Unsupervised Feature Selection with Structured Graph Optimization*

Feiping Nie[1] and **Wei Zhu**[1] and **Xuelong Li**[2]
[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P.R. China
[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an, 710119, Shaanxi, P. R. China
{feipingnie@gmail.com, zwvews@gmail.com, xuelong_li@opt.ac.cn}

Abstract

Since amounts of unlabelled and high-dimensional data needed to be processed, unsupervised feature selection has become an important and challenging problem in machine learning. Conventional embedded unsupervised methods always need to construct the similarity matrix, which makes the selected features highly depend on the learned structure. However real world data always contain lots of noise samples and features that make the similarity matrix obtained by original data can't be fully relied. We propose an unsupervised feature selection approach which performs feature selection and local structure learning simultaneously, the similarity matrix thus can be determined adaptively. Moreover, we constrain the similarity matrix to make it contain more accurate information of data structure, thus the proposed approach can select more valuable features. An efficient and simple algorithm is derived to optimize the problem. Experiments on various benchmark data sets, including handwritten digit data, face image data and biomedical data, validate the effectiveness of the proposed approach.

Introduction

Due to large amounts of data produced by rapid development of technology, the processing of high-dimensional data has become a big problem in many fields, such as computer vision, data mining, and pattern recognition. High-dimensional data often contain quite a lot noise features, which are detrimental to data processing. As one of the typical method to alleviate this problem, feature selection attracts more and more attentions. Feature selection aims at obtaining a subset of features which are selected from original feature space. Feature selection mainly uses three approaches: supervised, semi-supervised and unsupervised. Obviously, unsupervised feature selection is more difficult than the others for absence of label information, however, the large amount of unlabelled data makes the unsupervised feature selection practical.

Various methods of unsupervised feature selection have been proposed, and can be classified into three distinct types, i.e. filter method (He, Cai, and Niyogi 2005; Zhao and Liu 2007), wrapper method (Tabakhi, Moradi, and Akhlaghian 2014), and embedded method (Cai, Zhang, and He 2010; Zhao, Wang, and Liu 2010; Hou et al. 2014; Li et al. 2012; Wang, Tang, and Liu 2015). Embedded methods are superior to others in many respects, and have received more and more attentions. Since local manifold structure is considered better than global structure, most of embedded methods try to discover local structure. The classical graph based methods contain two independent steps. First, construct similarity matrix by exploring the local structure. Then, select valuable features by sparsity regularization.

However, there are at least two issues for conventional embedded methods. First, conventional spectral based feature selection methods construct similarity matrix and select features independently. Therefore, the similarity matrix is derived from original data and remains constant for the subsequent process, but real world data always contain lots of noise samples and features, which make the similarity matrix unreliable (Wang, Nie, and Huang 2015). The unreliable similarity matrix will damage the local manifold structure, and ultimately lead to suboptimal result. Second, similarity matrix obtained by conventional method is usually not an ideal neighbor assignment. The optimal similarity matrix should have exact c connected components, where c is the number of cluster. However, simply using k-nearest neighbors assignment hardly achieves the ideal state.

To mitigate the impact of above problems, we propose an unsupervised feature selection approach, called Structured Optimal Graph Feature Selection (SOGFS). It is worthwhile to highlight the main contributions of this paper as follows:

1. A novel embedded unsupervised feature selection approach is proposed. The approach performs feature selection and local structure learning simultaneously. It adaptively learns local manifold structure, and thus can select more valuable features.
2. A reasonable constraint is introduced to the approach. The similarity matrix obtained by local structure can be more accurate when we constrain similarity matrix to make it contain exact c connected components.
3. Comprehensive experiments on benchmark data sets show the effectiveness of the proposed approach, and demonstrate the advantage over other state-of-the-art methods.

*This work is supported by the Fundamental Research Funds for the Central Universities (Grant no. 3102015BJ(II)JJZ01).

Related Work

In this section, we briefly review recent studies of embedded unsupervised feature selection. Spectral analysis is the most common technique used in embedded methods. Cai, Zhang, and He (2010) propose Multi-Cluster Feature Selection (MCFS), MCFS captures the local manifold structure via spectral analysis, and then selects the features which can best preserve the clustering structure. Flexible Manifold Embedding (FME) (Nie et al. 2010b) is proposed as a general framework for dimensionality reduction, and has also been adopted by many feature selection methods. Based on FME and ℓ-2,1 norm, Hou et al. (2014) propose a general framework for feature selection termed as Joint Embedding Learning and Sparse Regression (JELSR). Li et al. (2012) jointly use FME, non-negative constraint, and ℓ-2,1 norm to perform Nonnegative Discriminative Feature Selection (NDFS). Qian and Zhai (2013) propose Robust Unsupervised Feature Selection (RUFS). RUFS uses FME, Nonnegative Matrix Factorization (NMF) and ℓ-2,1 norm to perform robust clustering and robust feature selection simultaneously. Robust Spectral Feature Selection (RSFS), which is proposed by Shi, Du, and Shen (2014), also can be considered as jointly using FME and ℓ-1 norm to perform robust feature selection. However, as previously mentioned, almost all of these methods have at least two problems, i.e. unreliable similarity matrix and improper neighbor assignment. These problems make the similarity matrix can't be fully relied, and eventually lead to suboptimal result.

Methodology

In this section, we introduce the proposed approach SOGFS by first formulating the optimization problem, and then presenting an efficient algorithm to tackle it.

We first introduce some notations that are used throughout the paper. For matrix $M \in \mathbb{R}^{r\times t}$, the (i,j)-th entry of M is denoted by m_{ij}, the transpose of the i-th row of M is denoted by $m_i \in \mathbb{R}^{t\times 1}$. The trace of M is denoted by $Tr(M)$. The transpose of matrix M is denoted by M^T. The F-norm of M is denoted by $\|M\|_F$. The $\ell_{2,1}$-norm of M is denoted by $\|M\|_{2,1} = \sum_{i=1}^{r}\sqrt{\sum_{j=1}^{t} m_{ij}^2}$.

Local Structure Learning

Inspired by the development of spectral analysis and manifold learning, many unsupervised feature selection methods try to preserve local manifold structure, which is believed better than global structure. Therefore, similarity matrix is crucial for the ultimate performance of spectral methods. Nevertheless, most methods construct similarity matrix simply from original features which contain many redundant and noise samples or features. This will inevitably damage the learned structure, and the similarity matrix is surely unreliable and inaccurate. Thus, in this paper, we apply an adaptive process to determine the similarity matrix with probabilistic neighbors (Nie, Wang, and Huang 2014) through the algorithm. In other words, we perform feature selection and local structure learning simultaneously.

Given a data matrix $X \in \mathbb{R}^{n\times d}$, where $x_i \in \mathbb{R}^{d\times 1}$ denotes the i-th sample. We adopt the data preprocessing proposed by Liu et al. (2013) and define that x_i can be connected by all the others with probability s_{ij}, where s_{ij} is an element of similarity matrix $S \in \mathbb{R}^{n\times n}$. The probability of two data to be neighbor can be regarded as the similarity between them. According to the common sense, closer samples are likely to have larger probability to connect, thus s_{ij} is inversely proportional to the distance between x_i and x_j (we adopt the square of Euclidean distance for simplicity, i.e. $\|x_i - x_j\|_2^2$). Therefore, determining the value of the probability s_{ij} can be seen as solving

$$\min_{s_i^T\mathbf{1}=1, 0\le s_{ij}\le 1} \sum_{i,j} (\|x_i - x_j\|_2^2 s_{ij} + \alpha s_{ij}^2), \qquad (1)$$

where α is the regularization parameter. Regularization term is used to avoid the trivial solution. Without the regularization term, the optimal solution is that the two samples which are nearest to each other should be neighbor with probability 1. We show how to determine α later.

Structured Optimal Graph

The ideal state of neighbor assignment is that similarity matrix contains exact c connected components, such assignment is obviously beneficial for subsequent processing. However, the similarity matrix S obtained by the solution of problem (1) is virtually impossible to be in such state, in most cases, there is only one connected component (Nie, Wang, and Huang 2014). Next, we show an efficient yet simple method to tackle this problem.

There is a basic but important equation in spectral analysis

$$\sum_{i,j} \|f_i - f_j\|_2^2 s_{ij} = 2Tr(F^T L_S F), \qquad (2)$$

where $F \in \mathbb{R}^{n\times c}$, and $L_S = D - \frac{S^T+S}{2}$ is called Laplacian Matrix, the degree matrix D is a diagonal matrix and the i-th entry is defined as $\sum_j \frac{(s_{ij}+s_{ji})}{2}$.

It can be proved that if $rank(L_S) = n - c$, the similarity matrix S will contain exact c connected components (Mohar et al. 1991). Thus we add the constraint to problem (1), then we have

$$\begin{aligned} &\min \sum_{i,j} (\|x_i - x_j\|_2^2 s_{ij} + \alpha s_{ij}^2). \\ &s.t. \quad \forall i, s_i^T\mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c \end{aligned} \qquad (3)$$

Because the constraint $rank(L_S) = n - c$ also depends on similarity matrix S, the problem (3) is difficult to solve. To tackle it, let $\sigma_i(L_S)$ denote the i-th smallest eigenvalue of L_S. As L_S is positive semi-definite, we get $\sigma_i(L_S) \ge 0$. It can be verified that $rank(L_S) = n - c$ indicates $\sum_{i=1}^{c}\sigma_i(L_S) = 0$. Due to the deviation of $\sum_{i=1}^{c}\sigma_i(L_S)$ is difficult to handle, considering of Ky Fan's Theorem (Fan 1949), we have

$$\sum_{i=1}^{c}\sigma_i(L_S) = \min_{F\in\mathbb{R}^{n\times c}, F^TF=I} Tr(F^T L_S F). \qquad (4)$$

Thus, we can rewrite problem (3) as (Nie, Wang, and Huang 2014; Wang et al. 2015)

$$\begin{aligned} &\min \sum_{i,j} (\|x_i - x_j\|_2^2 s_{ij} + \alpha s_{ij}^2) + 2\lambda Tr(F^T L_S F). \\ &s.t. \quad \forall i, s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, F \in \mathbb{R}^{n \times c}, \\ &\quad F^T F = I \end{aligned} \tag{5}$$

Obviously, as long as λ is large enough, the $Tr(F^T L_S F)$ will come infinitely close to zero. In fact, in each iteration, we can increase or decrease λ when the connected components are smaller or greater than c, respectively. Through transforming the constraint of matrix rank into trace, problem (5) is much easier to tackle than original one. By solving problem (5), S contains exact c connected components, thus captures more accurate information of local structure.

Structured Optimal Graph Feature Selection

According to the theory of manifold learning, there always exists a low-dimensional manifold that can express the structure of high-dimensional data. First, we aim at finding a linear combination of original features to best approximate the low-dimension manifold. Denote XW as this linear combination, where $W \in \mathbb{R}^{d \times m}$ is the projection matrix, d and m are the original dimension and projection dimension respectively. We apply it to Eq. (5), and finally we get

$$\begin{aligned} &\min \sum_{i,j} (\|W^T x_i - W^T x_j\|_2^2 s_{ij} + \alpha s_{ij}^2) \\ &\quad + \gamma \|W\|_{2,1} + 2\lambda Tr(F^T L_S F), \\ &s.t. \quad \forall i, s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, F \in \mathbb{R}^{n \times c}, \\ &\quad F^T F = I, W^T W = I \end{aligned} \tag{6}$$

where γ is the regularization parameter. The sparsity regularization on W makes it suitable for selecting valuable features. We adopt $\ell_{2,1}$-norm regularization (Nie et al. 2010a) on W to make it row sparse.

The importance of the i-th feature can be measured by $\|w_i\|_2$. The most important h features are selected by the sorted $\|w_i\|_2$, where h is the number of features that need to be selected. In feature selection tasks, the feature number of data is usually very high and we can't use PCA as data preprocessing, thus the covariance matrix of X, which is denoted by S_t, is often singular in practice. Therefore, even though the constraint $W^T S_t W = I$ is often adopted in feature extraction (Nie, Wang, and Huang 2014), it is by no means a good choice for feature selection. In this paper, we adopt the constraint $W^T W = I$ instead of $W^T S_t W = I$ to make the feature space distinctive after reduction (Wang, Nie, and Huang 2014). W is used for selecting features and S is used to capture local structure, thus the proposed approach performs feature selection and local structure learning simultaneously.

Optimization Algorithm

Since problem (6) contains ℓ-2,1 norm regularization and three different variables, it is hard to tackle it directly. Thus, we propose an alternative iterative algorithm to solve this problem.

Fix S update W

With fixed S, the problem (6) is transformed into

$$\min_{W^T W = I} \sum_{i,j} \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma \|W\|_{2,1}. \tag{7}$$

According to Eq. (2), and replace $\|W\|_{2,1}$ with $\sum_i \|w_i\|_2$, we rewrite problem (7) as

$$\min_{W^T W = I} Tr(W^T X^T L_S X W) + \gamma \sum_i \|w_i\|_2. \tag{8}$$

Obviously, $\|w_i\|_2$ can be zero in theory, however, it will make Eq. (8) non-differentiable. To avoid this condition, we transform $\|w_i\|_2$ into $\sqrt{(w_i)^T w_i}$, and regularize $\sqrt{(w_i)^T w_i}$ as $\sqrt{(w_i)^T w_i + \varepsilon}$, where ε is a small enough constant. Therefore, we have

$$\min_{W^T W = I} Tr(W^T X^T L_S X W) + \gamma \sum_i \sqrt{(w_i)^T w_i + \varepsilon}, \tag{9}$$

which is apparently equal to problem (7) when ε is infinitely close to zero. The Lagrangian function of the problem (9) is

$$\begin{aligned} \mathcal{L}(W, \Lambda) =& Tr(W^T X^T L_S X W) + \gamma \sum_i \sqrt{(w_i)^T w_i + \varepsilon} \\ &+ Tr(\Lambda(W^T W - I)), \end{aligned} \tag{10}$$

where Λ is the Lagrangian multiplier. Taking the derivative of $\mathcal{L}(W, \Lambda)$ w.r.t W, and setting the derivative to zero, we have

$$\frac{\partial \mathcal{L}(W, \Lambda)}{\partial W} = X^T L_S X W + \gamma Q W + W \Lambda = 0, \tag{11}$$

where $Q \in \mathbb{R}^{d \times d}$ is a diagonal matrix, and the i-th element is defined as

$$Q_{ii} = \frac{1}{2\sqrt{w_i^T w_i + \varepsilon}}. \tag{12}$$

Note that Q is also unknown and depend on W, thus we develop an iterative algorithm to solve problem (9). With fixed W, then Q is obtained by Eq. (12). And with fixed Q, it is easily to prove that solving Eq. (11) is equivalent to solving

$$\min_{W^T W = I} Tr(W^T X^T L_S X W) + \gamma Tr(W^T Q W), \tag{13}$$

and problem (13) can be solved directly to get W. The detail of the algorithm is described in Algorithm 1. Later, we will prove that Algorithm 1 can make problem (9) converge. Obviously, the converged solution satisfies KKT condition.

Fix S update F

With fixed S, the problem (6) is transformed into

$$\min_{F \in \mathbb{R}^{n \times c}, F^T F = I} Tr(F^T L_S F). \tag{14}$$

The optimal solution F is formed by the c eigenvectors of L_S corresponding to the c smallest eigenvalues.

Algorithm 1 Algorithm to solve problem (9)

Input: Data matrix $X \in \mathbb{R}^{n\times d}$, Laplacian Matrix $L_S \in \mathbb{R}^{n\times n}$, regularization parameter γ, projection dimension m

Initialize $Q \in \mathbb{R}^{d\times d}$ as $Q = I$;

repeat

1. With current Q, the optimal solution W by solving problem (13) is formed by the m eigenvectors of $(X^T L_S X + \gamma Q)$ corresponding to the m smallest eigenvalues.

2. With current W, Q is obtained by Eq. (12).

until converge

Output: Projection matrix $W \in \mathbb{R}^{d\times m}$.

Fix W and F update S

With fixed W and F, the problem (6) is transformed into

$$\begin{aligned} &\min \sum_{i,j} (\|W^T x_i - W^T x_j\|_2^2 s_{ij} + \alpha s_{ij}^2) \\ &\quad +2\lambda Tr(F^T L_S F). \\ &s.t. \quad \forall i, s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1 \end{aligned} \tag{15}$$

According to Eq. (2), we get

$$\begin{aligned} &\min \sum_{i,j} (\|W^T x_i - W^T x_j\|_2^2 s_{ij} + \alpha s_{ij}^2) \\ &\quad +\lambda \sum_{i,j} \|f_i - f_j\|_2^2 s_{ij}. \\ &s.t. \quad \forall i, s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1. \end{aligned} \tag{16}$$

Note that the similarity vector of each sample is independent, thus we can tackle the following problem for the i-th sample

$$\begin{aligned} &\min \sum_{j} (\|W^T x_i - W^T x_j\|_2^2 s_{ij} + \alpha s_{ij}^2) \\ &\quad +\lambda \sum_{j} \|f_i - f_j\|_2^2 s_{ij}. \\ &s.t. \quad s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1 \end{aligned} \tag{17}$$

For ease of representation, we denote matrix $M \in \mathbb{R}^{n\times n}$ with $m_{ij} = \|W^T x_i - W^T x_j\|_2^2$ and matrix $N \in \mathbb{R}^{n\times n}$with $n_{ij} = \|f_i - f_j\|_2^2$. Denote vector $d_i \in \mathbb{R}^{n\times 1}$ with $d_{ij} = m_{ij} + \lambda n_{ij}$. Then, we can rewrite problem (17) as follows

$$\min_{s_i^T \mathbf{1}=1, 0\le s_{ij}\le 1} \|s_i + \frac{1}{2\alpha} d_i\|_2^2. \tag{18}$$

The solution of this problem will be shown later. We summarize the detail algorithm in Algorithm 2. Noted that, in real life application, we can run only one iteration in Algorithm 1 in order to speed up Algorithm 2.

Convergence Analysis of Algorithm 1

The method proposed by Algorithm 1 can be used to find a locally optimal solution of problem (9). To prove the convergence, we need the lemma proposed by Nie et al. (2010a), which describes as follows.

Algorithm 2 Algorithm to solve problem (6)

Input: Data matrix $X \in \mathbb{R}^{n\times d}$, number of select features h, cluster number c, projection dimension m, regularization parameter γ, parameter α, a large enough λ

Initialize S by solving problem (1).

repeat

1. Update W via Algorithm (1).

2. Calculate $L_S = D - \frac{S^T+S}{2}$, where the degree matrix D is a diagonal matrix and the i-th element is defined as $\sum_j \frac{(s_{ij}+s_{ji})}{2}$.

3. Update F via solving problem (14), F is formed by the c eigenvectors of L_S corresponding to the c smallest eigenvalues.

4. Update S, each s_i is calculate by solving problem (18) individually.

until converge

Output: Calculate all $\|w_i\|_2 (i = 1 : \ldots : n)$ and sort in descending order, select top h ranked features as ultimate result.

Lemma 1. *The following inequality holds for any positive real number u and v.*

$$\sqrt{u} - \frac{u}{2\sqrt{v}} \le \sqrt{v} - \frac{v}{2\sqrt{v}}. \tag{19}$$

For detail and proof, see lemma 1 in (Nie et al. 2010a).

The convergence of Algorithm 1 can be proven by following theorem.

Theorem 1. *In Algorithm 1, updated W will decrease the objective value of problem (9) until converge.*

Proof. Suppose the updated W is $\widetilde{W}$, it's easy to know that

$$\begin{aligned} &Tr(\widetilde{W}^T X^T L_S X \widetilde{W}) + \gamma Tr(\widetilde{W}^T Q \widetilde{W}) \\ \le &Tr(W^T X^T L_S X W) + \gamma Tr(W^T Q W). \end{aligned} \tag{20}$$

We add $\gamma \sum_i \frac{\varepsilon}{2\sqrt{w_i^T w_i + \varepsilon}}$ to both sides of the inequality (20), and substitute the definition of Q in Eq. (12), then inequality (20) can be rewritten as

$$\begin{aligned} &Tr(\widetilde{W}^T X^T L_S X \widetilde{W}) + \gamma \sum_i \frac{\widetilde{w_i}^T \widetilde{w_i} + \varepsilon}{2\sqrt{w_i^T w_i + \varepsilon}} \\ \le &Tr(W^T X^T L_S X W) + \gamma \sum_i \frac{w_i^T w_i + \varepsilon}{2\sqrt{w_i^T w_i + \varepsilon}}, \end{aligned} \tag{21}$$

Based on Lemma 1, we know

$$\begin{aligned} &\gamma \sum_i \sqrt{\widetilde{w_i}^T \widetilde{w_i} + \varepsilon} - \gamma \sum_i \frac{\widetilde{w_i}^T \widetilde{w_i} + \varepsilon}{2\sqrt{w_i^T w_i + \varepsilon}} \\ \le &\gamma \sum_i \sqrt{w_i{}^T w_i + \varepsilon} - \gamma \sum_i \frac{w_i{}^T w_i + \varepsilon}{2\sqrt{w_i^T w_i + \varepsilon}} \end{aligned} \tag{22}$$

Sum over the inequality (21) and inequality (22), we arrive at

$$\begin{aligned} &Tr(\widetilde{W}^T X^T L_S X \widetilde{W}) + \gamma \sum_i \sqrt{\widetilde{w_i}^T \widetilde{w_i} + \varepsilon} \\ \le &Tr(W^T X^T L_S X W) + \gamma \sum_i \sqrt{w_i{}^T w_i + \varepsilon} \end{aligned} \tag{23}$$

which completes the proof. □

Determination of α

Let us first consider two extreme conditions of α for problem (1). One is $\alpha = 0$, it will make only one element of vector s_i not zero. Another is $\alpha = \infty$, it will make every element of vector s_i equal to $\frac{1}{n}$. Therefore, α can determine the number of sample's neighbors, and the optimal value of α should make most of s_i contain exact k non-zeros elements, where k is the number of neighbors connected to x_i.

In order to achieve above goals, we consider the Lagrangian Function of problem (18) as

$$\mathcal{L}(s_i, \theta, \varphi_i) = \frac{1}{2}\|s_i + \frac{d_i}{2\alpha_i}\|_2^2 - \theta(s_i^T \mathbf{1} - 1) - \varphi_i^T s_i, \quad (24)$$

where θ and φ_i are Lagrangian multipliers. Based on the KKT condition, the optimal solution of s_i is

$$s_{ij} = (-\frac{d_{ij}}{2\alpha_i} + \theta)_+, \quad (25)$$

where $\theta = \frac{1}{k} + \frac{1}{2k\alpha_i}\sum_{j=1}^{k} d_{ij}$ (Nie, Wang, and Huang 2014). Note that, problem (1) also can be solved in the same way.

To make s_i contain exact k non-zero elements, s_i must satisfies $s_{i,k+1} \leq 0 < s_{ik}$, thus α_i should be set as

$$\frac{k}{2}d_{ik} - \frac{1}{2}\sum_{j=1}^{k} d_{ij} < \alpha_i \leq \frac{k}{2}d_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k} d_{ij}, \quad (26)$$

where $d_{i1}, d_{i2}, \ldots, d_{in}$ are sorted in ascending order. Therefore, to get a good enough α which can make most of s_i has k non-zeros elements, we can set α to be

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}\alpha_i = \frac{1}{n}\sum_{i=1}^{n}(\frac{k}{2}d_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k} d_{ij}). \quad (27)$$

Experiments

In this section, we demonstrate the effectiveness of the proposed unsupervised feature selection method SOGFS on 8 benchmark data sets, and then show several analysis of experimental results.

Data Sets

The experiments are conducted on 8 different public available data sets, including handwritten digit (i.e. Binary Alphabet (BA) (Belhumeur, Hespanha, and Kriegman 1997), UMIST (Hou et al. 2014), USPS (Hull 1994)), human face (i.e. JAFFE (Lyons, Budynek, and Akamatsu 1999), ORL (Cai, Zhang, and He 2010)), object image (i.e. COIL20 (Nene, Nayar, and Murase 1996)), biology (i.e. SRBCT (Khan et al. 2001), Lung (Singh et al. 2002)). The detail of these data sets are summarized in Table 1.

Comparison Scheme

To validate the effectiveness of SOGFS, we compare it with several state-of-the-art unsupervised feature selection approaches, including Laplacian Score (LS) (He, Cai, and Niyogi 2005), Multi Cluster Feature Selection (MCFS) (Cai, Zhang, and He 2010), Unsupervised Discriminate Feature Selection (UDFS) (Yang et al. 2011), Robust Unsupervised Feature Selection (RUFS) (Qian and Zhai 2013) and Robust Spectral Feature Selection (RSFS) (Shi, Du, and Shen 2014). We also use all features to perform K-means as Baseline. We set parameters of all approaches in same strategy to make the experiments fair enough, i.e. $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. For the selected features, we use 5 times K-means clustering from different starting points, and only report optimal result to alleviate the stochastic effect caused by clustering method. To evaluate the result of selected features, we use clustering ACCurate (ACC) as evaluation metrics in this paper.

Clustering Result with Selected Features

The performance of feature selection approaches which are evaluated by ACC is shown in figure 1. Through the analysis of experimental results, we get some conclusions.

Generally speaking, as the size of feature subset increased, the performance of feature selection approaches shows the trend with first increase then decrease. Real life data sets always contain many redundant features, therefore the necessary information is just contained in a small size of feature set. If we increase the size excessively, lots of noise features will be brought into the final results which will surely decrease the performance. The trend indirectly validates the effective of feature selection methods.

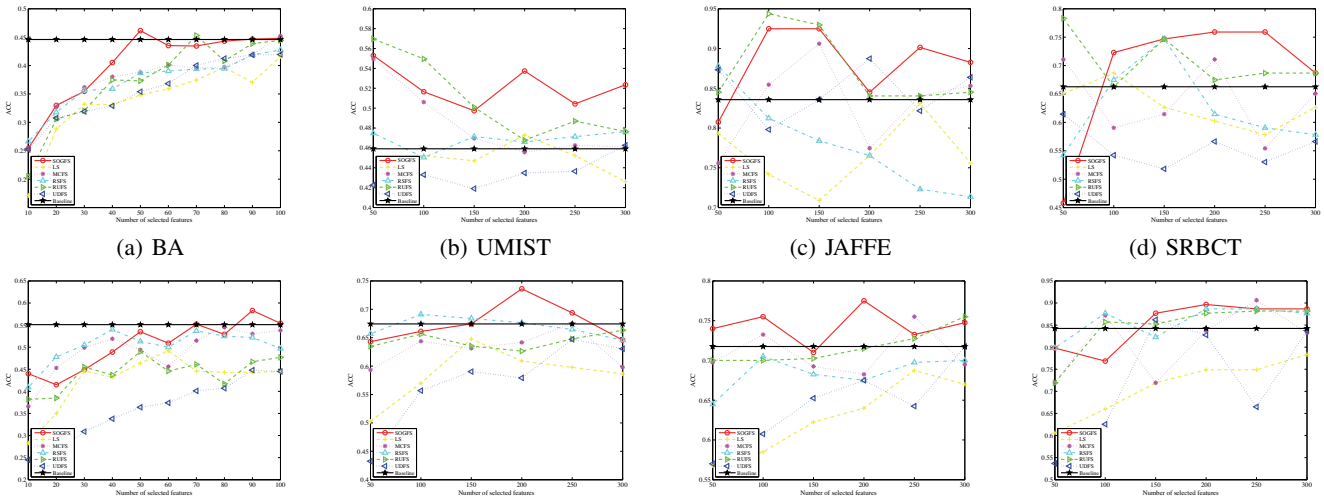
Through feature selection, we obtain refined data which contains more valuable information. Compared to baseline which uses all features to perform K-means, the results which use selected features become better in most cases. Especially our proposed method SOGFS, has more than 10% improvement in average. It directly validates that feature selection improves the data's quality.

Overall, the performance of our proposed method SOGFS exceeds other methods in ACC. To be specific, SOGFS has more than 7% improvement for human face data sets (i.e. JAFFE, ORL), compared to the second best approach RUFS. And more than 6.4% improvement for handwritten digit data sets (i.e. BA, USPS, UMIST), compared to RSFS, which is the second best approach in handwritten digit data sets. SOGFS also achieves pretty good performance for the rest data sets in average.

It seems that embedded approaches, such as MCFS, UDFS, RSFS, RUFS, have better performance than LS. With

Table 1: Data Set Description

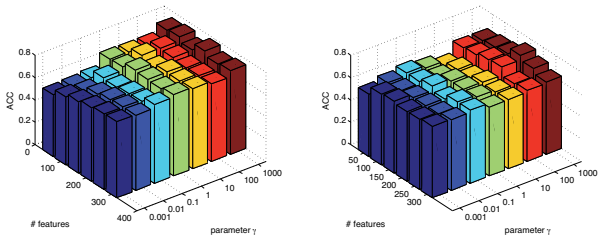
Data sets	Sample	Feature	Class	Select features
BA	1404	320	36	[10,20,. . .,100]
USPS	1000	256	10	[10,20,. . .,100]
UMIST	575	644	20	[50,100,. . .,300]
JAFFE	213	676	10	[50,100,. . .,300]
ORL	400	1024	40	[50,100,. . .,300]
COIL20	1440	1024	20	[50,100,. . .,300]
SRBCT	83	2308	4	[50,100,. . .,300]
LUNG	203	3312	5	[50,100,. . .,300]



(a) BA (b) UMIST (c) JAFFE (d) SRBCT

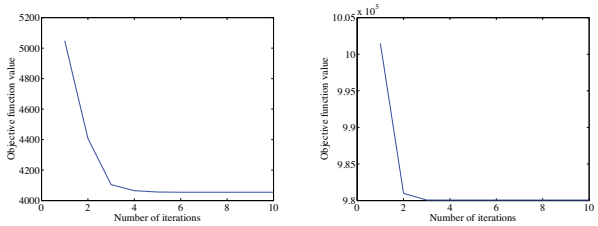
(e) USPS (f) COIL20 (g) ORL (h) LUNG

Figure 1: Clustering accuracy on 8 data sets with different number of selected features.



(a) ORL (b) COIL20

Figure 2: Clustering accuracy with different γ.



(a) UMIST (b) COIL20

Figure 3: Convergence curve of Algorithm 1

discriminate analysis and considering of noise data, RSFS and RUFS outperform MCFS. The proposed SOGFS uses structured optimal graph, and performs feature selection and local structure learning simultaneously, therefore, we obtain better results at last.

Parameter Sensitivity and Convergence Study

First, we investigate the impact of parameters in SOGFS. The parameter m influences performance slightly and is usually set empirically around $\frac{d}{3}$ to $\frac{2d}{3}$ in our experiments. Therefore, we only focus on the influence of parameter γ with fixing m. Parameter γ is used to control the row sparsity of W, and its value seriously influences the performance of SOGFS . As we vary the value of γ, the variance of performance is demonstrated in Figure 2. For brevity, we only show results on two data sets (i.e. ORL, COIL20). Results show that SOGFS is to some extent robust to γ. Nonetheless, we suggest to perform hierarchy grid search to get better result in real life application.

We propose Algorithm 1 to iteratively solve problem (9). We have already proven the convergence in the previous section, and now we experimentally study the speed of its convergence. For simplicity, we only show results on two data sets (i.e. UMIST, COIL20). The convergence curves of the objective value are shown in Figure 3. We can see that, Algorithm 1 converges very fast and almost within 10 iterations. The fast convergence of Algorithm 1 ensures the speed of the whole proposed approach.

Conclusions

In this paper, we propose a novel unsupervised feature selection approach named SOGFS, which performs feature selection and local structure learning simultaneously, to obtain optimal local structure, we propose a constraint to the approach. The selected features are obtained by analysis of projection matrix W. An efficient optimization algorithm is proposed to solve the problem. Comprehensive experiments on 8 benchmark data sets demonstrate the effectiveness of our approach. One of our future work is to replacing the regularization term $\gamma\|W\|_{2,1}$ with constraint $\|W\|_{20} = h$, which will greatly improve the usability of the proposed approach. The other is to use some technologies to speed up the approach.

References

Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19(7):711–720.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 333–342.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America* 35(11):652.

He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, 507–514.

Hou, C.; Nie, F.; Li, X.; Yi, D.; and Wu, Y. 2014. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Trans. Cybernetics* 44(6):793–804.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* 16(5):550–554.

Khan, J.; Wei, J. S.; Ringner, M.; Saal, L. H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C. R.; Peterson, C.; and Meltzer, P. S. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 7:673–679.

Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; and Lu, H. 2012. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Liu, Y.; Nie, F.; Wu, J.; and Chen, L. 2013. Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. *Neurocomputing* 105:12–18.

Lyons, M. J.; Budynek, J.; and Akamatsu, S. 1999. Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.* 21(12):1357–1362.

Mohar, B.; Alavi, Y.; Chartrand, G.; and Oellermann, O. 1991. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* 2:871–898.

Nene, S. A.; Nayar, S. K.; and Murase, H. 1996. Columbia Object Image Library (COIL-20). Technical report.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. Q. 2010a. Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization. In *Advances in Neural Information Processing Systems*, 1813–1821.

Nie, F.; Xu, D.; Tsang, I. W.; and Zhang, C. 2010b. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing* 19(7):1921–1932.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, 977–986.

Qian, M., and Zhai, C. 2013. Robust unsupervised feature selection. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*.

Shi, L.; Du, L.; and Shen, Y. 2014. Robust spectral learning for unsupervised feature selection. In *2014 IEEE International Conference on Data Mining, ICDM 2014*, 977–982.

Singh, D.; Febbo, P. G.; Ross, K.; Jackson, D. G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A. A.; D'Amico, A. V.; Richie, J. P.; Lander, E. S.; Loda, M.; Kantoff, P. W.; Golub, T. R.; and Sellers, W. R. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1(2):203 – 209.

Tabakhi, S.; Moradi, P.; and Akhlaghian, F. 2014. An unsupervised feature selection algorithm based on ant colony optimization. *Eng. Appl. of AI* 32:112–123.

Wang, X.; Liu, Y.; Nie, F.; and Huang, H. 2015. Discriminative unsupervised dimensionality reduction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 3925–3931. AAAI Press.

Wang, D.; Nie, F.; and Huang, H. 2014. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (TRACK). In *Machine Learning and Knowledge Discovery in Databases - European Conference*, 306–321.

Wang, D.; Nie, F.; and Huang, H. 2015. Feature selection via global redundancy minimization. *Knowledge and Data Engineering, IEEE Transactions on* 27(10):2743–2755.

Wang, S.; Tang, J.; and Liu, H. 2015. Embedded unsupervised feature selection. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 470–476.

Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, 1589–1594.

Zhao, Z., and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference*, 1151–1157.

Zhao, Z.; Wang, L.; and Liu, H. 2010. Efficient spectral feature selection with minimum redundancy. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*.