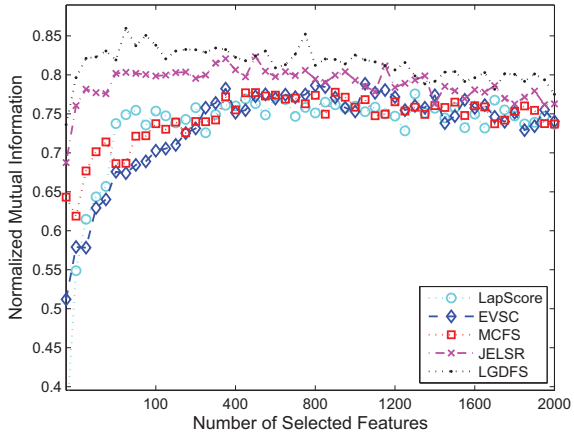

(a) ECOLI



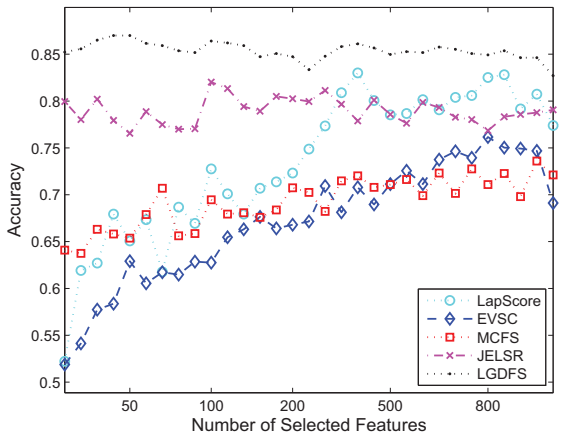

(b) ECOLI



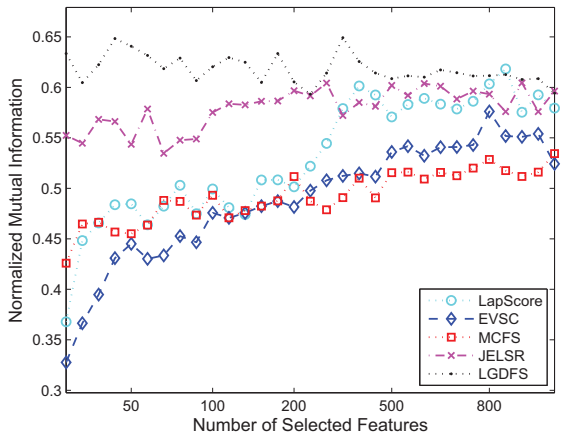

(c) CARCINOMAS




(d) CARCINOMAS

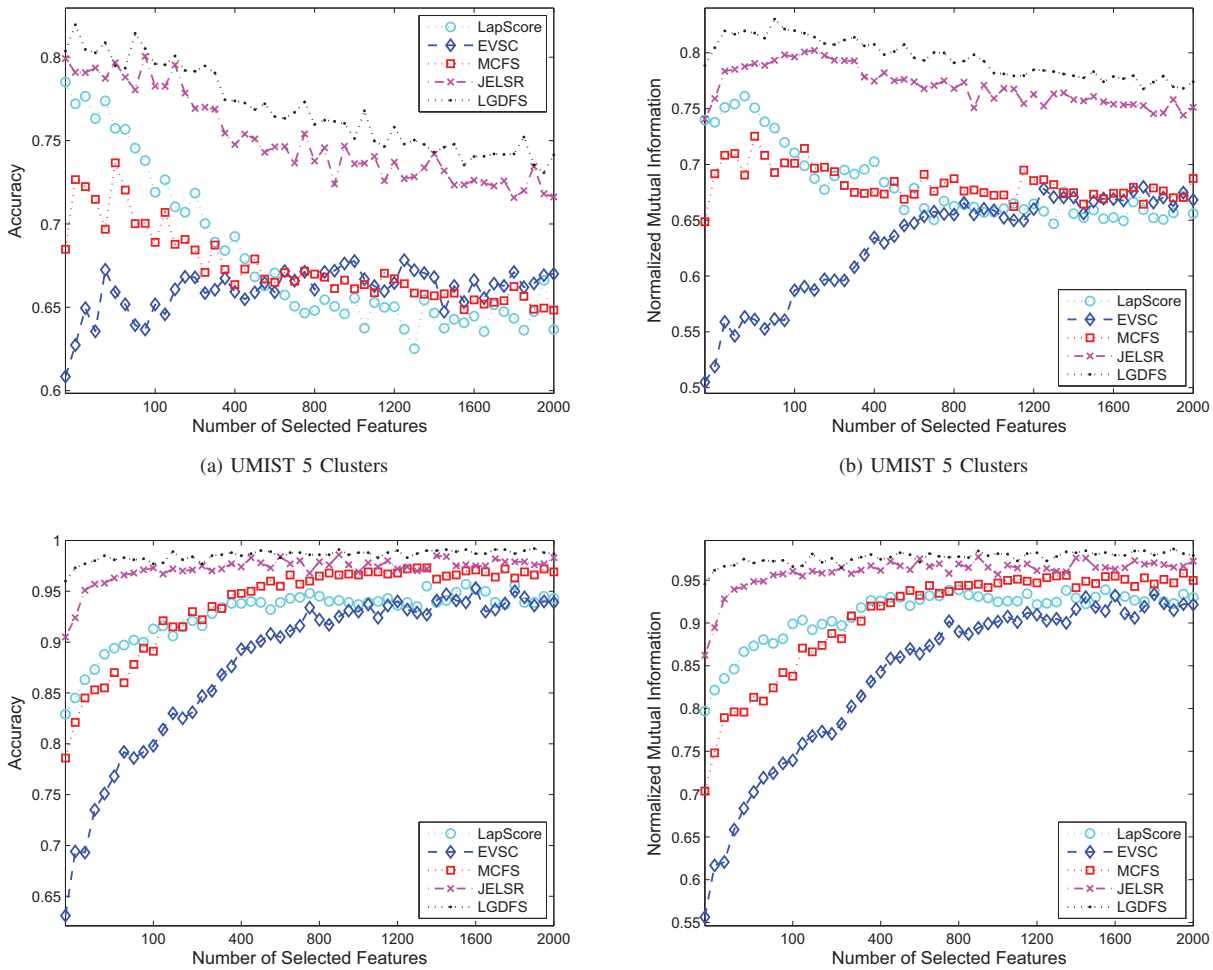



(e) LUNG




(f) LUNG

Figure 1: Clustering accuracy and normalized mutual information versus the number of selected features




(a) UMIST 5 Clusters

(b) UMIST 5 Clusters

(c) ORL 5 Clusters

(d) ORL 5 Clusters

Figure 2: Clustering accuracy and normalized mutual information versus the number of selected features on the UMIST and ORL data sets

performance, in terms of accuracy and normalized mutual information, versus the number of selected features.

In order to obtain stable results, similar to [6], we also report the clustering performance on the subsets of UMist and ORL with 5 clusters. For each data set, 20 tests are conducted on different randomly selected clusters, and the average performance over these 20 tests is reported. In each test, the K-means algorithm is applied 10 times with random initialization and the best result in terms of the objective function of K-means was recorded. Figure 2 shows the plots of clustering performance versus the number of selected features.

It can be seen that our LGDFS algorithm consistently outperforms the other four algorithms, and JELSR performs the second best in most of the cases. For example, when 100 features are selected, Compared with the second best algorithm, our method achieves 6.0% (3.3%), 6.8% (7.1%), 13.6% (17.9%), 2.7% (4.1%), and 2.8% (3.4%) relative improvement in clustering accuracy (normalized mutual information) on the ECOLI, CARCINOMAS, LUNG, x, UMIST and ORL data sets, respectively. This indicates that the joint embedding learning and sparse regression framework is generally capable of enhancing both clustering and feature selection. More importantly, the re-estimation of local structure in non-uniform weighted feature space and the exploration of local discriminative information can further improve the performance significantly. Although our algorithm performs the best in the entire scope, it is worthwhile to note that it performs especially well when there is limited number of features. Further, we observe that

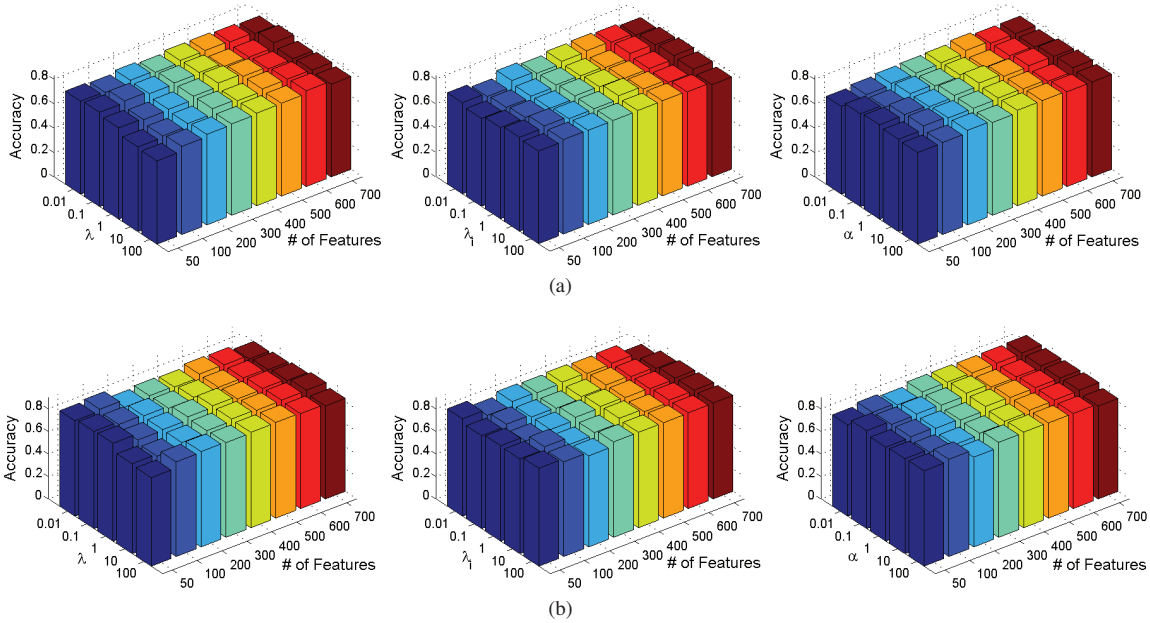



Figure 3: Clustering accuracy w.r.t. different parameters on CARCINOMAS (a) and LUNG (b).

the results of LapScore, EVSC, and MCFS on some high-dimensional data sets (e.g., UMIST) are less comparable. A plausible reason is that the gaussian kernel used in these algorithms did not capture the similarity well for very high-dimensional data (around 10K). This in turn means that it is necessary to remove the irrelevant or noisy features in the construction of graph Laplacian, which is one key contribution of our work.

### *E. Sensitivity to the Selection of Parameters*

Compared with other spectral feature selection algorithms, Our algorithm have three additional parameters, that are, the regularization parameters ($\lambda$ and $\lambda_i$) and the trade-off parameter ($\alpha$). In this section, we evaluate how the performance of LGDFS varies with different values of the parameters. The data set used for this test is CARCINOMAS and LUNG. Figure 3 shows the average clustering accuracy over selected features as a function of each of these three parameters. As we can see, the performance is not very sensitive to these parameters on different number of selected features.

## VI. Conclusion and Future Work

This paper presents a novel unsupervised feature selection algorithm, called LGDFS. It integrates a global linear regression model with weighted $\ell_2$-norm regularization and a set of locally linear ones through local learning into a unified learning framework. LGDFS characterizes the discriminative and geometrical information in a weighted Euclidean space, which alleviates the effects of irrelevant features. Thus, it can better estimate the cluster structure of the data. LGDFS evaluates the relevance of features through integrating both local and global learning. As a result, it can select the most representative features to best respect both the local and global cluster structure. Compared with state-of-the-art methods, namely LapScore, EVSC, MCFS, and JELSR, the experimental results validate that the new method achieves significantly higher performance for clustering.

In the future, we plan to continue this work on several issues as follows. First, as the computational complexity of the proposed method scales with the number of data points, we want to investigate the sampling techniques to speed up the computation. We may apply clustering techniques such as K-means to group the data points into clusters and select some representative points from each cluster. Our method is then applied only to the representative points. Second, we will try to derive a good and stable automatic parameter selection procedure for the regularization and trade-off parameters. Third, instead of using $k$-nearest neighbors measured on weighted feature space, we plan to determine it on the projection space.

## VII. Acknowledgments

This work is supported in part by NSFC grant 60970045 and China National 973 project 2013CB329305.

## References

[1] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," *NIPS*, vol. 23, pp. 1813–1821, 2010.

[2] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1610–1626, 2010.

[3] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *NIPS*, vol. 18, pp. 507–514, 2006.

[4] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th International Conference on Machine learning*, 2007, pp. 1151–1157.

[5] Y. Jiang and J. Ren, "Eigenvalue sensitive feature selection," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 89–96.

[6] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and data mining*. ACM, 2010, pp. 333–342.

[7] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010, pp. 673–678.

[8] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1294–1299.

[9] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{21}$-norm regularized discriminative feature selection for unsupervised learning," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1589–1594.

[10] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1324–1329.

[11] L. Du and Y.-D. Shen, "Joint clustering and feature selection," in *Web-Age Information Management*. Springer, 2013, pp. 241–252.

[12] M. Qian and C. Zhai, "Robust unsupervised feature selection," pp. 1621–1627, 2013.

[13] F. Bach and Z. Harchaoui, "Diffrac: a discriminative and flexible framework for clustering," *NIPS*, vol. 20, 2007.

[14] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2761–2773, 2010.

[15] L. Zhang, C. Chen, J. Bu, Z. Chen, S. Tan, and X. He, "Discriminative codeword selection for image representation," in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 173–182.

[16] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1, pp. 23–69, 2003.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[20] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," *NIPS*, vol. 23, pp. 1813–1821, 2010.

[21] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," *The Journal of Machine Learning Research*, vol. 6, pp. 1855–1887, 2005.

[22] M. Wu and B. Schölkopf, "A local learning approach for clustering," *NIPS*, vol. 19, pp. 1529–1536, 2007.

[23] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *NIPS*, vol. 2, pp. 849–856, 2002.

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

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

[26] X. He, M. Ji, C. Zhang, and H. Bao, "A variance minimization criterion to feature selection using laplacian regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 99, pp. 2013–2025, 2011.

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

[28] F. Nie, D. Xu, I. Tsang, and C. Zhang, "Spectral embedded clustering," in *Proceedings of the 21st International Joint Conference on Artifical intelligence*, 2009, pp. 1181–1186.

[29] A. Evgeniou and M. Pontil, "Multi-task feature learning," in *NIPS*, vol. 19, 2007, pp. 41–48.

[30] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 339–348.

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

[32] L. Lovász and M. Plummer, *Matching theory*, 1986, no. 121.