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 \( \ell_{2,1} \)-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 \( \ell_{2,1} \)-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). LapScor 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 \( \ell_1 \) 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 from $f_1, ..., f_m$ 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 $H_n = I - \frac{1}{n}1_n1_n^T$. For a matrix $A \in \mathbb{R}^{u \times v}$, its $l_2,1$-norm is defined as

$$\|A\|_{2,1} = \sum_{i=1}^u \sqrt{\sum_{j=1}^v A_{i,j}^2},$$

Consider that $x_i, ..., x_n$ 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 $F^T F = (Y^TY)^{1/2}Y^TY(Y^TY)^{-1/2} = I_c$. The total scatter matrix $S_t$ and the between-cluster scatter matrix $S_b$ are defined as (Fukunaga 2013)

$$S_t = \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T = \bar{X}\bar{X}^T,$$

$$S_b = \sum_{i=1}^c n_i(\mu_i - \mu)(\mu_i - \mu)^T = \bar{X}FF^T\bar{X}^T,$$

where $\mu$ is the mean of all data, $\mu_i$ is the mean of data in the $i$th cluster, $n_i$ is the number of data in the $i$th cluster, $X = XH_n$ 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)

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

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

$$\max_{W,F} Tr((W^TS_tW)^{-1}W^TS_bW) - \alpha\|W\|_{2,1},$$

$$s.t. F = Y(Y^TY)^{-1/2},$$

where the term $\|W\|_{2,1}$ is introduced to ensure that $W$ is sparse in rows, and $\alpha$ is a parameter to control the sparsity of $W$. To avoid the trivial solution of linear discriminant analysis, 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.

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

$$\min_{W,F} -T r((W^TS_bW)^{-1}W^TS_bW) + \alpha\|W\|_{2,1},$$

$$+ \beta Tr(W^TXFX^TW),$$

$$s.t. FF^T = I_c, F \geq 0,$$

where $L \in \mathbb{R}^{n \times n}$ is a matrix that conserves the local geometric structure of data, $\alpha$ and $\beta$ 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

\[ \min_{W,F} -\text{Tr}(W^T S_b W) + \alpha \|W\|_{2,1} + \beta \text{Tr}(W^T X L X^T W), \]  
\[ \text{s.t.} \quad F^T = I_c, F \geq 0, W^T S t W = I. \]  

(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

\[ \min_{W} \sum_{i,j=1}^{n} \|x_i^T W - x_j^T W\|_2^2 S_{ij}, \]  

(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

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

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

\[ \min_{W} \text{Tr}(W^T X L X^T W). \]  

(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.

\[ \min_{W,F} -\text{Tr}(W^T \tilde{X} F F^T \tilde{X}^T W) + \alpha \|W\|_{2,1} + \beta \text{Tr}(W^T X L X^T W) + \frac{\gamma}{2} \|F^T F - I_c\|^2_F, \]  
\[ \text{s.t.} \quad F \geq 0, W^T \tilde{X} \tilde{X}^T W = I, \]  

(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 \( B = \beta X L X^T - \tilde{X} F F^T \tilde{X} \).

\[ \min_{W} \text{Tr}(W^T B W) + \alpha \|W\|_{2,1}, \]  
\[ \text{s.t.} \quad W^T \tilde{X} \tilde{X}^T W = I. \]  

(12)

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

\[ U_{ii} = \frac{1}{2\|w_i\|^2_2}. \]  

(13)

Then, rewrite (12), we obtain

\[ \min_{W} \text{Tr}(W^T (B + \alpha U) W), \]  
\[ \text{s.t.} \quad W^T \tilde{X} \tilde{X}^T W = I. \]  

(14)

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

\[ (B + \alpha U) \tilde{w} = \lambda \tilde{X} \tilde{X}^T \tilde{w}. \]  

(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 \( (\tilde{W} \tilde{X} \tilde{X}^T W)_{ii} = 1, i = 1, ..., q \).

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

\[ \min_{F} -\text{Tr}(W^T \tilde{X} F F^T \tilde{X}^T W) + \frac{\gamma}{2} \|F^T F - I_c\|^2_F, \]  
\[ \text{s.t.} \quad F \geq 0. \]  

(16)

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

\[ \min_{F} \text{Tr}(F^T M F) + \frac{\gamma}{2} \|F^T F - I_c\|^2_F, \]  
\[ \text{s.t.} \quad F \geq 0. \]  

(17)

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

\[ F_{ij} \leftarrow F_{ij} \frac{(\gamma F)_{ij}}{(M + \gamma F F^T)_{ij}}. \]  

(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 objective function in (11) in each iteration. We denote the formulation in (11) as

\[ \Theta(W, F) = -\text{Tr}(W^T \tilde{X} F F^T \tilde{X}^T W) + \alpha \|W\|_{2,1} + \beta \text{Tr}(W^T X L X^T W) + \frac{\gamma}{2} \|F^T F - I_c\|^2_F. \]  

(19)

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

$$W^{k+1} = \min_{W \in \mathbb{R}^{d \times n}} Tr(W^T (B^{t+1} + \alpha U^{t+1}) W),$$

which indicates that

$$Tr((W^{t+1})^T (B^{t+1} + \alpha U) W^{t+1}) \leq Tr((W^t)^T (B^t + \alpha U) W^t).$$

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

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

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

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

Combing (22) and (23), we have

$$Tr((W^{t+1})^T B^{t+1} W^{t+1}) + \alpha \|W^{t+1}\|_{2,1}
\leq Tr((W^t)^T B^t W^t) + \alpha \|W^t\|_{2,1}.$$ (24)

That is

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

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of samples</th>
<th># of Features</th>
<th># of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMIST</td>
<td>575</td>
<td>644</td>
<td>20</td>
</tr>
<tr>
<td>ORL</td>
<td>400</td>
<td>1024</td>
<td>40</td>
</tr>
<tr>
<td>JAFFE</td>
<td>213</td>
<td>676</td>
<td>10</td>
</tr>
<tr>
<td>BA</td>
<td>1404</td>
<td>320</td>
<td>36</td>
</tr>
<tr>
<td>MNIST</td>
<td>2000</td>
<td>784</td>
<td>10</td>
</tr>
<tr>
<td>USPS</td>
<td>9298</td>
<td>256</td>
<td>10</td>
</tr>
<tr>
<td>Isolet5</td>
<td>1559</td>
<td>617</td>
<td>26</td>
</tr>
<tr>
<td>COIL20</td>
<td>1440</td>
<td>1024</td>
<td>20</td>
</tr>
</tbody>
</table>

1http://www.sheffield.ac.uk/eee/research/iel/research/face
2http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
3http://www.cs.nyu.edu/~roweis/data.html
4http://www.cs.nyu.edu/~roweis/data.html
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 perform 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 $\alpha$, $\beta$ 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 $\alpha$ and $\beta$ 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


458
Figure 1: Normalized Mutual Information (NMI) of GLFS with different $\alpha$ and feature numbers when $\beta = 10^2$.

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


