Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Global Discriminant Analysis for Unsupervised Feature Selection with Local Structure Preservation

Xiucai Ye, Kaiyang Ji, Tetsuya Sakurai
Department of Computer Science
University of Tsukuba, Japan

Abstract

Feature selection is an efficient technique for data dimension reduction in data mining and machine learning. Unsupervised feature selection is much more difficult than supervised feature selection due to the lack of label information. Discriminant analysis is powerful to select discriminative features, while local structure preservation is important to unsupervised feature selection. In this paper, we incorporate discriminant analysis, local structure preservation and $l_{2,1}$- norm regularization into a joint framework for unsupervised feature selection. The global structure of data is captured by the discriminant analysis, while the local manifold structure is revealed by the locality preserving projections. By imposing row sparsity on the transformation matrix, the resultant formulation optimizes for selecting the most discriminative features which can better capture both the global and local structure of data. We develop an efficient algorithm to solve the $l_{2,1}$-norm-based optimization problem in our method. Experimental results on different types of real-world data demonstrate the effectiveness of the proposed method.

Introduction

In the fields of data mining, machine learning, and computer vision, the data samples are often represented by a large number of features (Jain and Zongker 1997). The large number of features that often contain a lot of redundant and noisy information, make great challenges such as the curses of dimensionality and high computation cost. Feature selection is one main technique for dimensionality reduction that aims to extract the most useful features and eliminate the noisy ones (Guyon and Elisseeff 1997). Feature selection brings the immediate effects for applications including: speeding up the algorithms, reducing the risk of over fitting, and improving the accuracy of the predictive results (Dy and Brodley 2004). Based on the availability of label information, feature selection methods can be broadly classified into supervised and unsupervised methods (I. Guyon and Vapnik 2002). Unsupervised feature selection is considered as a more challenging problem, since the definition of relevance of features becomes unclear due to the lack of label information (Dy and Brodley 2000).

Unsupervised feature selection has attracted increasing attention in recent years (P. Zhu and Shiu 2015). Without the information of class label, unsupervised feature selection extracts features that effectively maintain the important underlying structure of data, such as the global structure (X. Liu and Liu 2014) and the local structure (Z. Zhao and Liu 2010). Many methods have been proposed to preserve the global structure of data, such as the Maximum Variance (MaxVar) method and the global pairwise similarity method (e.g., with a Gaussian kernel) (X. Liu and Liu 2014).

Instead of the global structure, a family of unsupervised feature selection methods choose features that preserve the local structure of data. The importance of preserving local structure has been well recognized in the recent development of unsupervised feature selection methods. Typical methods include: the Laplacian Score (i.e., LapScor) method (X. He and Niyogi 2006), the Multi-Cluster Feature Selection (i.e., MCFS) method (D. Cai and He 2010), Joint Embedding Learning and Sparse Regression (i.e., JELSR) method (C. Hou and Wu 2011). LapScore considers the local preserving property of individual feature while neglects the correlation among features (S. Alelyani and Liu 2013). MCFS selects the features that can best preserve the multi-cluster structure by manifold learning and l_1 regularization. JELSR uses the similarity via locally linear approximation to construct graph and unifies embedding learning and sparse regression to perform feature selection.

Compared with the global preserving unsupervised feature selection methods, the local preserving methods have been proved to perform better in many cases (Z. Zhao and Liu 2010). However, most of the local preserving unsupervised feature selection methods neglect the discriminative information of features. Discriminant analysis is important to unsupervised feature selection, which aims to select the discriminative features such that the within-class distance is as small as possible and the between-class distance is as large as possible (R. Duda and Stork 2001; Fukunaga 2013). Yang et al. (Y. Yang 2011) proposed a local discriminant analysis method (i.e., UDFS) for unsupervised feature selection. UDFS defines a local discriminative score to evaluate the within-class scatter and the between-class scatter for each data and its k nearest neighbors, in which the discriminative information mainly depends on the neighborhoods. Instead of the local discriminant analysis, Tang et al.

(J. Tang and Liu 2014) developed global discriminant analysis for unsupervised scenarios to select the discriminative features. However, this method only considers to preserve the global data structure but neglects to preserve the local data structure.

In this paper, we develop the global discriminant analysis for unsupervised feature selection, meanwhile, we consider the preservation of local data structure. That is, we incorporate discriminant analysis, local structure preservation and $l_{2,1}$-norm regularization into a joint framework for unsupervised feature selection. The global structure of data is captured by the discriminant analysis, and the local manifold structure is revealed by the locality preserving projections (LLP) (Niyogi 2004). Since we consider both global and local structure preservation, our proposed method is referred to as GLFS. The proposed GLFS method is flexible and extendable, since besides LLP there are a lot of local models can be incorporated to preserve the local data structure. To avoid the trivial solution of linear discriminant analysis for feature selection, we consider the nontrivial solution by a new formulation in GLFS. The resultant formulation of GLFS optimizes for selecting the most discriminative features which can better capture both the global and local data structure. We also proposed an iterative algorithm to effectively solve the optimization problem in the GLFS method. Many experimental results are provided for demonstration.

Related Methods

In this paper, we use x_1 , ..., x_n to denote the n unlabeled data samples, $x_i \in \mathbb{R}^m$ and $X = [x_1, ..., x_n] \in \mathbb{R}^{m\times n}$ is the data matrix. Let $\{f_1, ..., f_m\}$ be the set of features where m is the number of features. Feature selection is to select d features form $f_1, ..., f_m$ to represent the original data, where $d < m$. We use I to denote the identity matrix, and let $1_n \in \mathbb{R}^n$ denote a column vector with all of its elements being 1. The centering matrix is $H_n = I - \frac{1}{n}1_n 1_n^T$. For a matrix $A \in \mathbb{R}^{u\times v}$, its $l_{2,1}$-norm is defined as

$$\|A\|_{2,1} = \sum\nolimits_{i=1}^{u} \sqrt{\sum\nolimits_{j=1}^{v} A_{i,j}^2}. \tag{1}$$

Consider that x_1 , ..., x_n are sampled from c clusters. Let $Y = [y_1, ..., y_n]^T \in \{0, 1\}^{n\times c}$ denote the label matrix, where $y_i \in \{0, 1\}^{c\times 1}$ is the label vector of x_i. The j^{th} element of y_i is 1 if x_i is in the j^{th} cluster, and 0 otherwise. The scaled cluster indicator matrix F is defined as $F = [F_1, ..., F_n]^T = Y(Y^TY)^{-1/2}$. It is obvious that $F^TF = (Y^TY)^{-1/2}Y^TY(Y^TY)^{-1/2} = I_c$. The total scatter matrix S_t and the between-cluster scatter matrix S_b are defined as (Fukunaga 2013)

$$S_t = \sum\nolimits_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T, \tag{2}$$

$$S_b = \sum\nolimits_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \tilde{X}FF^T\tilde{X}^T, \tag{3}$$

where μ is the mean of all data, μ_i is the mean of data in the i^{th} cluster, n_i is the number of data in the i^{th} cluster, $\tilde{X} = XH_n$ is the data matrix after being centered.

The linear discriminant analysis is to find a linear transformation $W \in \mathbb{R}^{m\times q}$ ($q < m$) that projects X from the m-dimensional space to the q-dimensional space. In the lower dimensional space, the within-cluster distance is minimized while the between-cluster distance is maximized as (Fukunaga 2013)

$$\max_W Tr((W^TS_tW)^{-1}W^TS_bW). \tag{4}$$

Inspired by (Fukunaga 2013), Tang et al. (J. Tang and Liu 2014) utilized the linear discriminant analysis for unsupervised feature selection and formulated the optimization problem as

$$\begin{aligned} \max_{W,F} Tr((W^TS_tW)^{-1}W^TS_bW) - \alpha\|W\|_{2,1}, \\ s.t. F = Y(Y^TY)^{-1/2}, \end{aligned} \tag{5}$$

where the term $\|W\|_{2,1}$ is introduced to ensure that W is sparse in rows, and α is a parameter to control the sparsity of W. Let $W = [w_1, ..., w_n]^T \in \mathbb{R}^{m\times q}$, where w_i is the i^{th} row of W. Since w_i corresponds to the weight of feature f_i, the sparsity constraint on rows makes W suitable for feature selection. Each feature f_i is ranked according to $\|w_i\|_2$ in descending order and the top rank d features are selected.

However, Tao et al. (H. Tao and Yi 2015) have proved that (5) has a trivial solution of all zeros. The transformation matrix W may lose its function of selecting features if it leads to a solution near to the trivial solution. In this paper, we consider the nontrivial solution of (5), which also inherits the merit of selecting the most discriminative features. Meanwhile, we consider to preserve the local data structure in the low dimensional space by the transformation matrix W.

The Proposed Method

In this section, we propose a novel method for unsupervised feature selection, which is referred to as GLFS.

The Objective Function

By incorporating discriminant analysis, local structure preservation and $l_{2,1}$-norm regularization, the proposed GLFS method is formulated as

$$\begin{aligned} \min_{W,F} -Tr((W^TS_tW)^{-1}W^TS_bW) + \alpha\|W\|_{2,1} \\ +\beta Tr(W^TXLX^TW), \\ s.t. FF^T = I_c, F \geq 0, \end{aligned} \tag{6}$$

where $L \in \mathbb{R}^{n\times n}$ is a matrix that conserves the local geometric structure of data, α and β are two balanced parameters. We relax the condition of $F = Y(Y^TY)^{-1/2}$ to $FF^T = I_c$ in (6) as in (Y. Yang 2011). Since the nonnegative constraint of F can help to relieve the deviation from the true solution (Y. Yang and Zhou 2011), we constrain F to be nonnegative.

To avoid the trivial solution of all zeros in (6), we constrain the transformation matrix W to be uncorrelated with respect to S_t , i.e., $W^TS_tW = I$, similar to that considered

in (H. Tao and Yi 2015). The objective function of GLFS becomes

$$\begin{aligned} \min_{W,F} & -Tr(W^T S_b W) + \alpha \|W\|_{2,1} \\ & +\beta Tr(W^T X L X^T W), \\ s.t. & FF^T = I_c, F \geq 0, W^T S_t W = I. \end{aligned} \tag{7}$$

Note that in the objective function of GLFS in (7), many methods can be used to conserve the local data structure, such as locality preserving projections (LLP) (Niyogi 2004) and locally linear embedding (LLE) (Roweis and Saul 2000). For the sake of convenience, in this paper, we use the LLP method to conserve the local data structure. The LPP method aims to preserve the similarity of the original data in the lower dimensional space and forms the transformation matrix W by solving the following optimization problem

$$\min_W \sum_{i,j=1}^{n} \|x_i^T W - x_j^T W\|_2^2 S_{ij}, \tag{8}$$

where S_{ij} is the pairwise similarity between x_i and x_j. Based on the k-nearest neighbor graph, S_{ij} is calculated as

$$S_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}), & x_i \text{ and } x_j \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Let $L = D - S$ be the Laplacian matrix, where S is the similarity matrix with S_{ij} as its entries, D is the $n \times n$ diagonal matrix with $D_{ii} = \sum_{j=1}^{n} S_{ij}$ on the diagonal. Then, (10) can be equivalently expressed as

$$\min_W Tr(W^T X L X^T W). \tag{10}$$

The proposed GLFS method, i.e., the objective function in (7) integrates (10) to conserve the local geometric structure.

Optimization

In (7), the optimization problem is not convex when both W and F are optimized simultaneously, and the $l_{2,1}$-norm regularization term is non-smooth. To optimize the objective function, we propose an iterative algorithm, which divides the problem in (7) into two steps: learning the transformation matrix W while fixing the scaled cluster indicator matrix F, and learning F while fixing W.

According to (2), (3) and $FF^T = I_c$, we rewrite the objective function of GLFS as follows.

$$\begin{aligned} \min_{W,F} & -Tr(W^T \tilde{X} F F^T \tilde{X}^T W) + \alpha \|W\|_{2,1} \\ & +\beta Tr(W^T X L X^T W) + \frac{\gamma}{2} \|F^T F - I_c\|_F^2, \\ & s.t. F \geq 0, W^T \tilde{X} \tilde{X}^T W = I, \end{aligned} \tag{11}$$

where $\gamma > 0$ is a parameter which should be large enough to ensure the orthogonality.

When F is fixed, we need to solve the following problem by denoting $B = \beta X L X^T - \tilde{X} F F^T \tilde{X}^T$.

$$\begin{aligned} \min_W & Tr(W^T B W) + \alpha \|W\|_{2,1}, \\ & s.t. W^T \tilde{X} \tilde{X}^T W = I. \end{aligned} \tag{12}$$

By constructing an auxiliary function, $Tr((W^T B W) + \alpha \|W\|_{2,1}$ can be rewritten as $Tr((W^T B W) + \alpha Tr(W^T U W)$, where $U \in \mathbb{R}^{m \times m}$ is a diagonal matrix with the i^{th} diagonal element as

$$U_{ii} = \frac{1}{2\|w_i\|_2}. \tag{13}$$

Then, rewrite (12), we obtain

$$\begin{aligned} \min_W & Tr(W^T (B + \alpha U) W), \\ & s.t. W^T \tilde{X} \tilde{X}^T W = I. \end{aligned} \tag{14}$$

The solution of (14) can be obtained by solving the following generalized eigenproblem.

$$(B + \alpha U) \tilde{w} = \lambda \tilde{X} \tilde{X}^T \tilde{w}. \tag{15}$$

The matrix $W \in \mathbb{R}^{m \times q}$, containing the eigenvectors corresponding to the q smallest eigenvalues as the column vectors, is the solution of (14). Then, we normalize W such that $(W^T \tilde{X} \tilde{X}^T W)_{ii} = 1, i = 1, ..., q$.

Next, when W is fixed, we need to solve the following problem.

$$\begin{aligned} \min_F & -Tr(W^T \tilde{X} F F^T \tilde{X}^T W) + \frac{\gamma}{2} \|F^T F - I_c\|_F^2, \\ & s.t. F \geq 0. \end{aligned} \tag{16}$$

Since $Tr(W^T \tilde{X} F F^T \tilde{X}^T W) = Tr(F^T \tilde{X}^T W W^T \tilde{X} F)$, let $M = -\tilde{X}^T W W^T \tilde{X}$, (16) can be rewritten as

$$\begin{aligned} \min_F & Tr(F^T M F) + \frac{\gamma}{2} \|F^T F - I_c\|_F^2, \\ & s.t. F \geq 0. \end{aligned} \tag{17}$$

Following (Y. Yang and Zhou 2011), we update F as

$$F_{ij} \leftarrow F_{ij} \frac{(\gamma F)_{ij}}{(MF + \gamma F F^T F)_{ij}}. \tag{18}$$

Then, we normalize F such that $(F^T F)_{ii} = 1, i = 1, ..., n$.

We summarize the procedure of the proposed GLFS method in Algorithm 1. The most time consuming operation is to solve the generalized eigenproblem in (15). The time complexity of the operation is $O(m^3)$ approximately. Empirical results show that the convergence is fast and only several iterations (less than 10 iterations in the presented datasets) are needed to converge. Thus, the proposed method scales well in practice.

Convergence Analysis

Algorithm 1 will monotonically decrease the value of the objection function in (11) in each iteration.

We denote the formulation in (11) as

$$\begin{aligned} \Theta(W, F) = & -Tr(W^T \tilde{X} F F^T \tilde{X}^T W) + \alpha \|W\|_{2,1} \\ & +\beta Tr(W^T X L X^T W) + \frac{\gamma}{2} \|F^T F - I_c\|_F^2. \end{aligned} \tag{19}$$

We show that $\Theta(W^{t+1}, F^{t+1}) \leq \Theta(W^t, F^t)$.

Algorithm 1 The proposed GLFS method

Require:
Data matrix, $X \in \mathbb{R}^{m\times n}$; Parameters α, β, γ, k, c, q; Number of features to select d;

Ensure:
d selected features;

1: Construct the k-nearest neighbor graph and calculate L;
2: The iteration step $t = 1$; Initialize $F^1 \in \mathbb{R}^{n\times c}$ and set $U^1 \in \mathbb{R}^{m\times m}$ as an identity matrix;
3: Calculate $B^1 = \beta XLX^T - \tilde{X}F^1(F^1)^T\tilde{X}^T$;
4: Calculate W^1 by solving the generalized engenproblem $(B^1 + \alpha U^1)\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}$;
5: **repeat**
6: Calculate $M^t = -\tilde{X}^T W^t(W^t)^T\tilde{X}$;
7: $F_{ij}^{t+1} = F_{ij}^t \frac{(\gamma F^t)_{ij}}{(M^tF^t+\gamma F^t(F^t)^TF^t)_{ij}}$;
8: Update the diagonal matrix U^{t+1} with the i^{th} diagonal element as $U_{ii}^{t+1} = \frac{1}{2\|w_i^t\|_2}$;
9: Calculate $B^{t+1} = \beta XLX^T - \tilde{X}F^{t+1}(F^{t+1})^T\tilde{X}^T$;
10: Calculate W^{t+1} by solving the generalized engenproblem $(B^{t+1} + \alpha U^{t+1})\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}$;
11: t=t+1;
12: **until** Convergence
13: Sort each feature f_i according to $\|w_i\|_2$ in descending order and select the top d ranked ones.

We first prove $\Theta(W^{t+1}, F^t) \leq \Theta(W^t, F^t)$ where F^t is fixed. With F^t fixed, $\Theta(W^t, F^t) = Tr((W^t)^T B^t W^t) + \alpha\|W^t\|_{2,1}$. In the $(t+1)^{th}$ iteration

$$W^{k+1} = \min_{W, W^T\tilde{X}\tilde{X}^TW=I} Tr(W^T(B^t + \alpha U^t)W), \quad (20)$$

which indicates that

$$\begin{aligned} &Tr((W^{t+1})^T(B^{t+1} + \alpha U)W^{t+1}) \\ &\leq Tr((W^t)^T(B^t + \alpha U)W^t). \end{aligned} \quad (21)$$

Since $\|W\|_{2,1} = \sum_{i=1}^m \|w_i\|_2$, we obtain

$$\begin{aligned} &Tr((W^{t+1})^T B^{t+1}W^{t+1}) + \alpha\|W^{t+1}\|_{2,1} \\ &+\alpha\sum_{i=1}^m (\frac{\|w_i^{t+1}\|_2^2}{2\|w_i^t\|_2} - \|w_i^{t+1}\|_2) \leq Tr((W^t)^T B^t W^t) \\ &+\alpha\|W^t\|_{2,1} + \alpha\sum_{i=1}^m (\frac{\|w_i^t\|_2^2}{2\|w_i^t\|_2} - \|w_i^t\|_2). \end{aligned} \quad (22)$$

According to a Lemma in (F. Nie and Ding 2010), we know

$$\frac{\|w_i^{t+1}\|_2^2}{2\|w_i^t\|_2} - \|w_i^{t+1}\|_2 \geq \frac{\|w_i^t\|_2^2}{2\|w_i^t\|_2} - \|w_i^t\|_2. \quad (23)$$

Combing (22) and (23), we have

$$\begin{aligned} &Tr((W^{t+1})^T B^{t+1}W^{t+1}) + \alpha\|W^{t+1}\|_{2,1} \\ &\leq Tr((W^t)^T B^t W^t) + \alpha\|W^t\|_{2,1} \end{aligned} \quad (24)$$

That is

$$\Theta(W^{t+1}, F^t) \leq \Theta(W^t, F^t). \quad (25)$$

Table 1: Properties of Datasets

Dataset	# of samples	# of Features	# of Clusters
UMIST	575	644	20
ORL	400	1024	40
JAFFE	213	676	10
BA	1404	320	36
MNIST	2000	784	10
USPS	9298	256	10
Isolet5	1559	617	26
COIL20	1440	1024	20

Next, we can prove $\Theta(W^t, F^{t+1}) \leq \Theta(W^t, F^t)$ (W^t is fixed) by using the method in (Y. Yang and Zhou 2011).

According to (25), we have $\Theta(W^{t+1}, F^{t+1}) \leq \Theta(W^t, F^{t+1}) \leq \Theta(W^t, F^t)$. Thus, the procedure in Algorithm 1 is convergent.

Experiments

In this section, we conduct experiments to evaluate the performance of the proposed GLFS method. We test the performance in terms of clustering. After selecting the features, clustering is performed by using only the selected features.

Experiment Setup

In our experiment, we use a diversity of eight public datasets to compare the performance of different unsupervised feature selection methods. The datasets include three face image datasets, i.e., UMIST [1], ORL[2] and JAFFE[3], three handwritten digit datasets, i.e., Binary Alphabet (BA)[4], MNIST[2] and USPS [2], one spoken letter recognition data, i.e., Isolet5 [2], and one object dataset, i.e., COIL20 [2]. Their properties are summarized in Table 1.

We compare the proposed method with several well-known unsupervised feature selection methods, including LapScore (X. He and Niyogi 2006), MCFS (D. Cai and He 2010), JELSR (C. Hou and Wu 2011), and UDFS (Y. Yang 2011). We also compare these feature selection methods with the baseline method, which uses all the features for clustering. We set the number of nearest neighbors as k = 5 for all the compared methods. To fairly compare different unsupervised feature selection method, we tune the parameters from $\{10^{-6}, 10^{-4}, 10^2, 1, 10^2, 10^4, 10^6\}$. The number of selected features is ranged from $\{50, 100, 150, 200, 250, 300\}$. Two widely used evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI) (Strehl and Ghosh 2002), are applied to evaluate the clustering results. We report the best result of all the methods by using different parameters. We first perform each feature selection method to select features and then perform K-means based on the selected features. We repeat the clustering 20 times with ran-

[1] http://www.sheffield.ac.uk/eee/research/iel/research/face
[2] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
[3] http://www.cs.nyu.edu/ roweis/data.html
[4] http://www.cs.nyu.edu/∼ roweis/data.html

Table 2: Clustering Results (NMI % ± std) of Different Feature Selection Methods

Dataset	UMIST	ORL	JAFFE	BA	MNIST	USPS	Isolet5	COIL20
All Features	64.1±5.2	67.2±4.8	73.5±5.8	56.0±2.0	47.7±2.5	63.5±6.2	67.3±3.0	72.0±4.3
LapScore	60.2±4.8	61.2±4.8	72.8±6.2	56.3±1.8	48.2±2.8	62.5±3.2	45.3±3.8	65.9±4.4
MCFS	64.8±4.6	69.1±2.0	76.2±4.4	56.5±1.8	50.8± 2.3	64.1±5.1	70.6±1.8	68.2±4.5
JELSR	65.2±4.2	70.4±1.7	76.8±4.8	56.9±1.3	52.0±2.2	64.6±4.7	69.8±2.3	70.2±4.8
UDFS	65.0±4.9	68.8±1.8	75.3±4.6	57.7±1.5	51.2± 2.0	62.4±5.1	68.2±2.8	72.4±4.1
GLFS	**65.8±3.8**	**70.6±1.9**	**77.6±4.2**	**58.2±1.4**	**52.7±2.2**	**65.0±5.0**	**72.4±1.2**	**73.1±4.2**

Table 3: Clustering Results (ACC % ± std) of Different Feature Selection Methods

Dataset	UMIST	ORL	JAFFE	BA	MNIST	USPS	Isolet5	COIL20
All Features	43.0±3.7	45.6±6.0	68.2±6.5	38.5±3.1	52.4±5.0	60.2±3.6	45.7±4.5	57.5±3.2
LapScore	40.2±3.8	41.6±6.0	69.5±6.4	40.6±2.9	55.2±4.8	60.4±2.5	35.2±4.8	45.8±6.2
MCFS	42.8±3.6	48.4±5.2	72.4±5.8	41.2±2.8	56.8± 4.3	61.1±1.9	53.5±2.8	50.2±5.2
JELSR	44.9±3.2	50.0±4.8	72.6±5.4	40.3±3.0	57.1±4.2	61.2±2.0	51.5±3.7	56.2±4.3
UDFS	44.5±2.9	47.5±6.4	71.2±6.2	42.2±2.6	57.6± 4.0	60.8±2.6	51.2±4.5	57.2±2.8
GLFS	**45.2±3.0**	**50.5±4.7**	**73.1±5.0**	**43.0±2.4**	**58.8±3.8**	**62.0±2.4**	**55.2±2.6**	**57.8±2.7**

dom initializations and report the average results. All experiments were run in MATLAB 8.5.0 (R2015a) on Mac OS X 10.10.3 with core i7 (i7-4650u) CPU and 8GB ram.

Experimental results

First, we compare the performance of the feature selection methods and summarize the clustering results on the eight datasets in Table 2 and Table 3. We can see from the two tables that most of the unsupervised feature selection methods performs better than the baseline method. Feature selection can improve the accuracy of clustering results. Since the LapScore method neglects the correlation among features, it can not improve the accuracy of clustering results for many datasets. JELSR, UDFS and GLFS use $l_{2,1}$-norm regularization for sparsity constraint on the transformation matrix, while MCFS uses l_1-norm sparsity constraint. On most of the datasets, JELSR, UDFS and GLFS perform better than MCFS. Both UDFS and GLFS apply discriminant analysis for feature selection, which results in more accurate clustering than other methods on most of the data sets. The differences between UDFS and GLFS are that UDFS utilizes local discriminant analysis while GLFS utilizes global discriminant analysis. As shown in Table 2 and Table 3, the proposed GLFS method obtains best performance on all the eight datasets. That is because GLFS utilizes the global discriminant analysis and the local structure preservation simultaneously, which is able to select the most discriminative features to better capture both the global and local structure of data.

Then, we study the performance variation of GLFS with respect to the parameters α, β and the number of selected features. Due to the limited space, we only present the results in terms of NMI and objective values over UMIST, JAFFE, BA and Isolet5 datasets. The experimental results are shown in Fig. 1 and Fig. 2. We can see from these figures that the proposed GLFS method is not sensitive to the parameters α and β with wide range. On most of the datasets, the results are very stable when the number of selected features is larger.

Conclusion

In this paper, we propose a novel unsupervised feature selection method, which incorporates discriminant analysis, local structure preservation and $l_{2,1}$- norm regularization into a joint framework. The proposed method optimizes for selecting the most discriminative features which can better capture both the global and local structure of data. We derive an efficient algorithm to solve the optimization problem of the proposed method and show that the algorithm will monotonically decrease the objective until convergence. Experiments on various types of datasets demonstrate the advantages of the proposed method.

Acknowledgments

This work was supported in part by JST/CREST and MEXT KAKENHI (Grant No.25286097).

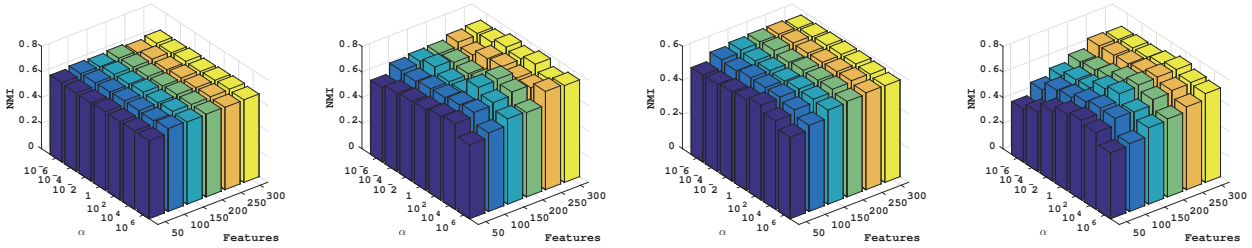
References

C. Hou, F. Nie, D. Y., and Wu, Y. 2011. Feature selection via joint embedding learning and sparse regression. In *Proceedings of International Joint Conference on Artificial Intelligence*, 1324–1329.

D. Cai, C. Z., and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 333–342.

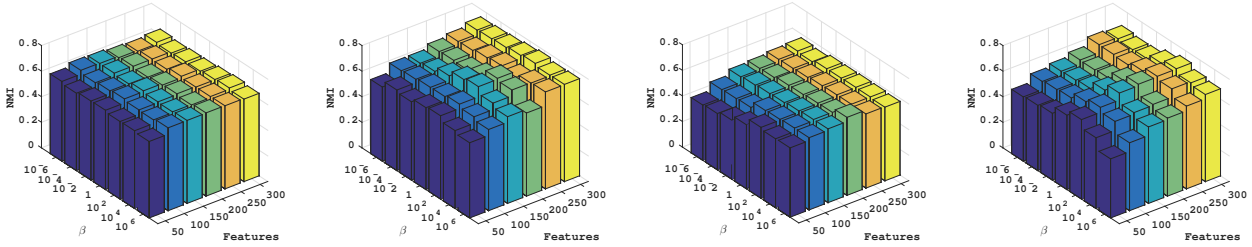
Dy, J. G., and Brodley, C. E. 2000. Visualization and interactive feature selection for unsupervised data. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 360–364.

Dy, J. G., and Brodley, C. E. 2004. Feature selection for unsupervised learning. *Journal of Machine Learning Research* 5:845–889.



(a) UMIST (b) JAFFE (c) BA (d) Isolet5

Figure 1: Normalized Mutual Information (NMI) of GLFS with different α and feature numbers when $\beta = 10^2$.



(a) UMIST (b) JAFFE (c) BA (d) Isolet5

Figure 2: Normalized Mutual Information (NMI) of GLFS with different β and feature numbers when $\alpha = 10^2$.

F. Nie, H. Huang, X. C., and Ding, C. 2010. Efficient and robust feature selection via joint l2,1-norms minimization. In *Advances in neural information processing systems*.

Fukunaga, K. 2013. Introduction to statistical pattern recognition. *Academic press*.

Guyon, I., and Elisseeff, A. 1997. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3:1157–1182.

H. Tao, C. Hou, F. N. Y. J., and Yi, D. 2015. Effective discriminative feature selection with nontrivial solution. *IEEE Transactions on Neural Networks and Learning Systems*.

I. Guyon, J.Weston, S. B., and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46:389–422.

J. Tang, Xi. Hu, H. G., and Liu, H. 2014. Discriminant analysis for unsupervised feature selection. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, 9–17.

Jain, A., and Zongker, D. 1997. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(2):153–158.

Niyogi, X. 2004. Locality preserving projections. *Neural information processing systems* 16:153.

P. Zhu, W. Zuo, L. Z. Q. H., and Shiu, S. C. 2015. Unsupervised feature selection by regularized self-representation. *Pattern Recognition* 48:438–446.

R. Duda, P. H., and Stork, D. 2001. Pattern classification. *Wiley New York*.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.

S. Alelyani, J. T., and Liu, H. 2013. Feature selection for clustering: A review. *Data Clustering: Algorithms and Applications* 48.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles–a knowledge reuse framework for combining multiple partitions,. *Journal of Machine Learning Research* 3:583?617.

X. He, D. C., and Niyogi, P. 2006. Laplacian score for feature selection. In *Advances in neural information processing systems*, 507–514.

X. Liu, L. Wang, J. Z. J. Y., and Liu, H. 2014. Global and local structure preservation for feature selection. *IEEE Transactions on Neural Networks and Learning Systems* 25:1083–1095.

Y. Yang, H. T. Shen, F. N. R. J., and Zhou, X. 2011. Nonnegative spectral clustering with discriminative regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Y. Yang, H. Shen, Z. M. Z. H. a. Z. 2011. L21-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, 1589–1594.

Z. Zhao, L. W., and Liu, H. 2010. Efficient spectral feature selection with minimum redundancy. In *Proceedings of the Twenty-4th AAAI Conference on Artificial Intelligence*, 1–6.