Self-Tuned Discrimination-Aware Method for Unsupervised Feature Selection

Xuelong Li, *Fellow, IEEE,* Mulin Chen, Qi Wang, *Senior Member, IEEE*

***Abstract*—Unsupervised feature selection is fundamentally important for processing unlabelled high-dimensional data, and several methods have been proposed on this topic. Most existing embedded unsupervised methods just emphasize the data structure in the input space, which may contain large noise. So they are limited to perceive the discriminative information implied within the low dimensional manifold. Besides, these methods always involve several parameters to be tuned, which is time-consuming. In this research, we present a Self-Tuned Discrimination-Aware (STDA) approach for unsupervised feature selection. The main contributions of this study are threefold: (1) it adopts the advantage of discriminant analysis technique to select the valuable features; (2) it learns the local data structure adaptively in the discriminative subspace to alleviate the effect of data noise; (3) it performs feature selection and clustering simultaneously with an efficient optimization strategy, and saves the additional efforts to tune parameters. Experimental results on a toy dataset and various real-world benchmarks justify the effectiveness of STDA on both feature selection and data clustering, and demonstrate its promising performance against the state-of-the-arts.**

***Index Terms*—Feature selection, graph learning, clustering, unsupervised learning, discriminant analysis**

I. Introduction

In many tasks of artificial intelligence, such as face recognition and image classification [1, 2], data is always characterized by high-dimensional features [3, 4]. However, the growth of dimensionality brings lots of noises and significantly increases the computational costs to process the data. To deal with this problem, plenty of algorithms have been developed to reduce the dimensionality of input data. Generally speaking, there are mainly two categories of dimensionality reduction methods in the literature: feature selection [5–7] and feature learning [8, 9]. Feature selection learns the most relevant feature subset for a compact representation, while feature learning methods project the features into a low dimensional subspace and creates new features. In comparison with feature learning approaches, feature selection methods have the advantage on retaining the original data representation [10]. Thus, feature selection has received a surge of interests in the past decades.

Based on the availability of labelled data, feature selection methods can be roughly grouped into supervised, semi-supervised and unsupervised feature selection. In real-world tasks, labels are expensive, which makes unsupervised feature selection especially practical. In addition, unsupervised feature selection can be further classified into filter [7, 11–15], wrapper [16–18] and embedded [19–21] methods. Filter methods select the features by examining their intrinsic properties and assigning each feature a score according to its ability to preserve the data structure. These methods consider each feature independently. Wrap methods generally solve a searching problem. They take clustering methods (e.g., GMM [22]) as the predictor and search the feature subset that maximizes the predictor performance. So they exactly remove the irrelevant features recursively. Embedded methods first learn a projection matrix $\mathbf{W}$ during the model learning procedure, and then select the features according to the transformation score $||\mathbf{W}_f||_2$ ($f = 1, ..., d$). Among them, embedded methods outperform the other two categories in many cases and have received increasing attentions [10, 23–25]. In this work, we mainly focus on the embedded algorithms.

From the perspective of manifold learning, it is important to discover the geometry structure lying within the high-dimensional data [3]. Thus, many embedded unsupervised methods [10, 26–29] are proposed to exploit the intrinsic local data structure. However, most of them just focus on the data graph in the input space, which is easily affected by noise. So it is necessary to investigate the data relationship in the low dimensional subspace, where the noise is alleviated. Additionally, the exploration of discriminant information is still not well-solved. Although some methods [24, 30–32] utilize Linear Discriminant Analysis (LDA) to select features, they inherit the limitations of traditional LDA method, such as the suboptimal solution and the neglect of local manifold. So it is important to adopt the merits of LDA while avoiding its drawbacks.

In this paper, we present a Self-Tuned Discrimination-Aware (STDA) method, which performs data graph learning and discriminant analysis simultaneously without any parameter to be tuned. The contributions of this research are summarized as follows:

(1) A Self-Tuned Discrimination-Aware (STDA) method is proposed, which reasonably incorporates discriminant analysis strategy into the unsupervised feature selection framework.

(2) The data relationship is adaptively learned in the desired discriminative subspace, so the noise in the input data space is alleviated.

(3) A rank constraint is introduced on the local relationship of data points, which ensures that both feature selection and clustering can be performed simultaneously. Moreover, there is no parameter to be tuned manually in the proposed method.

The rest of this paper is organized as follows. Section II re-

X. Li, M. Chen and Q. Wang are with the School of Computer Science and the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mail: xuelong_li@nwpu.edu.cn, chenmulin@mail.nwpu.edu.cn, crabwq@gmail.com.

visits the existing methods on unsupervised embedded feature selection. Section III introduces the proposed approach, and an alternative optimization strategy is also designed to solve it. Section IV conducts extensive experiments on various kinds of datasets and discusses the results from several aspects. The conclusions are given in Section V.

II. Related Works

In this section, some existing techniques on unsupervised embedded feature selection are reviewed.

Among the numerous unsupervised feature selection approaches, spectral-based algorithms have shown dominant performance in the past few years. Cai et al. [26] and Zhao et al. [33] captured the local data structure by spectral analysis, and ranked each feature with different sparse regression constraints. Nie et al. [28] developed a unified framework for feature selection. Qian et al. [27] employed Nonnegative Matrix Factorization (NMF) [34] to perform feature selection. Shi et al. [29] combined the local structure learning method with a sparse spectral regression strategy. Hou et al. [10] jointly performed feature selection and local structure learning. Although the performance of the above approaches are prominent in many occasions, they share the same drawback that the local manifold structure is learned in the input space, so they are easily affected by data noise. Recently, Nie et al. [23] remedied this problem by learning the similarity graph in the subspace, but they just pull the within-class samples together and neglect the between-class distance.

In order to select the discriminative features, some methods incorporate linear discriminant analysis into the framework of feature selection. Zhang et al. [30] performed multi-modal discriminative learning to capture the valuable features while exploring the local data relationship. Tao et al. [31] presented a Discriminative Feature Selection (DFS) method with a $\ell_{2,1}$ norm regularization. But these method require the data label as the input. To perform unsupervised discriminant analysis, Yang et al. [24] introduced pseudo labels to investigate the discriminant information. But this method inherits the shortcoming of Linear Discriminant Analysis (LDA) that the local data structure cannot be captured. To address this problem, Li et al. [25] and Tang et al. [32] first found the k neighbors of each point, and then performed discriminant analysis to select features. But the neighbor relationship in the input data space may be unreliable. It is also impractical to select an appropriate k for various applications. Moreover, these discriminative feature selection methods represent the LDA objective with a ratio trace form, which leads to the suboptimal solution.

In addition, all the above methods involve several parameters to be tuned, and the optimal parameters vary on different datasets, so they are not so practical on real-world applications.

III. Self-Tuned Discrimination-Aware Feature Selection

In this section, the Self-Tuned Discrimination-Aware (STDA) method is described. First, the Linear Discriminant Analysis (LDA) is introduced as the preliminary. Then, the objective function of STDA is presented and theoretically analyzed. Finally, an efficient optimization strategy is followed.

A. Preliminary

Here we revisit the classical LDA [35], and derive it to a variant. Given the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ with c classes, LDA aims to find a transformation matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ $(m \ll d)$ to project $\mathbf{x}_j$ into a m-dimensional representation: $\mathbf{y}_j = \mathbf{W}^T\mathbf{x}_j$ (T indicates the transpose operation). LDA assumes that the optimal $\mathbf{W}$ should pull the data points within the same class closer while pushing those from different classes far away, so the objective function is

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\sum_{i=1}^{c}\sum_{j=1}^{n_i}||\mathbf{W}^T(\mathbf{x}_j^i - \mu^i)||_2^2}{\sum_{i=1}^{c} n_i||\mathbf{W}^T(\mu^i - \mu)||_2^2}, \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix, n_i is the points number within class i, μ_i is the mean of points in class i, μ is the mean of all the points, x_j^i is the j-th point in class i. c is always given as a prior in machine learning. The orthogonal constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ avoids the arbitrary scaling and trivial solution. From problem 1, we can see that LDA just emphasizes the global data relationship, and neglects the local manifold structure.

According to Li et al. [36], problem (1) is equivalent to the following problem

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j,k=1}^{n_i}||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)||_2^2}{\frac{1}{n}\sum_{j,k=1}^{n}||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}. \tag{2}$$

Denote an indicator matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$, where $\mathbf{Z}_{jk}$ is 1 if j and k belong to the same class and $\mathbf{Z}_{jk}$ is 0 otherwise. Supposing that the number of points is equal for each class, problem (2) can be transformed into the following form:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\sum_{j,k=1}^{n}\mathbf{Z}_{jk}||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}{\sum_{j,k=1}^{n}||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}, \tag{3}$$

where n counts the number of points in all classes.

B. Problem Formulation

In this part, the proposed STDA is introduced. Since LDA has the capability to find the discriminative data subspace, we would like to adopt it into unsupervised feature selection scheme. However, LDA requires the data to be labelled, so it cannot be used directly unless the indicator matrix $\mathbf{Z}$ in problem (3) is given.

We consider that the data points belonging to the same category should have high similarity in the transformed subspace, so the matrix $\mathbf{Z}$ in problem (3) can be replaced with an affinity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, where S_{jk} is close to 1 if $||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)||_2^2$ is small and 0 otherwise. Then we need learn both the linear

transformation $\mathbf{W}$ and the affinity graph $\mathbf{S}$ with the following objective function

$$\min_{\mathbf{W},\mathbf{S}} \frac{\sum_{j,k=1}^{n} \mathbf{S}_{jk}^2 ||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}{\sum_{j,k=1}^{n} ||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}, \quad s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_k \mathbf{S}_{jk} = 1, \mathbf{S} \geq 0 \tag{4}$$

where the definitions are the same as those in problem (3). $\mathbf{S}_{jk}$ is squared to avoid the trivial solution, where $\mathbf{S}$ is 1 for the nearest points and 0 for the others. Nevertheless, the above objective cannot guarantee that the similarity of far away points is exact 0. Defining the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ as $\mathbf{D} - \frac{\mathbf{S}^T+\mathbf{S}}{2}$ ($\mathbf{D} \in \mathbb{R}^{n\times n}$ is the degree matrix), Mohar et al. [37] have proved that the data graph $\mathbf{S}$ will contain exact c connected components if the rank of $\mathbf{L}$ is $n - c$. If the above rank constraint is imposed on problem (4), the final $\mathbf{S}$ will contain c components, and the similarity for the far away points will be zero. So the objective function of STDA is

$$\min_{\mathbf{W},\mathbf{S}} \frac{\sum_{j,k=1}^{n} \mathbf{S}_{jk}^2 ||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}{\sum_{j,k=1}^{n} ||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}, \quad s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_k \mathbf{S}_{jk} = 1, \mathbf{S} \geq 0, rank(\mathbf{L}) = n - c, \tag{5}$$

where $rank()$ denotes the rank of a matrix.

Problem (5) is difficult to solve because the rank constraint depends on $\mathbf{S}$. So we transform the constraint into a different form. Following the proof in [38], denoting the i-the smallest eigenvalue of $\mathbf{L}$ as $\delta_i(\mathbf{L})$, then $rank(\mathbf{L}) = n - c$ is equivalent to enforcing $\sum_{j=1}^{c} \delta_i(\mathbf{L})$ to be zero. Furthermore, according to Ky Fan's Theorem [39], we get

$$\sum_{j=1}^{c} \delta_i(\mathbf{L}) = \min_{\mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F}=\mathbf{I}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), \tag{6}$$

where $\mathrm{Tr}()$ is the trace operator. Thus, problem (5) can be rewritten as

$$\min_{\mathbf{W},\mathbf{S},\mathbf{F}} \frac{\sum_{j,k=1}^{n} \mathbf{S}_{jk}^2 ||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2}{\sum_{j,k=1}^{n} ||\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)||_2^2} + \lambda \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), \quad s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_k \mathbf{S}_{jk} = 1, \mathbf{S} \geq 0, \mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \tag{7}$$

where λ is a parameter.

In problem (7), it can be seen that $\mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})$ will be infinitely close to zero if λ is large enough, then the rank constraint can be satisfied. The optimal $\mathbf{S}$ contains exact c components. Given the transformation matrix $\mathbf{W}$, $\mathbf{S}$ will be large for the points with small distance in the transformed space, so the local connectivity can be perceived. Then $\mathbf{W}$ can be updated with the learned data graph $\mathbf{S}$. Thus, the local manifold structure in the discriminative subspace can be exploited by optimizing $\mathbf{S}$ and $\mathbf{W}$ iteratively. With the optimal $\mathbf{W}$, we can calculate the score for each feature f ($f = 1, ..., d$) as $||\mathbf{W}_f||_2$, and select the features with large scores. Unlike existing works [24, 25, 30–32], which transform the objective function into the ratio-trace form, STDA performs discriminant analysis with the trace-ratio form, so the suboptimal solution can be avoided.

In addition, since $\mathbf{S}$ contains c connected components, it can be considered as an indicator matrix, where each component corresponds to a cluster. Therefore, the points can be partitioned into clusters once the optimal $\mathbf{S}$ is learned. Note that, the value of λ can be tuned in a heuristic way [38] automatically in each iteration. Specifically, If the number of zero eigenvalues in $\mathbf{L}$ is larger than c, which indicates that the number of the connected components of $\mathbf{S}$ is more than the desired cluster number, we decrease λ; otherwise we increase it. So the proposed method is totally self-tuned. This property is promising because the tuning of parameters is the most time-consuming part in the practical applications of feature selection methods. Similar to the most feature selection and clustering methods [24–26, 29, 34, 38, 40], the proposed method needs the class number c as the input, and the automatically estimating of c is not the focus of this research.

However, when transforming problem (2) to (3), the data is assumed to be with a balanced distribution over each class, which is not true in many tasks. Here we briefly discuss this confusion. In the proposed objective (7), we constrain the sum of each row of $\mathbf{S}$ to be 1, which effectively weights each class equally regardless of the point number within the class. Therefore, our method is able to deal with unbalanced data, and the experimental demonstration will be given in the experiments.

C. Optimization of STDA Algorithm

Problem (7) contains several different variables to be optimized, so we put forward an alternative algorithm to get the optimal solution. The similarity graph $\mathbf{S}$ is initialized with an efficient method [38].

Update F: when fixing $\mathbf{S}$ and $\mathbf{W}$, problem (7) becomes

$$\min_{\mathbf{F}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), \quad s.t. \mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}. \tag{8}$$

Denoting the j-th column of $\mathbf{F}$ as $\mathbf{F}_j$, because $\mathbf{F}$ is orthogonal, the minimum value of $\mathbf{F}_j^T\mathbf{L}\mathbf{F}_j$ equals to the smallest eigenvalue of $\mathbf{L}$. Thus, the optimal $\mathbf{F}$ is constructed with the c eigenvectors of $\mathbf{L}$, which correspond to the c smallest eigenvalues.

Update W: when fixing $\mathbf{S}$ and $\mathbf{F}$, denoting matrix $\tilde{\mathbf{S}}_w \in \mathbb{R}^{d\times d}$ and $\tilde{\mathbf{S}}_t \in \mathbb{R}^{d\times d}$ as

$$\tilde{\mathbf{S}}_w = \sum_i \sum_j \mathbf{S}_{jk}^2 (\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^T, \quad \tilde{\mathbf{S}}_t = \sum_i \sum_j (\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^T, \tag{9}$$

then problem (7) becomes a trace ratio problem:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\mathrm{Tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^T\tilde{\mathbf{S}}_t\mathbf{W})}. \tag{10}$$

The optimal $\mathbf{W}$ can be leaned with the optimization method in [13], which derives the trace ratio problem into a monotonically decreasing function, and then calculates the global minimum solution by binary searching.

Update S: when fixing $\mathbf{F}$ and $\mathbf{W}$, according to the spectral analysis theory [29], we can transform problem (7) into

$$\min_{\sum_k \mathbf{S}_{jk}=1,\mathbf{S}\geq 0} \frac{\sum_{j,k=1}^{n} \mathbf{S}_{jk}^2||\mathbf{W}^T(\mathbf{x}_j-\mathbf{x}_k)||_2^2}{\sum_{j,k=1}^{n}||\mathbf{W}^T(\mathbf{x}_j-\mathbf{x}_k)||_2^2}+ \lambda\sum_{j,k=1}^{n}\mathbf{S}_{jk}||\mathbf{F}_j-\mathbf{F}_k||_2^2. \tag{11}$$

It can be seen that the above problem is independent for different j. Thus, denoting a column vector $\mathbf{s}_j \in \mathbb{R}^{n\times 1}$ with its k-th element equal to $\mathbf{S}_{jk}$, we can optimize the following problem separately for each j

$$\min_{\mathbf{s}_j^T\mathbf{1}=1,\mathbf{s}_j\geq 0} \frac{\sum_{k=1}^{n} \mathbf{s}_{jk}^2||\mathbf{W}^T(\mathbf{x}_j-\mathbf{x}_k)||_2^2}{\sum_{j,k=1}^{n}||\mathbf{W}^T(\mathbf{x}_j-\mathbf{x}_k)||_2^2}+ \lambda\sum_{k=1}^{n}\mathbf{s}_{jk}||\mathbf{F}_j-\mathbf{F}_k||_2^2, \tag{12}$$

where $\mathbf{1} \in \mathbb{R}^{n\times 1}$ is a vector with all the elements as 1. For a clear representation, we define a constant variable a as $\sum_{j,k=1}^{n}||\mathbf{W}^T(\mathbf{x}_j-\mathbf{x}_k)||_2^2$ and a vector $\mathbf{b}_j$ with $\mathbf{b}_{jk}=||\mathbf{F}_j-\mathbf{F}_k||_2^2$. Denoting a vector $\mathbf{u}_j \in \mathbb{R}^{n\times 1}$ with $\mathbf{u}_{jk}=\frac{1}{2}\lambda a\mathbf{b}_{jk}$ and a diagonal matrix $\mathbf{V}\in\mathbb{R}^{n\times n}$ with the k-th diagonal element equal to $||\mathbf{W}^T(\mathbf{x}_j-\mathbf{x}_k)||_2^2$, then problem (12) can be simplified to

$$\min_{\mathbf{s}_j^T\mathbf{1}=1,\mathbf{s}_j\geq 0} \frac{1}{2}\mathbf{s}_j^T\mathbf{V}\mathbf{s}_j+\mathbf{s}_j^T\mathbf{u}_j. \tag{13}$$

The Lagrangian function of problem (13) is

$$\mathcal{L}(\mathbf{s}_j,\eta,\beta_j)=\frac{1}{2}\mathbf{s}_j^T\mathbf{V}\mathbf{s}_j+\mathbf{s}_j^T\mathbf{u}_j-\eta(\mathbf{s}_j^T\mathbf{1}-1)-\beta_j^T\mathbf{s}_j, \tag{14}$$

where $\eta \in \mathbb{R}^{1\times 1}$ and $\beta_j \in \mathbb{R}^{n\times 1}$ are the Lagrangian Multipliers. Setting the derivative of Eq. (14) w.r.t. $\mathbf{s}_j$ to 0, we have

$$\mathbf{V}\mathbf{s}_j+\mathbf{u}_j-\eta\mathbf{1}-\beta_j=0. \tag{15}$$

For the k-th element of $\mathbf{s}_j$, we have

$$\mathbf{V}_{kk}\mathbf{s}_{jk}+\mathbf{u}_{jk}-\eta-\beta_{jk}=0. \tag{16}$$

According to the KKT condition, $\mathbf{s}_{jk}\beta_{jk}$ is equal to 0, so we have

$$\mathbf{s}_{jk}=\max(\frac{\eta}{\mathbf{V}_{kk}}-\frac{\mathbf{u}_{jk}}{\mathbf{V}_{kk}},0). \tag{17}$$

According to [38], we can define a function $g_j(\eta)$ w.r.t. η as

$$g_j(\eta)=-1+\sum_i(\frac{\eta}{\mathbf{V}_{kk}}+\frac{\mathbf{u}_{jk}}{\mathbf{V}_{kk}})_+, \tag{18}$$

together with the constraint $\mathbf{s}_j^T\mathbf{1}=1$, we have

$$g_j(\eta)=0. \tag{19}$$

$g_j(\eta)$ is a monotonically increasing linear function. Using the Newton's method, we can easily get the optimal η such that $g_j(\eta)$ is 0. Once η is obtained, the optimal $\mathbf{s}_j$ can be calculated with Eq. (17).

The algorithm for solving problem (7) is outlined in Algorithm 1. Problem (7) is decomposed in to three sub-problems. When solving $\mathbf{F}$, the optimal solution is searched. When solving $\mathbf{W}$, the global minimum solution is obtained. When solving $\mathbf{S}$, the global optimal solution is achieved according to the KKT condition. The original problem is equivalent to the sub-problems when the irrelevant variables are fixed. So the objective value decreases during the optimization of each sub-problems, and finally reaches to a local optimal value. The convergence behavior will be proved experimentally in Section (IV-C).

Algorithm 1 Optimization algorithm of STDA

Input: Data matrix $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n]\in\mathbb{R}^{d\times n}$, projection dimension m.

1: Initialize data graph $\mathbf{S}$.
2: **repeat**
3: Update $\mathbf{F}$ by solving problem (8).
4: Update $\mathbf{W}$ by solving problem (10).
5: Update $\mathbf{S}$ by solving problem (13).
6: **until** Converge

Output: The optimal $\mathbf{F}$, $\mathbf{W}$ and $\mathbf{S}$.

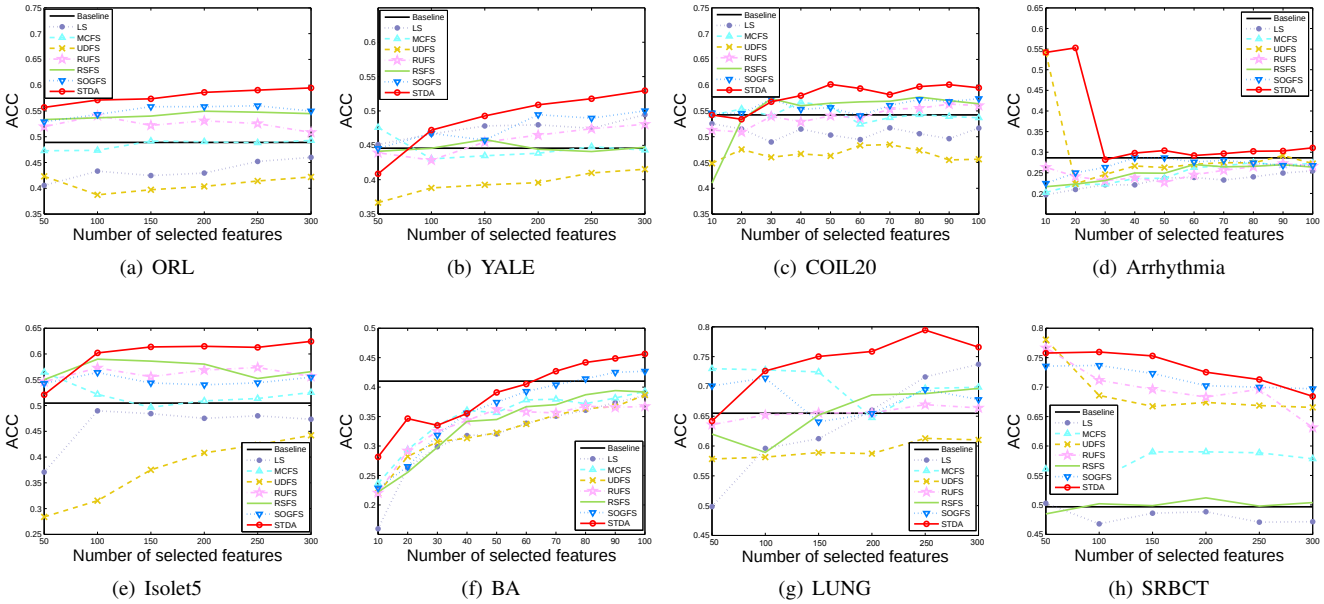
IV. Experiments

In this section, we conduct extensive experiments on various datasets to verify the effectiveness of the proposed STDA. First, several real-world datasets and one toy dataset are utilized to justify the performance on feature selection. Then, the clustering performance of STDA is evaluated.

A. Performance on Feature Selection

We first conduct feature selection experiments on real-world datasets. A standard evaluation scheme is to compare the clustering results with the features selected by different techniques. We perform k-means for 50 repetitions with random initialization and report the average results. Clustering ACCuracy (ACC) [41] and Normalized Mutual Information (NMI) [42] are taken as the evaluation metrics, which are widely used in clustering tasks.

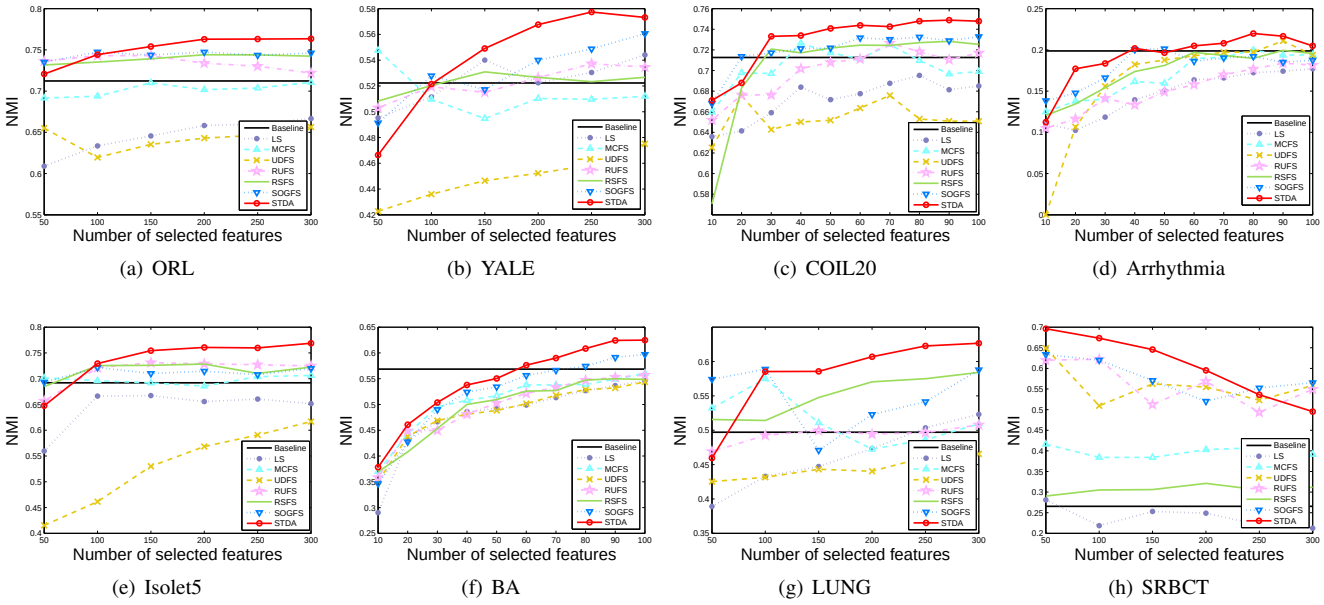
Datasets: The proposed STDA is evaluated on eight publicly available datasets: ORL [26], YALE [43], COIL20 [44], Arrhythmia and Isolet5 [45], Binary Alphabet (BA) [46], LUNG [47], and SRBCT [48]. Among them, ORL and YALE are face image datasets, COIL20 is an object image dataset,



(a) ORL (b) YALE (c) COIL20 (d) Arrhythmia

(e) Isolet5 (f) BA (g) LUNG (h) SRBCT

Fig. 1. Curves of ACC with different numbers of selected features on eight real-world datasets.



(a) ORL (b) YALE (c) COIL20 (d) Arrhythmia

(e) Isolet5 (f) BA (g) LUNG (h) SRBCT

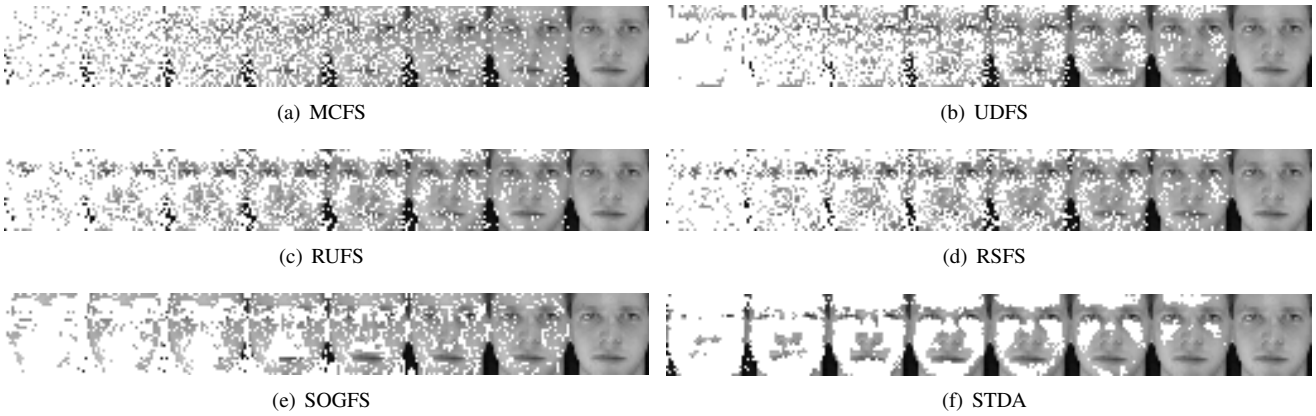
Fig. 2. Curves of NMI with different numbers of selected features on eight real-world datasets.

TABLE I
DATASET DESCRIPTION.

Dataset	Size	Feature	Classes	Selected features	IR
ORL	400	1024	40	[50,100,...,300]	1
YALE	165	1024	15	[50,100,...,300]	1
COIL20	1440	256	20	[10,20,30,...,100]	1
Arrhythmia	452	279	13	[10,20,30,...,100]	122.5
Isolet5	1560	617	26	[50,100,...,300]	1
BA	1404	320	36	[10,20,30,...,100]	1
LUNG	203	3312	5	[50,100,...,300]	23.2
SRBCT	83	2308	4	[50,100,...,300]	2.64

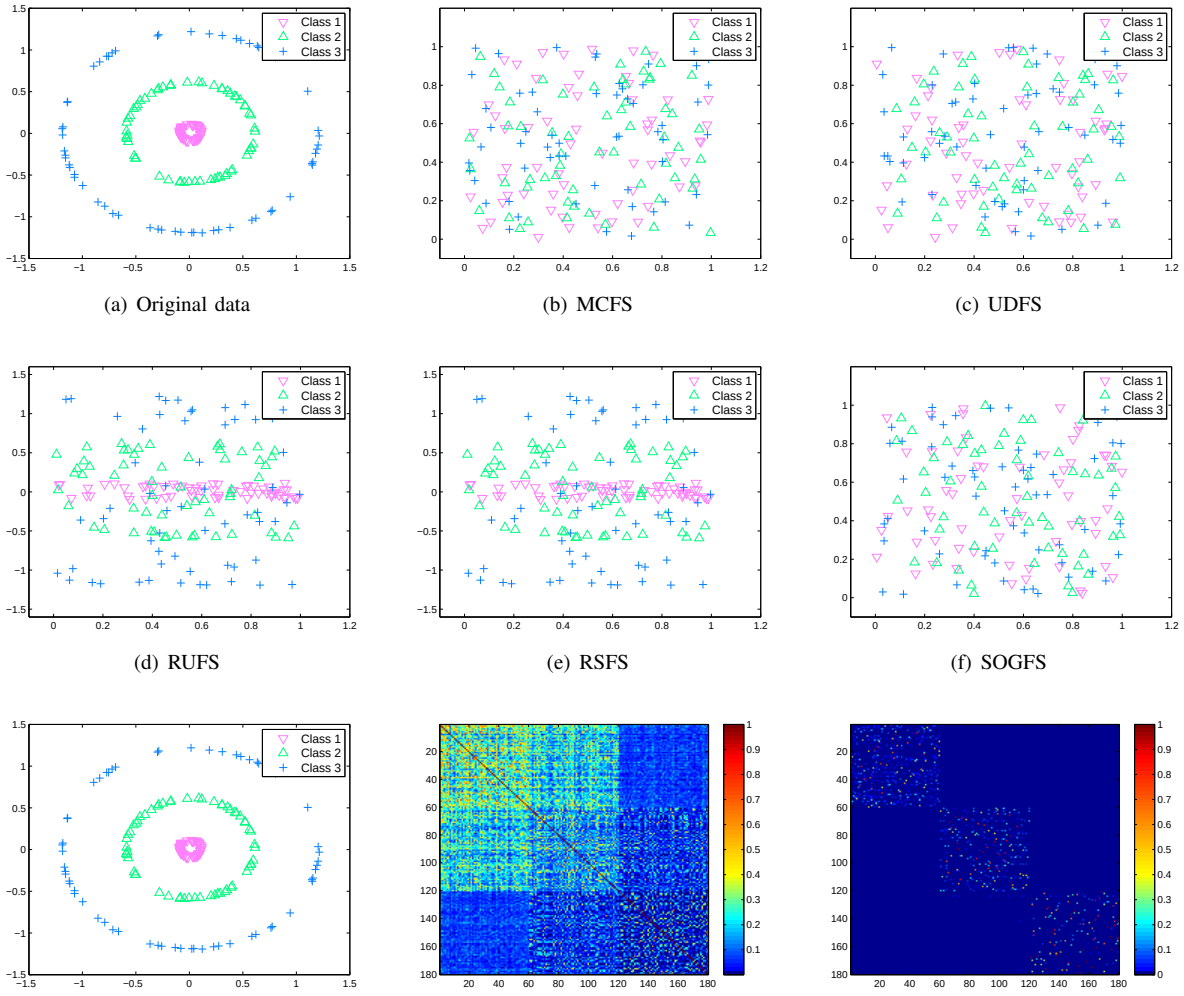
Arrhythmia and Isolet5 are from the UCI Machine Learning Repository, BA is a handwritten digit dataset, LUNG and SRBCT are biology datasets. The detailed description of these datasets is exhibited in Table I (Imbalance Rate is shorten as IR). As illustrated in the Table I, for the datasets with high dimensionality, the number of selected features starts with 50 and steps by 50 until reaching 300. For those with low dimensionality, the number starts with 10 and steps by 10 until reaching 100.

Competitors: To validate the effectiveness of STDA, six state-of-the-art competitors are taken for comparison. They are Laplacian Score (LS) [7], Multi-Cluster Feature Selection (MCFS) [26], Unsupervised Discriminate Feature Selection



(a) MCFS (b) UDFS (c) RUFS (d) RSFS (e) SOGFS (f) STDA

Fig. 3. Face features obtained by (a) MCFS, (b) UDFS, (c) RUFS, (d) RSFS, (e) SOGFS and (f) STDA with different number of selected features (from left to right).



(a) Original data (b) MCFS (c) UDFS

(d) RUFS (e) RSFS (f) SOGFS

(g) STDA (h) Gaussian graph (i) Graph learned by STDA

Fig. 4. (a) The first two dimensions of toy data. (b)-(g) visualize the two features selected by different methods, and the coordinate of each point is the corresponding feature values. (h) The initial data graph. (i) Graph learned by the proposed STDA. Best viewed in color.

TABLE II
CLUSTERING RESULTS ON REAL-WORLD DATASETS. THE BEST RESULTS ARE SHOWN IN BOLD FACE.

	ORL			YALE			COIL20			Arrhythmia		
	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
RCut	0.25	0.45	0.27	0.25	0.31	0.29	0.29	0.41	0.31	0.45	0.12	0.57
NCut	**0.61**	0.76	0.64	0.45	0.49	0.46	0.64	0.75	0.67	0.21	0.13	0.56
NMF	0.35	0.60	0.39	0.34	0.43	0.38	0.45	0.58	0.48	0.24	0.15	0.57
STDA	0.58	**0.83**	**0.77**	**0.47**	**0.52**	**0.49**	**0.85**	**0.93**	**0.88**	**0.58**	**0.21**	**0.62**

	Isolet5			BA			LUNG			SRBCT		
	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
RCut	0.21	0.34	0.23	0.26	0.39	0.27	0.53	0.35	0.77	0.37	0.07	0.37
NCut	**0.64**	**0.74**	0.67	0.35	0.51	0.38	0.53	0.39	0.79	0.38	0.10	0.41
NMF	0.33	0.44	0.37	0.22	0.34	0.23	0.52	0.31	0.73	0.45	0.15	0.48
STDA	0.48	0.63	**0.77**	**0.40**	**0.55**	**0.53**	**0.79**	**0.49**	**0.84**	**0.48**	**0.31**	**0.51**

(UDFS) [24], Robust Spectral Feature Selection (RSFS) [29] and Structured Optimal Graph Feature Selection (SOGFS) [23]. Among them, LS is a filter method and the others are embedded methods. For a fair comparison, the optimal parameters of each competitors are searched from $\{10^{-6}, 10^{-4}, ..., 10^{6}\}$, and the neighborhood size is set to be 5. Suggested by Nie et al. [23], we set the projected dimension m as $\frac{2}{3}d$ (d is the original dimensionality). Following Shi et al. [29], the best parameters are selected for each feature dimension. Each method is compared at its optimal projection dimension. The k-means clustering result with all features is taken as the baseline.

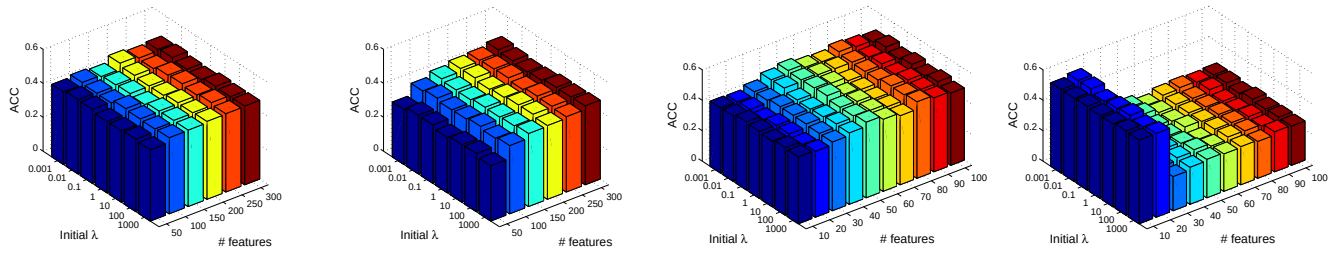
Performance: With different feature numbers, the ACC and NMI curves of all the methods are shown in Figure 1 and 2. Compared to the baseline (results with all features), the clustering results with selected features are always better, implying the fact that the real-world data contains redundant features. The performance of LS is unsatisfying because it learns the significance of each feature independently and neglects their intrinsic correlations. UDFS captures the discriminant features and selects them jointly, but it just focuses on the global relationship of data points and fails to exploit the local manifold. MCFS considers the local data structure, but it is sensitive to noise. RUFS, RSFS and SOGFS are more robust to some extent, but they fail to find the discriminant subspace where the features can be clearly classified. The proposed method captures the discriminative information, and optimizes the data graph adaptively in the learned subspace, where the data noise in the input space is avoided. So it is able to find the discriminative features with robustness, and achieves the best performance on all datasets. Moreover, there is no additional parameter to be tuned in our method, while all the competitors involve several parameters. Thus, the proposed method is more applicable than the others. In addition, Arrhythmia, Lung and SRBCT consist of unbalanced data. Especially for Arrhythmia, the point numbers for the 1st and 8th classes are 245 and 2 respectively. The proposed STDA performs well on these datasets, which validates that it is able to handle the unbalanced data.

We further visualize the features selected by MCFS, UDFS, RUFS, RSFS, SOGFS and the proposed STDA in Figure 3. We randomly choose one sample from the ORL dataset, and select $\{128, 256, 384, 512, 640, 768, 896, 1024\}$ features (from left to right). Each pixel is a feature. For a better illustration, the selected features keep their pixel values, and the unselected ones are set to white. As can be seen in Figure 3, compared to the competitors, STDA tends to capture more discriminative features, such as eyes, nose and mouth. Especially, compared to UDFS, which also performs discriminant analysis, our method drops more background (skin) pixels. This is because that STDA is able to exploit the local data structure. So we can say that the combination of discriminant analysis and data graph learning is helpful for feature selection.

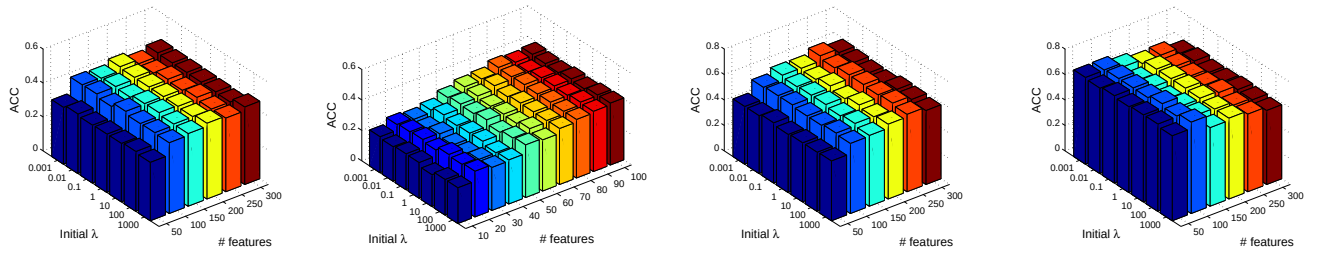
In addition, a toy dataset is introduced to further verify the capability of our method to select the discriminative features. As visualized in Figure 4 (a), the dataset is formed by the data points from three classes, each class contains 60 points. The data points reside in concentric circles at the first two dimensions, and the other eight dimensions are noises randomly generated in the range of 0 and 1. So only the first two dimensions are valuable features. To illustrate the effectiveness of local structure learning, we use the the self-tune Gaussian method [49] to construct the affinity matrix for all the methods. The top two features selected by different methods are shown in Figure 4 (b)-(g), where each point takes its selected features as the coordinate. it is manifest that the features selected by STDA correctly preserve the intrinsic data structure. As shown in Figure 4 (h) and (g), although the input graph contains large noise, our method still learns the optimal graph with clear cluster structure. The proposed STDA utilizes the merits of LDA and captures the local relationship of points adaptively, so it is able to find the discriminative features with robustness.

B. Performance on Clustering

As mentioned in Section III-B, the proposed method can be also used for clustering. So we compare its clustering performance with three widely used clustering methods, Ratio Cut (RCut) [50], Normalized Cut [40] and Non-negative

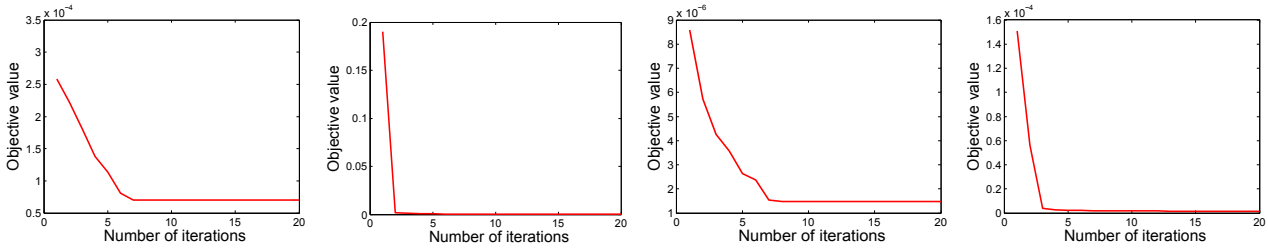


(a) ORL (b) YALE (c) COIL20 (d) Arrhythmia

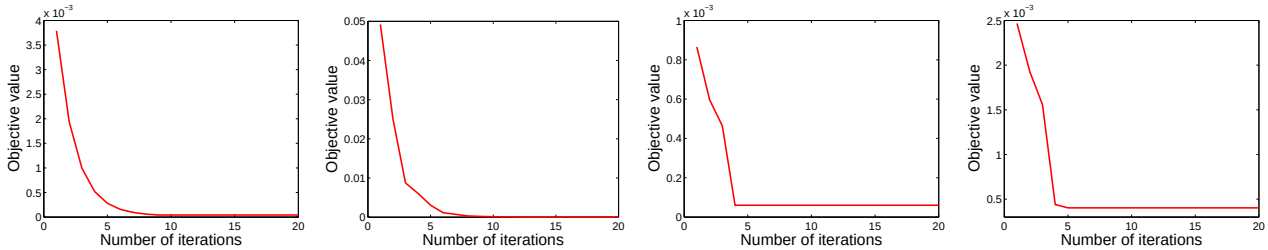


(e) Isolet5 (f) BA (g) LUNG (h) SRBCT

Fig. 5. Clustering accuracy with different initial value of parameter λ. The results are robust to the initial λ.



(a) ORL (b) YALE (c) COIL20 (d) Arrhythmia



(e) Isolet5 (f) BA (g) LUNG (h) SRBCT

Fig. 6. Convergence behavior of STDA on different datasets.

Matrix Factorization (NMF) [34]. For RCut and NCut, the self-tune Gaussian method [49] is utilized to construct the affinity matrix. Since RCut and NCut are sensitive to the initialization, we repeat them for 50 repetitions and report their mean results. The clustering ACCuracy (ACC) [41], Normalized Mutual Information (NMI) and Purity [42] are taken as measurements. NMI measures the mutual information between the predicted labels and the ground-truth, and Purity indicates the extent to which the points within the same cluster come from the same class. The clustering results on the real-world datasets are reported in Table II .

Table II shows that the proposed STDA achieves the best performance in most cases, and outperforms RCut and NMF on all the datasets. On the Isolet5, NCut shows better performance than STDA because this dataset contains less redundant features and the data structure is clear. On the other datasets, STDA performs well. RCut and NCut highly depend on the initial affinity graph, so they tend to be affected the data noise. NMF just emphasizes the global data structure and neglects the local aspect. On the other hand, the proposed STDA learns the data graph adaptively in the optimization procedure, and captures the most discriminative features, so it shows better performance. In addition, the results of both RCut and NCut are unstable (i.e., different outputs for every time of running),

while our method achieves stable clustering performance as it does not involve k-means as the post-processing.

C. Influence of Initial λ and Convergence Behavior

Here we discuss the performance variation of STDA with different initial λ. The parameter λ in Eq. (7) balances the importance of the rank constraint. Since λ is self-tuned, we show the variance of feature selection performance versus the initial λ. In Figure 5, it can be seen that the results are insensitive to the initial λ. We simply set the initial value to 1 in our experiments.

The convergence behavior of the proposed optimization algorithm is also demonstrated experimentally. Figure 6 shows the convergence curves on all the datasets, where the objective value decreases during the iterations. As shown in the figure, the optimization method converges fast (within 10 iterations) on all occasions.

V. Conclusion

In this work, we present an unsupervised feature selection method called Self-Tuned Discrimination-Aware (STDA), which is able to capture the discriminative features. An efficient optimization strategy is developed to solve the problem. Different from existing works, STDA jointly incorporates the merits of Linear Discriminant Analysis and data graph learning, so it can exploits the local manifold structure in the discriminative subspace. Moreover, it accomplishes data clustering at the same time, and saves the efforts for tuning parameters. Experimental results on different kinds of datasets demonstrate the promising performance of STDA on both feature selection and data clustering tasks.

Acknowledge

This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant 61773316 and 61871470, Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

References

[1] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1279–1289, 2016.

[2] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5910–5922, 2018.

[3] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Advances in Neural Information Processing Systems*, 2010, pp. 1813–1821.

[4] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[5] D. Koller and M. Sahami, "Toward optimal feature selection," in *International Conference on Machine Learning*, 1996, pp. 284–292.

[6] P. Mitra, C. Murthy, and S. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[7] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems*, 2005, pp. 507–514.

[8] M. Masaeli, G. Fung, and J. Dy, "From transformation-based dimensionality reduction to feature selection," in *International Conference on Machine Learning*, 2010, pp. 751–758.

[9] F. Nie, D. Xu, X. Li, and S. Xiang, "Semisupervised dimensionality reduction and classification through virtual label regression," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 41, no. 3, pp. 675–685, 2011.

[10] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 793–804, 2014.

[11] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering - A filter solution," in *IEEE International Conference on Data Mining*, 2002, pp. 115–122.

[12] Q. Huang, D. Tao, X. Li, L. Jin, and G. Wei, "Exploiting local coherent patterns for unsupervised feature ranking," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 41, no. 6, pp. 1471–1482, 2011.

[13] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *AAAI Conference on Artificial Intelligence,*, 2008, pp. 671–676.

[14] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *International Conference on Machine Learning*, 2007, pp. 1151–1157.

[15] Y. Jiang and J. Ren, "Eigenvalue sensitive feature selection," in *International Conference on Machine Learning*, 2011, pp. 89–96.

[16] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[17] V. Roth and T. Lange, "Feature selection in clustering problems," in *Advances in Neural Information Processing Systems*, 2003, pp. 473–480.

[18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[19] C. Constantinopoulos, M. Titsias, and A. Likas, "Bayesian feature and model selection for gaussian mixture models," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 28, no. 6, pp. 1013–1018, 2006.

[20] S. Yang, S. Yan, C. Zhang, and X. Tang, "Bilinear analysis for kernel selection and nonlinear feature extraction,"

IEEE Transactions on Neural Networks, vol. 18, no. 5, pp. 1442–1452, 2007.

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

[22] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized gaussian mixture model for data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1406–1418, 2011.

[23] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1302–1308.

[24] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l2,1-norm regularized discriminative feature selection for unsupervised learning," in *International Joint Conference on Artificial Intelligence*, 2011, pp. 1589–1594.

[25] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *AAAI Conference on Artificial Intelligence*, 2012, pp. 1026–1032.

[26] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333–342.

[27] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 1621–1627.

[28] F. Nie, D. Xu, I. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1921–1932, 2010.

[29] L. Shi, L. Du, and Y. Shen, "Robust spectral learning for unsupervised feature selection," in *IEEE International Conference on Data Mining*, 2014, pp. 977–982.

[30] Z. Zhang and N. Ye, "Locality preserving multimodal discriminative learning for supervised feature selection," *Knowledge and Information Systems*, vol. 27, no. 3, pp. 473–490, 2011.

[31] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, p. 796, 2016.

[32] J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection," in *International Conference on Data Mining*, 2014, pp. 938–946.

[33] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *AAAI Conference on Artificial Intelligence*, 2010.

[34] D. Li and H. Seung, "Algorithms for nonnegative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, no. 6, pp. 556–562, 2000.

[35] M. Friedman and A. Kandel, *Introduction to Pattern Recognition - Statistical, Structural, Neural and Fuzzy Logic Approaches*. WorldScientific, 1999, vol. 32.

[36] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 2201–2207.

[37] B. Mohar, Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, "The laplacian spectrum of graphs," in *Graph Theory, Combinatorics, and Applications*, 2001, pp. 871–898.

[38] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1969–1976.

[39] K. Fan, "On a theorem of weyl concerning eigenvalues of linear transformations i." *Proc Natl Acad Sci U S A*, vol. 36, no. 1, pp. 652–655, 1949.

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

[41] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4147–4153.

[42] Q. Wang, M. Chen, and X. Li, "Quantifying and detecting collective motion by manifold learning," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4292–4298.

[43] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[44] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (coil-20)," *Technical Report*, 1996.

[45] M. Lichman, "UCI machine learning repository," *University of California, Irvine, School of Information and Computer Sciences*, 2013.

[46] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[47] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, and J. Richie, "Gene expression correlates of clinical prostate cancer behavior." *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[48] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[49] L. Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, 2004, pp. 1601–1608.

[50] L. Hagen and A. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE Transactions on CAD of Integrated Circuits and Systems*, vol. 11, no. 9, pp. 1074–1085, 1992.

Xuelong Li (M'02-SM'07-F'12) is currently a Full Professor with the School of Computer Science and with the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China.

Mulin Chen received the B.E. degree in software engineering and the M.E. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2016 respectively. He is currently pursuing the Ph.D. degree with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research intersts include computer vision and machine learning.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.