Unsupervised Feature Selection with Adaptive Structure Learning

Liang Du
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
School of Computer and Information Technology, Shanxi University
duliang@ios.ac.cn

Yi-Dong Shen
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
ydshen@ios.ac.cn

ABSTRACT

The problem of feature selection has raised considerable interests in the past decade. Traditional unsupervised methods select the features which can faithfully preserve the intrinsic structures of data, where the intrinsic structures are estimated using all the input features of data. However, the estimated intrinsic structures are unreliable/inaccurate when the redundant and noisy features are not removed. Therefore, we face a dilemma here: one need the true structures of data to identify the informative features, and one need the informative features to accurately estimate the true structures of data. To address this, we propose a unified learning framework which performs structure learning and feature selection simultaneously. The structures are adaptively learned from the results of feature selection, and the informative features are reselected to preserve the refined structures of data. By leveraging the interactions between these two essential tasks, we are able to capture accurate structures and select more informative features. Experimental results on many benchmark data sets demonstrate that the proposed method outperforms many state of the art unsupervised feature selection methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms

Keywords

unsupervised feature selection; adaptive structure learning

1. INTRODUCTION

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783345.

Real world applications usually involve big data with high dimensionality, presenting great challenges such as the curse of dimensionality, huge computation and storage cost. To tackle these difficulties, feature selection techniques are developed to keep a few relevant and informative features. According to the availability of label information, these algorithms can be categorized into supervised [25], [22], [20], [17], semi-supervised [33], [27] and unsupervised algorithms [6], [4]. Compared to supervised or semi-supervised counterparts, unsupervised feature selection is generally more challenging due to the lack of supervised information to guild the search of relevant features.

Unsupervised feature selection has attracted much attention in recent years and a number of algorithms have been proposed [8, 4, 36, 28, 16]. Without class label, unsupervised feature selection chooses features that can effectively reveal or maintain the underlying structure of data. Recent research on feature selection and dimension reduction has witnessed that several important structures should be preserved by the selected features. These important structures include, but not limited to, the global structure [36, 16], the local manifold structure [9, 10] and the discriminative information [28, 14]. And these structures can be captured by widely used models in the form of graph, such as, the sample pairwise similarity graph [36], the k-nn graph [8], the global integration of local discriminant model [28, 31], the local linear embedding (LLE) [16].

Clearly, many of existing unsupervised feature selection methods rely on the structure characterization through some kind of graph, which can be computed within the *original* feature space. And once the graph is determined, it is *fixed* in the next procedures, such as sparse spectral regression [3], to guide the search of informative features. Consequently, the performance of feature selection is largely determined by the effectiveness of graph construction. Ideally, such graphs should be constructed only using the informative feature subset rather than all candidate features. Unfortunately, the desired subset of features is unknown in advance, and the irrelevant or noisy features would be inevitably introduced in many real applications. As a result, unrelated or noisy features will have an adverse effect on the characterization of the structures and henceforth hurt the following feature selection performance.

In unsupervised scenario, this is actually the chicken-and-egg problem between *structure characterization* and *feature selection*. Facing with such dilemma, we propose to perform

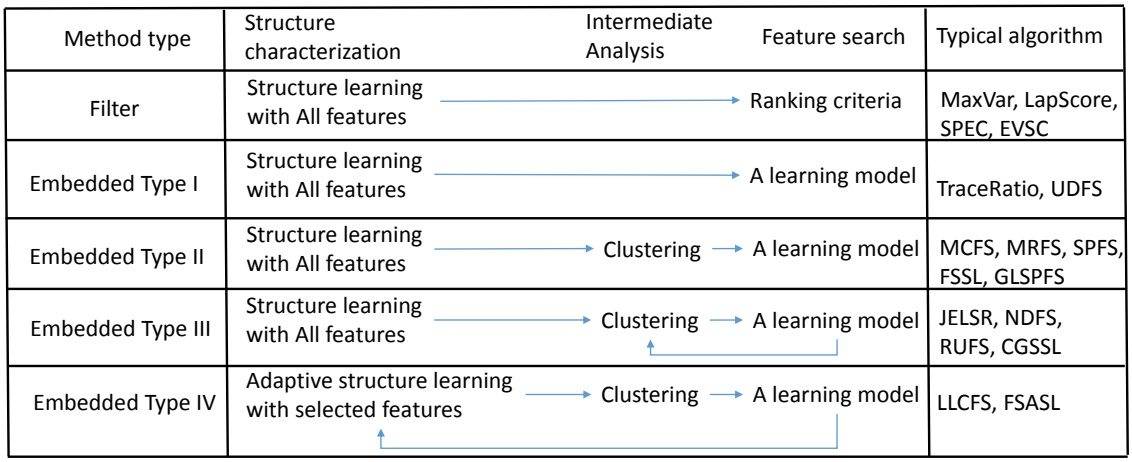

Figure 1: An illustration of unsupervised filter methods and four type embedded methods.

feature selection and structure learning in a unified framework, where each sub task can be iteratively boosted by using the result of the other one. Concretely, the global structure of data is captured within the sparse representation framework, where the reconstruction coefficient is learned from the selected features. The local manifold structure is revealed by a probabilistic neighborhood graph, where the pairwise relationship is also determined by the selected features. When the global and local structures are given in the form of graph Laplacians, we seek the relevant features via sparse spectral regression with the help of graph embedding for cluster analysis. In this way, both the global and local structure of data can be better captured by only using the selected features; Moreover, with the refined characterization of the structure, a better search of the informative features could also be expected.

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

1. Based on the different learning paradigms for unsupervised feature selection, we investigate most of existing unsupervised embedded methods and further classify them into four closely related but different types. These analyses provide more insight into what should be further emphasized on the development of more essential unsupervised feature selection algorithm.

2. We propose a novel unified learning framework, called unsupervised Feature Selection with Adaptive Structure Learning (FSASL in short), to fulfil the gap between two essential sub tasks, i.e. structure learning and feature learning. In this way, these two tasks can be mutually improved.

3. Comprehensive experiments on benchmark data sets show that our method achieves statistically significant improvement over state-of-the-art feature selection methods, suggesting the effectiveness of the proposed method.

2. RELATED WORKS

In this section, we mainly review most existing unsupervised feature selection methods, i.e. filter and embedded methods. Unsupervised filter methods pick the features one by one based on certain evaluation criteria, where no learning algorithm is involved. The typical methods include: max variance (MaxVar) [12], Laplacian score (LapScore) [8], spectral feature selection (SPEC) [34], feature selection via eigenvalue sensitive criterion (EVSC) [4]. A common limitation of these approaches is the correlation among features is neglected [1].

Unsupervised embedded approaches are developed to perform feature selection and fit a learning model simultaneously. Based on the different sub-steps involved in the feature selection procedure, these embedded methods can be further divided into four different types as illustrated in Figure 1.

The first type of embedded methods first detect the structure of the data and then directly select those features which is used to best preserve the enclosed structure. The typical methods include: trace ratio (TraceRatio) [19] and unsupervised discriminative feature selection (UDFS) [28]. TraceRatio is prone to select redundant features [16] and the learning model of UDFS is often too restrictive [21].

The second type of embedded methods first construct various graph Laplacians to capture the data structure, then flat the cluster structure via graph embedding, and finally use the sparse spectral regression [3] to select those features that are best aligned to the embedding. Instead of directly selecting features as the first type, these approaches resorted to an *intermediate cluster analysis sub-step* to reveal the cluster structure for guiding the selection of features. The cluster structure discovered by either the graph spectral embedding or other clustering module can be seen as an approximation of the unseen labels. The typical methods include: multi-cluster feature selection (MCFS) [4], minimum redundancy spectral feature selection (MRSF) [35], similarity preserving feature selection (SPFS) [36], and joint feature selection and subspace learning (FSSL) [7], global and local structure preserving feature selection (GLSPFS) [16].

Unlike the second type methods, the clustering analysis in the third type of embedded methods is co-determined by the embedding of the graph Laplacian and the *adaptive discriminative regularization* [29], [31], which can be obtained from the result of sparse spectral regression. By using the feedback from feature selection, the whole learn-

ing procedure can provide better cluster analysis, and vice versa. The typical methods include: joint embedding learning and spectral regression (JELSR) [11], [10], nonnegative discriminative feature selection (NDFS) [14], robust unsupervised feature selection (RUFS) [21], feature selection via clustering-guided sparse structural learning (CGSSL) [13].

The fourth type of embedded methods try to feed the result of feature selection into the structure learning procedure for improving the quality of structure learning. In [32], a feature selection method is proposed for local learning-based clustering (LLCFS), which incorporates the relevance of each feature into the built-in regularization of the local learning model, where the induced graph Laplacian can be iteratively updated. However, LLCFS uses the discrete k-nearest neighbor graph and does not optimize the same objective function in structure learning and feature search.

It can be seen that all these above methods (except LLCFS) share a common drawback: they use all features to estimate the underlying structures of data. Since the redundant and noisy features are unavoidable in real world applications, that is also why we need feature selection, the learned structures using all features will also be contaminated, which would degrade the performance of feature selection. By leveraging the coherent interactions between structure learning and feature selection, our proposed method seamlessly integrates them into a unified framework, where the result of one task is used to improve the other one.

3. UNSUPERVISED FEATURE SELECTION WITH ADAPTIVE STRUCTURE LEARNING

Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathcal{R}^{d\times n}$ denotes the data matrix, whose columns correspond to data instances and rows to features. The generic problem of unsupervised feature selection is to find the most informative features. With the absence of class label to guild the search of relevant features, the data represented with the selected features should well preserve the intrinsic structure as the data represented by all the original features.

To achieve this goal, we propose to jointly perform unsupervised feature selection and data structure learning simultaneously, where both global and local structure are adaptively updated using the result of current feature selection.

We first summarize some notations used throughout this paper. We use bold uppercase characters to denote matrices, bold lowercase characters to denote vectors. For an arbitrary matrix $\mathbf{A} \in \mathcal{R}^{r\times t}$, $\mathbf{a}_i$ means the i-th column vector of $\mathbf{A}$ and $\mathbf{a}_j^T$ means the j-th row vector of $\mathbf{A}$, $\mathbf{A}_{ij}$ denotes the (i,j)-th entry of $\mathbf{A}$. The $\ell_{2,1}$-norm is defined as $||\mathbf{A}||_{21} = \sum_{i=1}^{r}\sqrt{\sum_{j=1}^{t}\mathbf{A}_{ij}^2}$.

3.1 Adaptive Global Structure Learning

Over the past decades, a large number of algorithms have been proposed based on the analysis of the global structure of data, such as the Principal Component Analysis (PCA) and the Maximum Variance (MaxVar). Recently, the global pairwise similarity (e.g., with a Gaussian kernel) between high-dimensional samples has demonstrated promising performance for unsupervised feature selection [36, 16]. However, such dense similarity becomes less discriminative for high dimension data, especially when there are many unfavorable features in the original high dimensional space.

Inspired by the recent development on compressed sensing and sparse representation [26], we use the sparse reconstruction coefficients to extract the global structure of data. In sparse representation, each data sample $\mathbf{x}_i$ can be approximated as a linear combination of all the other samples, and the optimal sparse combination weight matrix $\mathbf{S} \in \mathcal{R}^{n\times n}$ can be obtained by solving the following problem

$$\min_{\mathbf{S}} \quad \sum_{i=1}^{n}\left(||\mathbf{x}_i - \mathbf{X}\mathbf{s}_i||^2 + \alpha||\mathbf{s}_i||_1\right) \quad \text{s.t.} \quad \mathbf{S}_{ii} = 0 \qquad (1)$$

where α is used to balancing the sparsity and the reconstruction error. Compared with the pairwise similarity, the sparse representation is naturally discriminative: among all the candidates samples, it selects the samples which most compactly expresses the target and rejects all other possible but less compact candidates [26].

Clearly, the selected features should preserve such global and sparse reconstruction structure. To achieve this, we introduce a row sparse feature selection and transformation matrix $\mathbf{W} \in \mathcal{R}^{d\times c}$ to the reconstruction process, and get

$$\min_{\mathbf{S},\mathbf{W}} \quad \sum_{i=1}^{n}||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{X}\mathbf{s}_i||^2 + \alpha||\mathbf{S}||_1 + \gamma||\mathbf{W}||_{21} \qquad (2)$$
$$\text{s.t.} \quad \mathbf{S}_{ii} = 0, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

where γ is regularization parameter. Compared with the Eq.(1), the benefits of Eq.(2) are two folds: 1) The global structure captured by $\mathbf{S}$ can be used to guide the search of relevant features; 2) By largely eliminating the adverse effect of noisy and unfavorable features, the global structure can also be better estimated.

3.2 Adaptive Local Structure Learning

The importance of preserving local manifold structure has been well recognized in the recent development of unsupervised feature selection algorithms, especially considering that high-dimensional data often presents a low-dimensional manifold structure [8, 4, 16]. To detect the underlying local manifold structure, these algorithms usually first construct a k-nearest neighbor graph and then compute the graph Laplacian with different models. Clearly, both the k-nn graph and the graph Laplacian are determined by *all the relevant and irrelevant features*. As a result, the manifold structure captured by such graph Laplacian would be inevitably affected by the redundant and noisy features. Moreover, the iterative updating of discrete neighborhood relationship using the result of feature selection still suffers from the lack of theoretical guarantee of its convergence [32, 24].

Instead of using the graph Laplacian with the determinate neighborhood relationship, we introduce to directly learn a euclidean distance induced probabilistic neighborhood matrix [18]. For each data sample $\mathbf{x}_i$, all the data points $\{\mathbf{x}_j\}_{j=1}^n$ are considered as the neighborhood of $\mathbf{x}_i$ with probability $\mathbf{P}_{ij}$, where $\mathbf{P} \in \mathcal{R}^{n\times n}$ can be determined by solving the following problem:

$$\min_{\mathbf{P}} \sum_{i,j}(||\mathbf{x}_i - \mathbf{x}_j||_2^2\mathbf{P}_{ij} + \mu\mathbf{P}_{ij}^2), \text{ s.t. } \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \geq 0 \quad (3)$$

where μ is the regularization parameter. The regularization term is used to 1) avoid the trivial solution; 2) add a prior

of uniform distribution. It can be found that a large distance $||\mathbf{x}_i - \mathbf{x}_j||_2^2$ will lead to a small probability $\mathbf{P}_{ij}$. With such nice property, the estimated weight matrix $\mathbf{P}$ and the induced Laplacian $\mathbf{L_P} = \mathbf{D_P} - (\mathbf{P} + \mathbf{P}^T)/2$ can be used for local manifold characterization, where $\mathbf{D_P}$ is a diagonal matrix whose i-th diagonal element is $\sum_j (\mathbf{P}_{ij} + \mathbf{P}_{ji})/2$.

To leverage the result of feature selection and iteratively improve the probabilistic neighborhood relationship, we also introduce the feature selection and transformation matrix $\mathbf{W}$ as used in global structure adaptive learning, and we get

$$\min_{\mathbf{P},\mathbf{W}} \quad \sum_{i,j}^{n} (||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||_2^2 \mathbf{P}_{ij} + \mu \mathbf{P}_{ij}^2) + \gamma ||\mathbf{W}||_{21} \quad (4)$$
$$\text{s.t.} \quad \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \geq \mathbf{0}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

With the sparsity on $\mathbf{W}$, the irrelevant and noisy features can be largely removed, thus we can learn a better probabilistic neighborhood graph for local structure characterization based on the result of feature selection, i.e. $\mathbf{W}^T\mathbf{X}$. Moreover, we aim to seek those features to preserve the local structure encoded by $\mathbf{P}$. Thus, the optimization problem in Eq. (4) can be used to perform feature selection and local structure learning, simultaneously.

3.3 Unsupervised Feature Selection with Adaptive Structure Learning

Based on the two adaptive structure learning models presented in Eq. (2) and Eq. (4), we propose a novel unsupervised feature selection method by solving the following optimization problem,

$$\min_{\mathbf{W},\mathbf{S},\mathbf{P}} \quad \left(||\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{S}||^2 + \alpha||\mathbf{S}||_1\right) \quad (5)$$
$$+ \quad \beta \sum_{i,j}^{n} \left(||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||^2 \mathbf{P}_{ij} + \mu \mathbf{P}_{ij}^2\right) + \gamma ||\mathbf{W}||_{21}$$
$$\text{s.t.} \quad \mathbf{S}_{ii} = 0, \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \geq \mathbf{0}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

where β and γ are regularization parameters balancing the fitting error of global and local structure learning in the first and second group and the sparsity of the feature selection matrix in the third group.

It can be seen that when both $\mathbf{S}$ and $\mathbf{P}$ are given, our method selects those features to well respect both the global and local structure of data. When the feature selection matrix $\mathbf{W}$ is given, our method learns the global and local structure of data in a transformed space, i.e. $\mathbf{W}^T\mathbf{X}$, where the adverse effect of noisy features is largely alleviated with sparse regularization. In this way, these two essential tasks can be boosted by the other one within a unified learning framework. Since both the global and local structure can be adaptively refined according to the result of feature selection, we call Eq. (5) unsupervised Feature Selection with Adaptive Structure Learning (FSASL).

3.4 Optimization Algorithm

Because the optimization problem in Eq. (5) comprises three different variables with different regularizations and constraints, it is hard to derive its closed solution directly. Thus we derive an alternative iterative algorithm to solve the problem, which converts the problem with a couple of variables ($\mathbf{S}$, $\mathbf{P}$ and $\mathbf{W}^T$) into a series of sub problems where only one variable is involved.

First, when $\mathbf{W}$ and $\mathbf{P}$ are fixed, we need to solve n decoupled sub problems in the following form:

$$\min_{\mathbf{s}_i} \quad ||\mathbf{x}_i^{'} - \mathbf{X}^{'}\mathbf{s}_i||^2 + \alpha|\mathbf{s}_i|, \quad \text{s.t.} \quad \mathbf{S}_{ii} = 0 \quad (6)$$

where $\mathbf{X}^{'}$ is the new transformed data by projecting the relevant features into a low dimension space, and $\mathbf{X}^{'} = \mathbf{W}^T\mathbf{X}$. The above LASSO problem can be efficiently solved by routine optimization tools, e.g. proximal methods [2, 15].

Next, when $\mathbf{W}^T$ and $\mathbf{S}$ are fixed, we need to solve n decoupled sub problems in the following form:

$$\min_{\mathbf{p}_i^T} \quad \sum_{j=1}^{n} ||\mathbf{x}_i^{'} - \mathbf{x}_j^{'}||^2 \mathbf{P}_{ij} + \mu ||\mathbf{P}_{ij}||^2, \quad (7)$$
$$\text{s.t.} \quad \mathbf{1}_n^T \mathbf{p}_i = 1, \mathbf{P}_{ij} \geq 0$$

Denote $\mathbf{A} \in \mathcal{R}^{n \times n}$ be a square matrix with $\mathbf{A}_{ij} = -\frac{1}{2\mu}||\mathbf{x}_i^{'} - \mathbf{x}_j^{'}||^2$, then the above problem can be written as follows

$$\min_{\mathbf{p}_i^T} \quad \frac{1}{2}||\mathbf{p}_i^T - \mathbf{a}_i^T||^2, \quad \text{s.t.} \quad \mathbf{p}_i^T\mathbf{1}_n = 1, 0 \leq \mathbf{p}_{ij}^T \leq 1 \quad (8)$$

where $\mathbf{p}_i^T$ is the i-th row of $\mathbf{P}$. The above euclidean projection of a vector onto the probability simplex can be efficiently solved by Algorithm 1 without iterations. More details can be found in Eq. (19).

Algorithm 1 The optimization algorithm of Eq. (8)

Input: $\mathbf{a}$
sort $\mathbf{a}$ into $\mathbf{b}$ where $b_1 \geq b_2 \geq, ..., b_n$
find $\rho = \max\{1 \leq j \leq n : b_j + \frac{1}{j}(1 - \sum_{i=1}^{j} b_i) > 0\}$
define $z = \frac{1}{\rho}(1 - \sum_{i=1}^{\rho} b_i)$
Output: $\mathbf{p}$ with $p_j = \max\{a_j + z, 0\}, j = 1, ..., n$

Next, when $\mathbf{S}$ and $\mathbf{P}$ are fixed, we need to solve the following problem:

$$\min_{\mathbf{W}} \quad ||\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{S}||^2 + \beta \sum_{i,j}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||^2 \mathbf{P}_{ij} + \gamma ||\mathbf{W}||_{21}$$
$$\text{s.t.} \quad \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I} \quad (9)$$

Using $\mathbf{L_S} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T$, $\mathbf{L_P} = \mathbf{D_P} - (\mathbf{P} + \mathbf{P}^T)/2$ and let $\mathbf{L} = \mathbf{L_S} + \beta\mathbf{L_P}$, the above problem can be rewritten as

$$\min_{\mathbf{W}} \quad Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) + \gamma ||\mathbf{W}||_{21} \quad (10)$$
$$\text{s.t.} \quad \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

Due to the non-smooth regularization, it is hard to obtain the close form solution. We solve it in an iterative way. Given the t-th estimation $\mathbf{W}^t$ and denote $\mathbf{D}_{\mathbf{W}^t}$ be a diagonal matrix with the i-th diagonal element as $\frac{1}{2||\boldsymbol{w}_i^t||^2}$, Eq. (10) can be rewritten as:

$$\min_{\mathbf{W}} \quad Tr\left(\mathbf{W}^T\mathbf{X}(\mathbf{L} + \gamma\mathbf{D}_{\mathbf{W}^t})\mathbf{X}^T\mathbf{W}\right) \quad (11)$$
$$\text{s.t.} \quad \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

The optimal solution of $\mathbf{W}$ are the eigenvectors corresponding to the c smallest eigenvalues of generalized eigen-problem:

$$\mathbf{X}(\mathbf{L} + \gamma\mathbf{D}_{\mathbf{W}^t})\mathbf{X}^T\mathbf{W} = \Lambda\mathbf{X}\mathbf{X}^T\mathbf{W} \quad (12)$$

where Λ is a diagonal matrix whose diagonal elements are eigenvalues. To get a stable solution of this eigen-problem, the matrices $\mathbf{X}\mathbf{X}^T$ is required to be non-singular which is not true when the number of features is larger than the number of samples. Moreover, the computational complexity of this approach scales as $O(d^3 + nd^2)$, which is costly for high dimensional data. Thus, such solution is less attractive in real world applications. To improve the effectiveness and the efficiency to optimize Eq. (10), we further resort to a two steps procedure inspired from [3].

THEOREM 1. *Let $\mathbf{Y} \in \mathcal{R}^{n\times c}$ be a matrix of which each column is an eigenvector of eigen-problem $\mathbf{L}\mathbf{y} = \lambda\mathbf{y}$. If there exists a matrix $\mathbf{W} \in \mathcal{R}^{d\times c}$ such that $\mathbf{X}^T\mathbf{W} = \mathbf{Y}$, then each column of $\mathbf{W}$ is an eigenvector of the generalized eigen-problem $\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{w}$ with the same eigenvalue λ.*

PROOF. *With $\mathbf{X}^T\mathbf{W} = \mathbf{Y}$, the following equation holds*

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} = \mathbf{X}\mathbf{L}\mathbf{y} = \mathbf{X}\lambda\mathbf{y} = \lambda\mathbf{X}\mathbf{y} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{w} \tag{13}$$

Thus, $\mathbf{y}$ is the eigenvector of the generalized eigen-problem $\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{w}$ with the same eigenvalue λ. □

Theorem 1 shows that instead of solving the generalized eigen-problem in Eq. (12), $\mathbf{W}$ can be obtained by the following two steps:

1. Solve the eigen-problem $\mathbf{L}\mathbf{Y} = \Lambda\mathbf{Y}$ to get $\mathbf{Y}$ corresponding to the c smallest eigenvalues;
2. Find $\mathbf{W}$ which satisfies $\mathbf{X}^T\mathbf{W} = \mathbf{Y}$. Since such $\mathbf{W}$ may not exist in real applications, we resort to solve the following optimization problem:

$$\min_{\mathbf{W}} \quad ||\mathbf{Y} - \mathbf{X}^T\mathbf{W}||^2 + \gamma||\mathbf{W}||_{21} \tag{14}$$

   The optimal solution of Eq. (14) can also be obtained from routine optimization tools, such as the iterative re-weighted method and the proximal method [15].

The complete algorithm to solve FSASL is summarized in algorithm 2.

Algorithm 2 The optimization algorithm of FSASL

Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, the regularization parameters α, β, γ, μ, the dimension of the transformed data c.

repeat

For each i, update the i-th column of $\mathbf{S}$ by solving the problem in Eq. (6);

For each i, update the i-th row of $\mathbf{P}$ using Algorithm 1;

Compute the overall graph Laplacian $\mathbf{L} = \mathbf{L_S} + \beta\mathbf{L_P}$;

Compute $\mathbf{W}$ by Eq. (12) or Eq. (14);

until Converges

Output: Sort all the d features according to $||\mathbf{w}_i||_2 (i = 1, ..., d)$ in descending order and select the top m ranked features.

3.5 Convergence Analysis

FSASL is solved in an alternative way, the optimization procedure will monotonically decrease the objective of the problem in Eq. (5) in each iteration. Since the objective function has lower bounds, such as zero, the above iteration converges. Besides, the experimental results show that it converges fast, the time of iteration is often less than 20.

3.6 The determination of parameter μ

Since the parameter μ is used to control the trade off between the trivial solution ($\mu = 0$) and the uniform distribution ($\mu = \infty$), we would like to keep only top-k neighbors for local manifold structure characterization as the k-nn graph [18]. Inspired by recent work on adaptive clustering in [18], we provide an effective method to achieve this. For each sub problem in Eq. (8), the Lagrangian function is

$$\frac{1}{2}||\mathbf{p}_i^T - \mathbf{a}_i^T||^2 - \tau(\mathbf{p}_i^T\mathbf{1}_n - 1) - \eta_i^T\mathbf{p}_i \tag{15}$$

where τ and η_i are the Lagrangian multipliers. According to KKT condition, the optimal value can be obtained by

$$\mathbf{P}_{ij} = (\mathbf{A}_{ij} + \tau)_+ \tag{16}$$

By sorting each row of $\mathbf{A}$ into $\mathbf{B}$ with ascending order, the following inequality holds

$$\begin{cases} B_{ik'} + \tau > 0 & \text{for } k' = 1, ..., k \\ B_{ik'} + \tau \leq 0 & \text{for } k' = k+1, ..., n \end{cases} \tag{17}$$

Considering the simplex constraint on $\mathbf{p}_i^T$, we further get

$$\tau = \frac{1}{k}(1 - \sum_{k'=1}^{k} B_{ik'}) \tag{18}$$

By replacing Eq. (18) into Eq. (16), the optimal value of $\mathbf{P}$ can be obtained by

$$\mathbf{P}_{ij} = (\mathbf{A}_{ij} - \frac{1}{k}(1 - \sum_{k'=1}^{k} B_{ik'}))_+ \tag{19}$$

Since $B_{ik'} = -\frac{1}{2\mu}||\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_{k'}||^2 = d_{ik'}^{\mathbf{W}}$, for each subproblem we have

$$\frac{k}{2}d_{ik}^{\mathbf{W}} - \frac{1}{2}\sum_{k'=1}^{k} d_{ik'}^{\mathbf{W}} < \mu \leq \frac{k}{2}d_{i,k'+1}^{\mathbf{W}} - \frac{1}{2}\sum_{k'=1}^{k} d_{ik'}^{\mathbf{W}} \tag{20}$$

When μ satisfies the above inequality for i-th example, the corresponding $\mathbf{p}_i^T$ has k non-zero component. Therefore the average non-zero elements in each row of $\mathbf{P}$ is close to k when we set

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}d_{i,k'+1}^{\mathbf{W}} - \frac{1}{2}\sum_{k'=1}^{k} d_{ik'}^{\mathbf{W}}\right) \tag{21}$$

In this way, the search of parameter μ can be better handled by searching the neighborhood size k, which is more intuitive and easy to tune.

4. DISCUSSION

In this section, we discuss some approaches which are closely related to our method.

Zeng and Cheung [32] proposed to integrate feature selection within the regularization of local learning-based clustering (LLCFS), which involves two sub steps:

1. It constructs the k-nearest neighbor graph in the weighted feature space.

2. It performs joint clustering and feature weight learning by solving the following problem

$$\min_{\mathbf{Y},\{\mathbf{W}^i,\mathbf{b}^i\}_{i=1}^n,\mathbf{z}} \sum_{i=1}^{n}\sum_{c'=1}^{c}\Bigg[\sum_{\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}}\beta(\mathbf{Y}_{ic'}-\mathbf{x}_j^T\mathbf{W}_{c'}^i-\mathbf{b}_{c'}^i)^2 + (\mathbf{W}_{c'}^i)^T\mathrm{diag}(\mathbf{z}^{-1})\mathbf{W}_{c'}^i\Bigg] \quad (22)$$
$$\text{s.t.} \quad \mathbf{1}_d^T\boldsymbol{z}=1,\mathbf{z}\geq 0$$

where $\mathbf{z}$ is the feature weight vector and $\mathcal{N}_{\mathbf{x}_i}$ is the k-nearest neighbor of $\mathbf{x}_i$ based on $\mathbf{z}$ weighted features.

Compared with LLCFS, FSASL performs both the global and local structure learning in an adaptive manner, where only local structure is explored by LLCFS. Moreover, LLCFS uses the discrete k-nearest graph and does not optimize the same objective function in structure learning and feature search, while FSASL is optimized within a unified framework with the probabilistic neighborhood relationship.

Hou et al. [10] proposed the joint embedding learning and sparse regression (JELSR) method, which can be formulated as follows:

$$\min_{\mathbf{W},\mathbf{Y}^T\mathbf{Y}=\mathbf{I}} \mathrm{tr}(\mathbf{Y}^T\mathbf{L}_2\mathbf{Y})+\lambda_1(||\mathbf{Y}-\mathbf{X}^T\mathbf{W}||^2+\lambda_2||\mathbf{W}||_{21}) \quad (23)$$

Comparing the formulation in Eq. (5) and Eq. (23), the main differences between FSASL and JELSR include: 1) FSASL selects those features to respect both the global and local manifold structure, while JELSR only incorporates the local manifold structure; 2) The local structure in JELSR is based on k-nearest neighbor graph, while FSASL learns a probabilistic neighborhood graph, which can be easily refined according the result of feature selection. 3)JELSR iteratively perform spectral embedding for clustering and sparse spectral regression for feature selection as illustrated in Fig. (1). However, the local structure itself (i.e. $\mathbf{L}_2$) is not changed during iterations. FSASL can adaptively improve both the global and local structure characterization using selected features.

Most recently, Liu et al. [16] proposed a global and local structure preservation framework for feature selection (GLSPFS). It first constructs the pairwise sample similarity matrix $\mathbf{K}$ with Gaussian kernel function to capture the global structure of data, then decompose $\mathbf{K}=\mathbf{Y}\mathbf{Y}^T$. Using $\mathbf{Y}$ as the regression target, GLSPFS solve the following problem:

$$\min_{\mathbf{W}} ||\mathbf{Y}-\mathbf{X}^T\mathbf{W}||^2+\lambda_1\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}_3\mathbf{X}^T\mathbf{W})+\lambda_2||\mathbf{W}||_{21} \quad (24)$$

The main differences between FSASL and GLSPFS include: 1) GLSPFS uses the Gaussian kernel, while FSASL captures the global structure within sparse representation, which is more discriminant; 2) Both the global and local structures (i.e. $\mathbf{K}$ and $\mathbf{L}_3$) in GLSPFS are based on all features, while FSASL refines these structures with selected features.

5. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of the proposed FSASL for the task of unsupervised feature selection.

5.1 Data Sets

The experiments are conducted on 8 publicly available datasets, including handwritten and spoken digit/letter recognition data sets (i.e., MFEA from UCI reporsitory and USPS49 [32] which is a two class subset of USPS), three face image data sets (i.e., UMIST [10], JAFFE [14, 23], AR [30]), one object data set (i.e. COIL [4, 5]) and one biomedical data sets (i.e., LUNG [17], TOX). The details of these benchmark data sets are summarized in Table 3.

Table 3: Summary of the benchmark data sets and the number of selected features

Data Sets	sample	feature	class	selected features
MFEA	2000	240	10	$[5,10,\ldots,50]$
USPS49	1673	256	2	$[5,10,\ldots,50]$
UMIST	575	644	20	$[5,10,\ldots,50]$
JAFFE	213	676	10	$[5,10,\ldots,50]$
AR	840	768	120	$[5,10,\ldots,50]$
COIL	1440	1024	20	$[5,10,\ldots,50]$
LUNG	203	3312	5	$[10,20,\ldots,150]$
TOX	171	5748	4	$[10,20,\ldots,150]$

5.2 Experiment Setup

To validate the effectiveness of our proposed FSASL[1] , we compare it with one baseline (i.e., AllFea) and states-of-the-art unsupervised feature selection methods,

- LapScore[2] [8], which evaluates the features according to their ability of locality preserving of the data manifold structure.
- MCFS[3] [4], which selects the features by adopting spectral regression with ℓ_1-norm regularization.
- LLCFS [32], which incorporates the relevance of each feature into the built-in regularization of the local learning-based clustering algorithm.
- UDFS[4] [28], which exploits local discriminative information and feature correlations simultaneously.
- NDFS[5] [14], which selects features by a joint framework of nonnegative spectral analysis and $\ell_{2,1}$-norm regularized regression.
- SPFS[6] [36], which selects a feature subset with which the pairwise similarity between high dimensional samples can be maximally preserved.
- RUFS[7] [21], which performs robust clustering and robust feature selection simultaneously to select the most important and discriminative features.

[1]For the purpose of reproducibility, we provide the code at https://github.com/csliangdu/FSASL
[2]http://www.cad.zju.edu.cn/home/dengcai/Data/code/LaplacianScore.m
[3]http://www.cad.zju.edu.cn/home/dengcai/Data/code/MCFS_p.m
[4]http://www.cs.cmu.edu/~yiyang/UDFS.rar
[5]https://sites.google.com/site/zcliustc
[6]https://sites.google.com/site/alanzhao
[7]https://sites.google.com/site/qianmingjie

Table 1: Aggregated clustering results measured by Accuracy (%) of the compared methods.

Data Sets	AllFea	LapScore	MCFS	LLCFS	UDFS	NDFS	SPFS	RUFS	JELSR	GLSPFS	FSASL
MFEA	68.73	51.78 ± 5.51 0.00	51.04 ± 8.13 0.00	60.38 ± 8.58 0.00	48.94 ± 3.32 0.00	67.13 ± 7.53 0.01	**68.20 ± 9.43 0.22**	64.58 ± 7.99 0.00	67.01 ± 8.37 0.01	61.00 ± 8.70 0.00	**69.94 ± 7.19 1.00**
USPS49	77.70	69.21 ± 8.95 0.00	53.74 ± 3.50 0.00	94.96 ± 1.44 0.03	94.05 ± 1.13 0.00	68.12 ± 8.18 0.00	83.43 ± 6.66 0.00	85.86 ± 2.58 0.00	95.16 ± 0.55 0.00	94.75 ± 0.61 0.00	**95.95 ± 0.48 1.00**
UMIST	42.40	36.73 ± 1.18 0.00	44.46 ± 3.26 0.00	47.31 ± 0.83 0.00	48.04 ± 1.92 0.00	52.80 ± 2.26 0.00	46.72 ± 1.70 0.00	50.87 ± 1.95 0.00	53.52 ± 1.54 0.01	50.53 ± 0.59 0.00	**54.92 ± 1.89 1.00**
JAFFE	71.57	67.62 ± 8.49 0.00	73.56 ± 4.83 0.00	64.79 ± 4.08 0.00	75.48 ± 1.63 0.00	74.98 ± 2.15 0.00	73.93 ± 2.85 0.00	75.75 ± 2.53 0.00	77.77 ± 1.87 0.00	75.46 ± 1.61 0.00	**79.29 ± 2.24 1.00**
AR	30.26	25.29 ± 2.89 0.00	29.05 ± 1.19 0.00	**34.22 ± 2.70 0.05**	30.87 ± 0.35 0.00	32.34 ± 1.52 0.00	31.06 ± 2.14 0.00	34.84 ± 1.90 0.04	34.19 ± 2.52 0.02	34.12 ± 1.60 0.00	**36.11 ± 0.75 1.00**
COIL	59.17	45.60 ± 6.16 0.00	51.50 ± 5.38 0.00	50.84 ± 3.76 0.00	31.40 ± 16.89 0.00	44.22 ± 6.33 0.00	56.94 ± 3.43 0.00	59.20 ± 3.28 0.00	59.53 ± 4.01 0.03	57.96 ± 2.27 0.00	**60.93 ± 2.50 1.00**
LUNG	72.46	58.97 ± 5.24 0.00	70.42 ± 3.41 0.00	71.58 ± 5.85 0.00	65.46 ± 3.88 0.00	75.52 ± 1.57 0.00	73.49 ± 3.43 0.00	77.35 ± 2.62 0.00	77.86 ± 3.12 0.00	77.83 ± 2.70 0.00	**81.93 ± 1.63 1.00**
TOX	43.65	40.25 ± 0.65 0.00	43.10 ± 1.86 0.00	39.28 ± 0.49 0.00	47.14 ± 0.75 0.00	38.28 ± 1.64 0.00	39.93 ± 1.13 0.00	47.67 ± 0.83 0.00	43.96 ± 1.56 0.00	47.38 ± 1.93 0.00	**49.17 ± 0.67 1.00**
Average	58.24	49.43	52.11	57.92	55.17	56.67	59.21	62.02	63.63	62.38	**66.03**

- JELSR[8] [11, 10], which joins embedding learning with sparse regression to perform feature selection.

- GLSPFS[9] [16], which integrates both global pairwise sample similarity and local geometric data structure to conduct feature selection.

There are some parameters to be set in advance. For all the feature selection algorithms except SPFS, we set $k = 5$ for all the datasets to specify the size of neighborhoods [4, 13]. The weight of k-nn graph for LapScore and MCFS, and the pairwise similarity for SPFS and GLSPFS is based on the Gaussian kernel, where the kernel width is searched from the grid $\{2^{-3}, 2^{-2}, \ldots, 2^{3}\}\delta_0$, where δ_0 is the mean distance between any two data examples. For GLSPFS, we report the best results among three local manifold models, that is locality preserving projection (LPP), LLE and local tangent space alignment (LTSA) as in [16]. For LLCFS, UDFS, NDFS, RUFS, JELSR, GLSPFS and FSASL, the regularization parameters are searched from the grid $\{10^{-5}, 10^{-4}, \ldots, 10^{5}\}$. And the regularization parameter for γ is searched from the grid $\{0.001, 0.005, 0.01, 0.05, 0.1\}\gamma_{max}$, where γ_{max} is automatically computed from SLEP [15]. For FSASL, μ is determined by Eq. (21) with $k = 5$ and c is set to be the true number of classes. To fairly compare different unsupervised feature selection algorithms, we tune the parameters for all methods by the grid-search strategy [21, 16, 5].

With the selected features, we evaluate the performance in terms of k-means clustering by two widely used metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI). The results of k-means clustering depend on the initialization. For all the compared algorithms with different parameters and different number of selected features, we first repeat the clustering 20 times with random initialization and record the average results.

[8] `http://www.escience.cn/people/chenpinghou`

[9] We also use the implementation provided by the authors.

5.3 Clustering with Selected Features

Since the optimal number of selected features is unknown in advance, to better evaluate the performance of unsupervised feature selection algorithms, we finally report the averaged results over different number of selected features (the range of selected features for each data set can be found in Table 3) with standard derivation. For all the algorithms (except for AllFea), we also report its p-value by the paired t-test against the best results. The best one and those having no significant difference ($p > 0.05$) from the best one are marked in bold.

The clustering results in terms of ACC and NMI are reported in Table 1 and Table 2, respectively. For different feature selection algorithms, the results in each cell of Table 1 and 2 are the mean ± standard deviation and the p-value. The last row of Table 1 and Table 2 shows the averaged results of all the algorithms over the 8 datasets.

Compared with clustering using all features, these unsupervised feature selection algorithms not only can largely reduce the number of features facilitating the latter learning process, but can also often improve the clustering performance. In particular, our method FSASL achieves 11.8% and 15.04% improvement in terms of accuracy and NMI respectively with less than 10% features. These results can well demonstrate the effectiveness and efficiency of unsupervised feature selection algorithm. It can also be observed

Table 2: Aggregated clustering results measured by Normalized Mutual Information (%) of the compared methods.

Data Sets	AllFea	LapScore	MCFS	LLCFS	UDFS	NDFS	SPFS	RUFS	JELSR	GLSPFS	FSASL
MFEA	70.33	53.74 ± 4.77 0.00	54.72 ± 9.14 0.00	52.77 ± 9.76 0.00	49.19 ± 3.83 0.00	64.97 ± 7.54 0.03	**64.92 ± 8.27 0.11**	63.98 ± 7.22 0.00	**64.51 ± 9.07 0.06**	59.26 ± 7.59 0.00	**66.70 ± 6.71 1.00**
USPS49	23.51	15.88 ± 17.98 0.00	4.60 ± 2.57 0.00	72.03 ± 5.56 0.03	68.12 ± 4.46 0.00	12.27 ± 9.62 0.00	38.10 ± 16.66 0.00	41.73 ± 7.23 0.00	72.28 ± 2.24 0.00	70.43 ± 2.57 0.00	**75.88 ± 2.28 1.00**
UMIST	64.15	55.57 ± 2.32 0.00	63.46 ± 4.93 0.00	63.42 ± 1.42 0.00	65.19 ± 2.96 0.00	71.19 ± 2.77 0.01	64.90 ± 3.06 0.00	68.19 ± 2.61 0.00	71.33 ± 2.06 0.00	69.16 ± 0.97 0.00	**72.39 ± 2.39 1.00**
JAFFE	81.52	77.28 ± 8.98 0.00	79.04 ± 5.88 0.00	66.97 ± 3.47 0.00	84.25 ± 1.74 0.00	82.53 ± 3.49 0.00	80.01 ± 3.06 0.00	82.00 ± 3.56 0.00	85.23 ± 3.31 0.00	83.20 ± 3.17 0.00	**86.42 ± 3.34 1.00**
AR	65.48	63.59 ± 2.36 0.00	66.41 ± 0.85 0.00	69.01 ± 1.45 0.01	67.49 ± 0.27 0.00	67.89 ± 0.89 0.00	66.94 ± 1.11 0.00	69.54 ± 1.10 0.01	69.02 ± 1.32 0.00	69.44 ± 0.84 0.00	**70.78 ± 0.63 1.00**
COIL	75.58	62.21 ± 4.98 0.00	66.19 ± 6.78 0.00	64.04 ± 4.34 0.00	44.27 ± 12.61 0.00	56.29 ± 6.91 0.00	69.91 ± 4.38 0.00	70.54 ± 4.48 0.00	71.37 ± 4.97 0.00	69.89 ± 4.00 0.00	**72.93 ± 4.44 1.00**
LUNG	60.37	50.14 ± 4.13 0.00	55.68 ± 2.31 0.00	60.12 ± 4.65 0.00	54.88 ± 4.21 0.00	60.57 ± 1.54 0.00	61.75 ± 3.32 0.00	65.47 ± 1.87 0.00	63.54 ± 2.94 0.00	63.50 ± 2.99 0.00	**66.78 ± 1.72 1.00**
TOX	15.87	10.92 ± 0.68 0.00	16.53 ± 2.68 0.00	9.68 ± 0.75 0.00	22.16 ± 1.36 0.00	9.07 ± 1.87 0.00	10.13 ± 1.03 0.00	23.58 ± 1.60 0.00	17.46 ± 3.36 0.00	23.49 ± 2.77 0.00	**25.79 ± 1.62 1.00**
Average	57.10	48.67	50.83	57.26	56.94	53.10	57.08	60.63	64.34	63.55	**67.21**

that FSASL consistently produces better performance than the other nine feature selection algorithms, and the improvement is in the range from 3.63% to 25.14% in terms of clustering accuracy and from 4.27% to 27.59% in terms of NMI. This can be mainly explained by the following reasons. First, both global and local structure are used to guide the search of relevant features. Second, the structure learning and feature selection are integrated into a unified framework. Third, both the global and local structures can be adaptively updated using the results of selected features.

5.4 Effect of Adaptive Structure Learning

Here, we investigate the effect of adaptive structure learning by empirically answering the following questions:

1. What kind of structure should be captured and preserved by the selected features, either global or local or both of these structures?

2. Does the adaptive structure learning lead to select more informative features?

We conduct different settings of FSASL on USPS200, which consists the first 100 samples in USPS49. We solve the optimization problem in Eq. (2), Eq. (4) and Eq. (5), which uses global, local, and both global and local structures, respectively. We also distinguish these problems with and without adaptive structure learning. Thus, we have 6 settings in total. Figure 3 and Figure 4 show the results of these different settings with different number of selected features. The aggregated result over different number of selected features is also provided in Table 4.

From these results, we can see that: 1) The exploitation of both global and local structures (i.e., Eq. (5) + $\mathbf{W}$) outperform another two alternatives with only global (i.e., Eq. (2) + $\mathbf{W}$) or local (i.e., Eq. (4) + $\mathbf{W}$) structure. It validates that the integration of both global and local structure is better than the single one. 2) With the update of structure learning (i.e., Eq. (2) + $\mathbf{W}, \mathbf{S}$, Eq. (4) + $\mathbf{W}, \mathbf{P}$ and Eq. (5) + $\mathbf{W}, \mathbf{S}, \mathbf{P}$) is better than their counterparts without adaptive structure learning respectively. It shows that the adaptive learning in either global and/or local structure learning can further improve the result of feature selection.

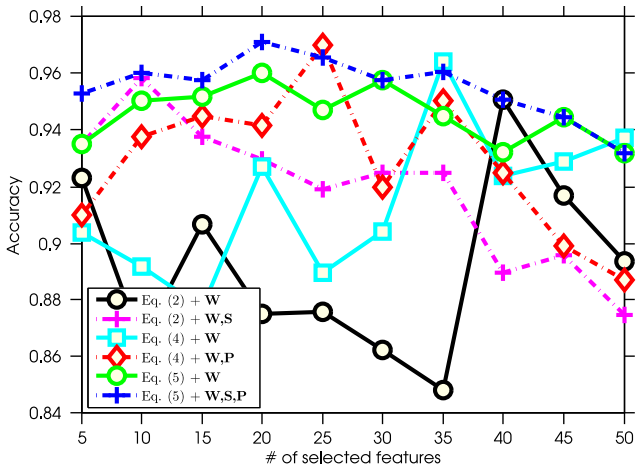

Figure 3: Clustering accuracy w.r.t. 6 different settings of FSASL on USPS200.

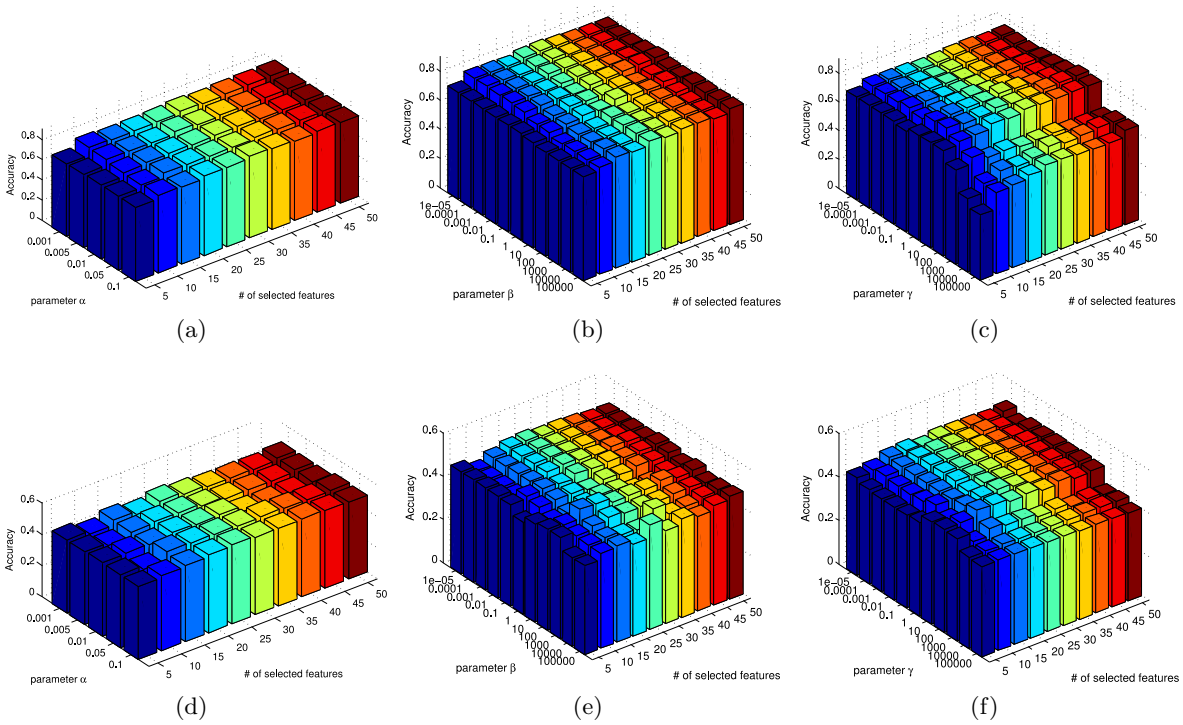

Figure 2: Clustering accuracy w.r.t. different parameters on JAFFE (a-c) and TOX (d-f).

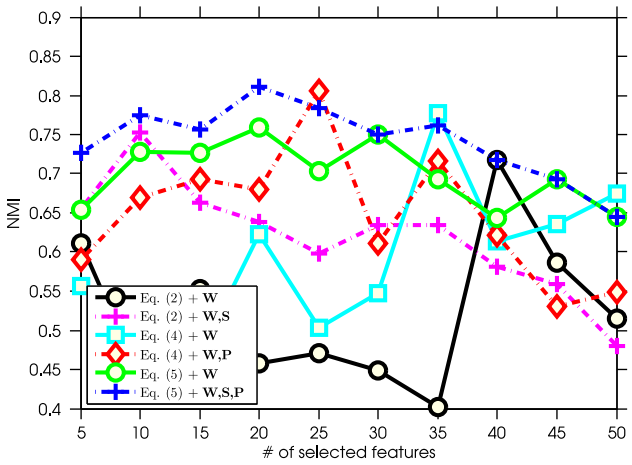

Figure 4: Clustering NMI w.r.t. 6 different settings of F-SASL on USPS200.

In the next, we investigate the sensitivity with respect to the regularization parameters α, β and γ.

5.5 Parameter Sensitivity

When we vary the value of one parameter, we keep the other parameters fixed at the optimal value. We plot the clustering accuracy with respect to these parameters on JAFFE and TOX in Figure 2. The experimental results show that our method is not very sensitive to α, β and γ with wide ranges. However, the performance is relatively sensitive to the number of selected features, which is still an open problem.

Table 4: Aggregated clustering results (%) of 6 different settings of FSASL on USPS200.

Problem	Variables	ACC	NMI
Eq. (2)	$\mathbf{W}$	89.17 ± 3.22	52.01 ± 9.69
Eq. (2)	$\mathbf{W}, \mathbf{S}$	91.90 ± 2.51	61.95 ± 7.21
Eq. (4)	$\mathbf{W}$	91.48 ± 2.62	59.10 ± 9.31
Eq. (4)	$\mathbf{W}, \mathbf{P}$	92.86 ± 2.53	64.65 ± 8.30
Eq. (5)	$\mathbf{W}$	94.65 ± 1.24	69.94 ± 4.22
Eq. (5)	$\mathbf{W}, \mathbf{S}, \mathbf{P}$	95.53 ± 1.10	74.20 ± 4.83

6. CONCLUSION

In this paper, we proposed a novel unsupervised feature selection method to simultaneously perform feature selection and the structure learning. In our new method, the global structure learning and feature selection are integrated within the framework of sparse representation; the local structure learning and feature selection are incorporated into the probabilistic neighborhood relationship learning framework. By combining both the global and local structure learning and feature selection, our method can boost both these two essential tasks, i.e., structure learning and feature selection, by using the result of the other task. We derive an efficient algorithm to optimize the proposed method and discuss the connections between our method and other feature selection methods. Extensive experiments have been conducted on real-world benchmark data sets to demonstrate the superior performance of our method.

7. ACKNOWLEDGMENTS

We would like to thank Prof. Feiping Nie and Prof. Mingyu Fan for their helpful suggestions to improve this paper. This work is supported in part by the China National 973 program 2014CB340301, the Natural Science Foundation of China (NSFC) grant 61379043, 61322211 and Program for New Century Excellent Talents in University (No. NCET-12-1031).

8. REFERENCES

[1] S. Alelyani, J. Tang, and H. Liu. Feature selection for clustering: A review. *Data Clustering: Algorithms and Applications*, 29, 2013.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[3] D. Cai, X. He, and J. Han. Spectral regression: A unified approach for sparse subspace learning. In *ICDM*, pages 73–82, 2007.

[4] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *SIGKDD*, pages 333–342, 2010.

[5] L. Du, Z. Shen, X. Li, P. Zhou, and Y. Shen. Local and global discriminative learning for unsupervised feature selection. In *ICDM*, pages 131–140, 2013.

[6] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *JMLR*, 5:845–889, 2004.

[7] Q. Gu, Z. Li, and J. Han. Joint feature selection and subspace learning. In *IJCAI*, pages 1294–1299, 2011.

[8] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *NIPS*, 18:507–514, 2006.

[9] X. He, M. Ji, C. Zhang, and H. Bao. A variance minimization criterion to feature selection using laplacian regularization. *PAMI*, (99):2013–2025, 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. *Cybernetics, IEEE Transactions on*, 44(6):793–804, 2014.

[11] C. Hou, F. Nie, D. Yi, and Y. Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI*, pages 1324–1329, 2011.

[12] W. Krzanowski. Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics*, pages 22–33, 1987.

[13] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu. Clustering-guided sparse structural learning for unsupervised feature selection. *TKDE*, 26(9):2138–2150, 2014.

[14] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.

[15] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[16] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu. Global and local structure preservation for feature selection. *IEEE Transactions on NNLS*, 25(6):1083–1095, 2014.

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

[18] F. Nie, X. Wang, and H. Huang. Clustering and projected clustering with adaptive neighbors. In *SIGKDD*, pages 977–986, 2014.

[19] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan. Trace ratio criterion for feature selection. In *IJCAI*, volume 2, pages 671–676, 2008.

[20] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 27(8):1226–1238, 2005.

[21] M. Qian and C. Zhai. Robust unsupervised feature selection. In *IJCAI*, pages 1621–1627, 2013.

[22] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[23] L. Shi, L. Du, and Y. Shen. Robust spectral learning for unsupervised feature selection. In *ICDM*, pages 977–982, 2014.

[24] I. Takeuchi and M. Sugiyama. Target neighbor consistent feature weighting for nearest neighbor classification. In *NIPS*, pages 576–584, 2011.

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

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009.

[27] Z. Xu, I. King, M.-T. Lyu, and R. Jin. Discriminative semi-supervised feature selection via manifold regularization. *Neural Networks, IEEE Transactions on*, 21(7):1033–1047, 2010.

[28] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. ℓ_{21}-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.

[29] Y. Yang, H. T. Shen, F. Nie, R. Ji, and X. Zhou. Nonnegative spectral clustering with discriminative regularization. In *AAAI*, 2011.

[30] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image clustering using local discriminant models and global integration. *TIP*, 19(10):2761–2773, 2010.

[31] Y. Yang, Y. Yang, H. T. Shen, Y. Zhang, X. Du, and X. Zhou. Discriminative nonnegative spectral clustering with out-of-sample extension. *TKDE*, 25(8):1760–1771, 2013.

[32] H. Zeng and Y. Cheung. Feature selection and kernel learning for local learning-based clustering. *PAMI*, 33(8):1532–1547, 2011.

[33] Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In *SDM*, pages 641–646, 2007.

[34] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, pages 1151–1157, 2007.

[35] Z. Zhao, L. Wang, and H. Liu. Efficient spectral feature selection with minimum redundancy. In *AAAI*, pages 673–678, 2010.

[36] Z. Zhao, L. Wang, H. Liu, and J. Ye. On similarity preserving feature selection. *TKDE*, 25(3):619–632, 2013.