Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Unsupervised Feature Selection Using Nonnegative Spectral Analysis

Zechao Li[†], Yi Yang[‡], Jing Liu[†], Xiaofang Zhou[♯], Hanqing Lu[†]
[†]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Science
[‡]School of Computer Science, Carnegie Mellon University
[♯]School of Information Technology and Electrical Engineering, The University of Queensland
{zcli, jliu, luhq}@nlpr.ia.ac.cn, yiyang@cs.cmu.edu, zxf@itee.uq.edu.au

Abstract

In this paper, a new unsupervised learning algorithm, namely Nonnegative Discriminative Feature Selection (NDFS), is proposed. To exploit the discriminative information in unsupervised scenarios, we perform spectral clustering to learn the cluster labels of the input samples, during which the feature selection is performed simultaneously. The joint learning of the cluster labels and feature selection matrix enables NDFS to select the most discriminative features. To learn more accurate cluster labels, a nonnegative constraint is explicitly imposed to the class indicators. To reduce the redundant or even noisy features, $\ell_{2,1}$-norm minimization constraint is added into the objective function, which guarantees the feature selection matrix sparse in rows. Our algorithm exploits the discriminative information and feature correlation simultaneously to select a better feature subset. A simple yet efficient iterative algorithm is designed to optimize the proposed objective function. Experimental results on different real world datasets demonstrate the encouraging performance of our algorithm over the state-of-the-arts.

Introduction

The dimension of data is often very high in many domains (Jain and Zongker 1997; Guyon and Elisseeff 2003), such as image and video understanding (Wang et al. 2009a; 2009b), and bio-informatics. In practice, not all the features are important and discriminative, since most of them are often correlated or redundant to each other, and sometimes noisy (Duda, Hart, and Stork 2001; Liu, Wu, and Zhang 2011). These features may result in adverse effects in some learning tasks, such as over-fitting, low efficiency and poor performance (Liu, Wu, and Zhang 2011). Consequently, it is necessary to reduce dimensionality, which can be achieved by feature selection or transformation to a low dimensional space. In this paper, we focus on feature selection, which is to choose discriminative features by eliminating the ones with little or no predictive information based on certain criteria.

Many feature selection algorithms have been proposed, which can be classified into three main families: filter, wrapper, and embedded methods. The filter methods (Duda, Hart, and Stork 2001; He, Cai, and Niyogi 2005; Zhao and Liu 2007; Masaeli, Fung, and Dy 2010; Liu, Wu, and Zhang 2011; Yang et al. 2011a) use statistical properties of the features to filter out poorly informative ones. They are usually performed before applying classification algorithms. They select a subset of features only based on the intrinsic properties of the data. In the wrapper approaches (Guyon and Elisseeff 2003; Rakotomamonjy 2003), feature selection is "wrapped" in a learning algorithm and the classification performance of features is taken as the evaluation criterion. Embedded methods (Vapnik 1998; Zhu et al. 2003) perform feature selection in the process of model construction. In contrast with filter methods, wrapper and embedded methods are tightly coupled with in-built classifiers, which causes that they are less generality and computationally expensive. In this paper, we focus on the filter feature selection algorithm.

Because of the importance of discriminative information in data analysis, it is beneficial to exploit discriminative information for feature selection, which is usually encoded in labels. However, how to select discriminative features in unsupervised scenarios is a significant but hard task due to the lack of labels. In light of this, we propose a novel unsupervised feature selection algorithm, namely Nonnegative Discriminative Feature Selection (NDFS), in this paper. We perform spectral clustering and feature selection simultaneously to select the discriminative features for unsupervised learning. The cluster label indicators are obtained by spectral clustering to guide the feature selection procedure. Different from most of the previous spectral clustering algorithms (Shi and Malik 2000; Yu and Shi 2003), we explicitly impose a nonnegative constraint into the objective function, which is natural and reasonable as discussed later in this paper. With nonnegative and orthogonality constraints, the learned cluster indicators are much closer to the ideal results and can be readily utilized to obtain cluster labels. Our method exploits the discriminative information and feature correlation in a joint framework. For the sake of feature selection, the feature selection matrix is constrained to be sparse in rows, which is formulated as $\ell_{2,1}$-norm minimization term. To solve the proposed problem, a simple yet effective iterative algorithm is proposed. Extensive experiments are conducted on different datasets, which show that the proposed approach outperforms the state-of-the-arts in different applications.

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Nonnegative Discriminative Feature Selection

Preliminaries

We first summarize some notations. Throughout this paper, we use bold uppercase characters to denote matrices, bold lowercase characters to denote vectors. For an arbitrary matrix $\mathbf{A}$, $\mathbf{a}_i$ means the i-th row vector of $\mathbf{A}$, A_{ij} denotes the (i,j)-th entry of $\mathbf{A}$, $\|\mathbf{A}\|_F$ is Frobenius norm of $\mathbf{A}$ and Tr[$\mathbf{A}$] is the trace of $\mathbf{A}$ if $\mathbf{A}$ is square. For any $\mathbf{A} \in \mathcal{R}^{r\times t}$, its $\ell_{2,1}$-norm is defined as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{r} \sqrt{\sum_{j=1}^{t} A_{ij}^2}. \quad (1)$$

Assume that we have n samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ denote the data matrix, in which $\mathbf{x}_i \in \mathcal{R}^d$ is the feature descriptor of the i-th sample. Suppose these n samples are sampled from c classes. Denote $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \{0,1\}^{n\times c}$, where $\mathbf{y}_i \in \{0,1\}^{c\times 1}$ is the cluster indicator vector for $\mathbf{x}_i$. The same as (Yang et al. 2011b), the scaled cluster indicator matrix $\mathbf{F}$ is defined as

$$\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n]^T = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}, \quad (2)$$

where $\mathbf{f}_i$ is the scaled cluster indicator of $\mathbf{x}_i$. It turns out that

$$\mathbf{F}^T\mathbf{F} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}_c, \quad (3)$$

where $\mathbf{I}_c \in \mathcal{R}^{c\times c}$ is an identity matrix.

The Objective Function

In this work, we propose a general approach for spectral analysis-based feature selection. To select the discriminative features for unsupervised learning, we propose to utilize the cluster labels (which can be regarded as pseudo class labels) based on the data structure. Spectral clustering techniques have been demonstrated effective methods to detect the cluster structure of data and have received significant research attention recently (Shi and Malik 2000; Ng, Jordan, and Weiss 2001). Therefore, we make use of spectral clustering to learn the pseudo class labels, which are leveraged to guide the process of inferring the feature selection matrix. In our framework, the features which are most related to the pseudo class labels are selected. To this end, we assume that there is a linear transformation between features and pseudo labels. We propose to learn the scaled cluster indicator matrix $\mathbf{F} \in \mathcal{R}^{n\times c}$ and the feature selection matrix $\mathbf{W} \in \mathcal{R}^{d\times c}$ simultaneously.

Given a spectral clustering method with criterion $\mathcal{J}(\mathbf{F})$, we propose to optimize the following objective function for feature selection:

$$\min_{\mathbf{F},\mathbf{W}} \mathcal{J}(\mathbf{F}) + \alpha(\|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}) \quad \text{s.t.} \quad \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}, \quad (4)$$

where α and β are parameters. In (4), the $\ell_{2,1}$-norm regularization term is imposed to ensure $\mathbf{W}$ sparse in rows. In that way, the proposed method is able to handle correlated and noisy features (Kong, Ding, and Huang 2011; Nie et al. 2010). Let $\mathbf{w}_j$ denote the j-th row of $\mathbf{W}$. The joint minimization of the regression model and $\ell_{2,1}$-norm regularization term enables $\mathbf{W}$ to evaluate the correlation between pseudo labels and features, making it particularly suitable for feature selection. More specifically, $\mathbf{w}_j$ shrinks to zero if the j-th feature is less correlated to the pseudo labels $\mathbf{F}$. Therefore, the features corresponding to zero rows of $\mathbf{W}$ will be discarded when performing feature selection.

Clearly, an effective cluster indicator matrix is more capable to reflect the discriminative information of the input data. The local geometric structure of data plays an important role in data clustering, which has been exploited by many spectral clustering algorithms (Shi and Malik 2000; Yu and Shi 2003). Note that there are many different algorithms to uncover local data structure. In this work, we use the strategy proposed in (Shi and Malik 2000; Belkin and Niyogi 2001; Yu and Shi 2003) to be the criterion for its simplicity. The local geometric structure can be effectively modeled by a nearest neighbor graph on a scatter of data points. To construct the affinity graph $\mathbf{S}$, we define

$$S_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}) & \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{N}_k(\mathbf{x})$ is the set of k-nearest neighbors of $\mathbf{x}$. The local geometrical structure can be utilized by minimizing the following (Shi and Malik 2000; Yu and Shi 2003):

$$\min_{\mathbf{F}} \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}\|\frac{\mathbf{f}_i}{\sqrt{A_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{A_{jj}}}\|_2^2 = \text{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}], \quad (5)$$

where $\mathbf{A}$ is a diagonal matrix with $A_{ii} = \sum_{j=1}^n S_{ij}$ and $\mathbf{L} = \mathbf{A}^{-1/2}(\mathbf{A} - \mathbf{S})\mathbf{A}^{-1/2}$ is the normalized graph Laplacian matrix. Therefore $\mathcal{J}(\mathbf{F})$ is defined as

$$\mathcal{J}(\mathbf{F}) = \text{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}]. \quad (6)$$

Combining (4) and (6), we have

$$\min_{\mathbf{F},\mathbf{W}} \text{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha(\|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}) \quad \text{s.t.} \quad \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}. \quad (7)$$

According to the definition of $\mathbf{F}$, its elements are constrained to be discrete values, making the optimization of (7) an NP-hard problem (Shi and Malik 2000). A well-known solution is to relax it from discrete values to continuous ones (Shi and Malik 2000; Yu and Shi 2003), i.e., the objective function (7) is relaxed to

$$\min_{\mathbf{F},\mathbf{W}} \text{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha(\|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}) \quad \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \quad (8)$$

where the orthogonal constraint shown in (3) is kept. In (8), the first term learns the pseudo class labels using spectral analysis while the second term and the third term try to learn the feature selection matrix by a regression model with $\ell_{2,1}$-norm regularization.

Note that all the elements of $\mathbf{F}$ are nonnegative by definition. However, the optimal $\mathbf{F}$ of (8) has mixed signs, which violates its definition. In addition, since we have no discrete

(a) Mixed-signs $\mathbf{F}$ (b) Nonnegative $\mathbf{F}$ (c) $\mathbf{W}$

Figure 1: The visualization of the learned $\mathbf{F}$ and $\mathbf{W}$. (a) and (b): Each row is a sample and each column is a cluster indicator vector. (c): Each row is the ℓ_2-norm value of each row of $\mathbf{W}$. The results are normalized for a clearer illustration. The data used are from the JAFFE dataset.

process, the mixed signs make $\mathbf{F}$ severely deviate from the ideal cluster indicators. As a result, we cannot directly assign labels to data using the cluster indicator matrix $\mathbf{F}$. To address this problem, it is natural and reasonable to impose a nonnegative constraint into the objective function. When both nonnegative and orthogonal constraints are satisfied, there is only one element in each row of $\mathbf{F}$ is greater than zero and all of the others are zeros. In that way, the learned $\mathbf{F}$ is more accurate, and more capable to provide discriminative information. Therefore, we rewrite (8) and the objective function of NDFS is given by

$$\min_{\mathbf{F},\mathbf{W}} \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha(\|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}) \quad \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c,\ \mathbf{F} \geq 0. \tag{9}$$

It is worth noting that we adopt $\mathbf{L}$ defined in (5) for simplicity while other sophisticated Laplacian matrices e.g., the one proposed in (Yang et al. 2011a), can be used here as well.

Next, we take the JAFFE dataset (Lyons, Budynek, and Akamatsu 1999) as an example to illustrate the effectiveness of the nonnegative constraint and $\ell_{2,1}$-norm regularization term in the objective function (9). In Fig. 1 (a) and Fig. 1 (b), we plot the normalized absolute values of the optimal $\mathbf{F}$ corresponding to (8) and (9), respectively. From Fig. 1 (a), we can see that it is unclear how to directly assign cluster labels according to $\mathbf{F}$ without nonnegative constraint. It can be observed from Fig. 1 (b) that in each row of $\mathbf{F}$, only one element is positive and all of the others are 0, when nonnegative and orthogonal constraints are satisfied. Thus, cluster labels of the input data can be readily obtained according to $\mathbf{F}$. With the accurate cluster labels, NDFS is able to exploit the discriminative information. The $\ell_{2,1}$-norm minimization enforces $\mathbf{W}$ sparse in rows, as shown in Fig. 1 (c).

Optimization Algorithm

An Efficient Iterative Algorithm

In this subsection, we present an iterative algorithm to solve the optimization problem of NDFS. The $\ell_{2,1}$-norm regularization term is non-smooth and the objective function is not convex in $\mathbf{W}$ and $\mathbf{F}$ simultaneously. To optimize the objective function, we propose an iterative optimization algorithm. First, we rewrite the objective function of NDFS as follows

$$\min_{\mathbf{F},\mathbf{W}} \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha(\|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}) + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2 \quad \text{s.t.} \quad \mathbf{F} \geq 0. \tag{10}$$

where $\gamma > 0$ is a parameter to control the orthogonality condition. In practice, γ should be large enough to insure the orthogonality satisfied. For the ease of representation, let us define

$$\mathscr{L}(\mathbf{F},\mathbf{W}) = \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha(\|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}) + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2. \tag{11}$$

Setting $\frac{\partial \mathscr{L}(\mathbf{F},\mathbf{W})}{\partial \mathbf{W}} = 0$, we have

$$\frac{\partial \mathscr{L}(\mathbf{F},\mathbf{W})}{\partial \mathbf{W}} = 2\alpha(\mathbf{X}(\mathbf{X}^T\mathbf{W} - \mathbf{F}) + \beta\mathbf{D}\mathbf{W}) = 0 \Rightarrow \mathbf{W} = (\mathbf{X}\mathbf{X}^T + \beta\mathbf{D})^{-1}\mathbf{X}\mathbf{F}. \tag{12}$$

Here $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \frac{1}{2\|\mathbf{w}_i\|_2}$. [1] Substituting $\mathbf{W}$ by (12), the problem (10) is rewritten as

$$\min_{\mathbf{F},\mathbf{W}} \mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2 \quad \text{s.t.} \quad \mathbf{F} \geq 0, \tag{13}$$

where $\mathbf{M} = \mathbf{L} + \alpha(\mathbf{I}_n - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \beta\mathbf{D})^{-1}\mathbf{X})$ and $\mathbf{I}_n \in \mathcal{R}^{n\times n}$ is an identity matrix. Following (Lee and Seung 1999; 2001; Liu, Jin, and Yang 2006), we introduce multiplicative updating rules. Letting ϕ_{ij} be the Lagrange multiplier for constraint $F_{ij} \geq 0$ and $\Phi = [\phi_{ij}]$, the Lagrange function is

$$\mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2 + \mathrm{Tr}(\Phi\mathbf{F}^T). \tag{14}$$

Setting its derivative with respect to F_{ij} to 0 and using the Karush-Kuhn-Tuckre (KKT) condition (Kuhn and Tucker 1951) $\phi_{ij}F_{ij} = 0$, we obtain the updating rules:

$$F_{ij} \leftarrow F_{ij}\frac{(\gamma\mathbf{F})_{ij}}{(\mathbf{M}\mathbf{F} + \gamma\mathbf{F}\mathbf{F}^T\mathbf{F})_{ij}}. \tag{15}$$

Then, we normalize $\mathbf{F}$ such that $(\mathbf{F}^T\mathbf{F})_{ii} = 1, i = 1, \cdots, n$. Based on the above analysis, we summarize the detailed optimization algorithm in Algorithm 1.

Convergence Analysis

In this subsection, we prove the convergence of the proposed iterative procedure in Algorithm 1.

Theorem 1 *The alternate updating rules in Algorithm 1 monotonically decrease the objective function value of (10) in each iteration.*

Proof: In the iterative procedure, for $\mathbf{F}$ and $\mathbf{W}$ we update one while keeping the other one fixed. For convenience, let us denote

$$h(\mathbf{F}) = \mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \frac{\gamma}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2. \tag{16}$$

[1] In practice, $\|\mathbf{w}_i\|_2$ could be close to zero but not zero. Theoretically, it could be zeros. For this case, we can regularize $D_{ii} = \frac{1}{2\sqrt{(\mathbf{w}_i^T\mathbf{w}_i + \epsilon)}}$, where ϵ is very small constant.

Algorithm 1 Nonnegative Discriminative Feature Selection

Input:

Data matrix $\mathbf{X} \in \mathcal{R}^{d\times n}$; Parameters α, β, γ, k, c and p

1: Construct the k-nearest neighbor graph and calculate $\mathbf{L}$;
2: The iteration step $t = 1$; Initialize $\mathbf{F}^t \in \mathcal{R}^{n\times c}$ and set $\mathbf{D}^t \in \mathcal{R}^{d\times d}$ as an identity matrix;
3: **repeat**
4: $\quad \mathbf{M}^t = \mathbf{L} + \alpha(\mathbf{I}_n - \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \beta\mathbf{D}^t)^{-1}\mathbf{X})$;
5: $\quad F_{ij}^{t+1} = F_{ij}^t \frac{(\gamma\mathbf{F}^t)_{ij}}{(\mathbf{M}^t\mathbf{F}^t + \gamma\mathbf{F}^t(\mathbf{F}^t)^T\mathbf{F}^t)_{ij}}$;
6: $\quad \mathbf{W}^{t+1} = (\mathbf{X}\mathbf{X}^T + \beta\mathbf{D}^t)^{-1}\mathbf{X}\mathbf{F}^{t+1}$;
7: $\quad$ Update the diagonal matrix $\mathbf{D}$ as

$$\mathbf{D}^{t+1} = \begin{bmatrix} \frac{1}{2\|\mathbf{w}_1^t\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\|\mathbf{w}_d^t\|_2} \end{bmatrix};$$

8: $\quad$ t=t+1;
9: **until** Convergence criterion satisfied

Output:

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

With $\mathbf{W}^t$ fixed, we have $\mathscr{L}(\mathbf{F}^t, \mathbf{W}^t) = h(\mathbf{F}^t)$. By introducing an auxiliary function of h as in (Lee and Seung 1999; 2001), it is easy to prove $h(\mathbf{F}^{t+1}) \leq h(\mathbf{F}^t)$. Thus, we have

$$\mathscr{L}(\mathbf{F}^{t+1}, \mathbf{W}^t) \leq \mathscr{L}(\mathbf{F}^t, \mathbf{W}^t). \tag{17}$$

It can easily verified that Eq. (12) is the solution to the following problem.

$$\min_{\mathbf{W}} \|\mathbf{X}^T\mathbf{W} - \mathbf{F}\|_F^2 + \beta\mathrm{Tr}[\mathbf{W}^T\mathbf{D}\mathbf{W}] \tag{18}$$

Accordingly, in the t-th iteration, with $\mathbf{F}^t$ fixed we have

$$\begin{aligned} &\mathbf{W}^{t+1} = \min_{\mathbf{W}} \|\mathbf{X}^T\mathbf{W} - \mathbf{F}^t\|_F^2 + \beta\mathrm{Tr}[\mathbf{W}^T\mathbf{D}^t\mathbf{W}] \\ \Rightarrow\ &\|\mathbf{X}^T\mathbf{W}^{t+1} - \mathbf{F}^t\|_F^2 + \beta\mathrm{Tr}[(\mathbf{W}^{t+1})^T\mathbf{D}^t\mathbf{W}^{t+1}] \\ \leq\ &\|\mathbf{X}^T\mathbf{W}^t - \mathbf{F}^t\|_F^2 + \beta\mathrm{Tr}[(\mathbf{W}^t)^T\mathbf{D}^t\mathbf{W}^t]. \end{aligned} \tag{19}$$

That is to say,

$$\begin{aligned} &\|\mathbf{X}^T\mathbf{W}^{t+1} - \mathbf{F}^t\|_F^2 + \beta\sum_i \frac{\|\mathbf{w}_i^{t+1}\|_2^2}{2\|\mathbf{w}_i^t\|_2} \\ \leq\ &\|\mathbf{X}^T\mathbf{W}^t - \mathbf{F}^t\|_F^2 + \beta\sum_i \frac{\|\mathbf{w}_i^t\|_2^2}{2\|\mathbf{w}_i^t\|_2} \\ \Rightarrow\ &\|\mathbf{X}^T\mathbf{W}^{t+1} - \mathbf{F}^t\|_F^2 + \beta\|\mathbf{W}^{t+1}\|_{2,1} \\ &-\beta(\|\mathbf{W}^{t+1}\|_{2,1} - \sum_i \frac{\|\mathbf{w}_i^{t+1}\|_2^2}{2\|\mathbf{w}_i^t\|_2}) \\ \leq\ &\|\mathbf{X}^T\mathbf{W}^t - \mathbf{F}^t\|_F^2 + \beta\|\mathbf{W}^t\|_{2,1} \\ &-\beta(\|\mathbf{W}^t\|_{2,1} - \sum_i \frac{\|\mathbf{w}_i^t\|_2^2}{2\|\mathbf{w}_i^t\|_2}). \end{aligned} \tag{20}$$

According to the Lemmas in (Nie et al. 2010), $\sqrt{a} - \frac{a}{2\sqrt{b}} \leq \sqrt{b} - \frac{b}{2\sqrt{b}}$ and $\|\mathbf{W}^{t+1}\|_{2,1} - \sum_i \frac{\|\mathbf{w}_i^{t+1}\|_2^2}{2\|\mathbf{w}_i^t\|_2} \leq \|\mathbf{W}^t\|_{2,1} - \sum_i \frac{\|\mathbf{w}_i^t\|_2^2}{2\|\mathbf{w}_i^t\|_2}$. Thus, we obtain

$$\begin{aligned} &\|\mathbf{X}^T\mathbf{W}^{t+1} - \mathbf{F}^t\|_F^2 + \beta\|\mathbf{W}^{t+1}\|_{2,1} \\ \leq\ &\|\mathbf{X}^T\mathbf{W}^t - \mathbf{F}^t\|_F^2 + \beta\|\mathbf{W}^t\|_{2,1}. \end{aligned} \tag{21}$$

Therefore, according to Eq. (11), we arrive at

$$\mathscr{L}(\mathbf{F}^t, \mathbf{W}^{t+1}) \leq \mathscr{L}(\mathbf{F}^t, \mathbf{W}^t). \tag{22}$$

Based on Eq. (17) and Eq. (22), we obtain

$$\mathscr{L}(\mathbf{F}^{t+1}, \mathbf{W}^{t+1}) \leq \mathscr{L}(\mathbf{F}^{t+1}, \mathbf{W}^t) \leq \mathscr{L}(\mathbf{F}^t, \mathbf{W}^t). \tag{23}$$

Thus, $\mathscr{L}(\mathbf{F}, \mathbf{W})$ monotonically decreases using the updating rules in Algorithm 1 and Theorem 1 is proved. ■

According to Theorem 1, we can see that the iterative approach in Algorithm 1 converges to local optimal $\mathbf{F}$ and $\mathbf{W}$. The proposed optimization algorithm is efficient. In the experiment, we observe that our algorithm usually converges around only 30 iterations.

Discussions

To exploit the discriminative information in unsupervised scenarios, clustering-based feature selection is also studied in Multi-Cluster Feature Selection (MCFS) (Cai, Zhang, and He 2010). MCFS uses a two-step strategy to select features according to spectral clustering. The first step is to learn $\mathbf{F}$ using spectral clustering and then $\mathbf{W}$ is learned by a regression model with ℓ_1-norm regularization in the second step. However, it ignores the nonnegative constraint, increasing difficulty in getting the cluster labels. The mixed signs from eigenvalue decomposition make $\mathbf{F}$ deviate from the ideal solution as shown in Fig. 1 (a). Our NDFS algorithm differs MCFS from the following aspects. First, the proposed NDFS is a one-step algorithm and learns $\mathbf{F}$ and $\mathbf{W}$ simultaneously. When $\alpha \to 0$, our method leads to a two-step algorithm for feature selection. The first step is spectral clustering and the second step is a regression model with $\ell_{2,1}$-norm regularization. Thus, NDFS is more general. Second, $\mathbf{F}$ is constrained to be nonnegative. When both nonnegative and orthogonal constraints are satisfied, only one element in each row of $\mathbf{F}$ is positive and all the others are 0, which is much closer to the ideal clustering result, and the solution can be directly obtained without discretization. Finally, in our framework, we perform clustering and feature selection simultaneously, which explicitly enforces that $\mathbf{F}$ can be linearly approximated by the selected features, making the results more accurate. The experimental results in Section 5 demonstrate that our NDFS is better that MCFS in a variety of applications.

In the optimization problem (10), if we do not constraint $\mathbf{F}$ to be nonnegative, when $\alpha \to +\infty$ and $\alpha\beta \nrightarrow +\infty$, we have $\mathbf{F} = \mathbf{X}^T\mathbf{W}$ and the following objective function.

$$\min_{\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}=\mathbf{I}_c} \mathrm{Tr}[\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}] + \alpha\beta\|\mathbf{W}\|_{2,1} \tag{24}$$

If we remove the nonnegative constraint, our objective function and that of Unsupervised Discriminative Feature Selection (UDFS) (Yang et al. 2011a) have similar fashions. In this extreme case, $\mathbf{F}$ is enforced to be linear, i.e., $\mathbf{F} =$

Table 1: Dataset Description.

Dataset	# of Samples	# of Features	# of Classes
UMIST	575	644	20
AT&T	400	644	40
JAFFE	213	676	10
Poingting4	2790	1120	15
MNIST	5000	784	10
BA	1404	320	36
WebKB	814	4029	7
Lung	203	12600	5

$\mathbf{X}^T\mathbf{W}$. However, as indicated in (Shi and Malik 2000), it is likely that $\mathbf{F}$ is nonlinear in many applications. Hence, NDFS is superior to UDFS due to its flexibility of linearity. Additionally, $\mathbf{F}$ is constrained to be nonnegative, making it more accurate than the one with mixed signs. Therefore, compared with UDFS, NDFS is more capable to select discriminative feature subset, which is also verified by our experiments.

Experimental Analysis

In this section, we conduct extensive experiments to evaluate the performance of the proposed NDFS, which can be applied to many applications, such as clustering and classification. Following previous unsupervised feature selection work (Cai, Zhang, and He 2010; Yang et al. 2011a), we only evaluate the performance of NDFS for feature selection in terms of clustering due to space limit.

Datasets

The experiments are conducted on 8 publicly available datasets, including four face image datasets, i.e., UMIST[2], AT&T (Samaria and Harter 1994), JAFFE (Lyons, Budynek, and Akamatsu 1999) and Pointing4 (Gourier, Hall, and Crowley 2004), two handwritten digit datasets, i.e., a subset of MNIST[3] and Binary Alphabet (BA)[4], one text database WebKB collected by the University of Texas (Craven et al. 1998), and one cancer database Lung (Hong and Yang 1991). Datasets from different areas serve as a good test bed for a comprehensive evaluation. Table 1 summarizes the details of the datsets used in the experiments.

Experimental Settings

To validate the effectiveness of NDFS for feature selection, we compare it with the following unsupervised feature selection methods.

1. **Baseline**: All original features are adopted;
2. **MaxVar**: Features corresponding to the maximum variance are selected to obtain the best expressive features;
3. **LS**: Features consistent with Gaussian Laplacian matrix are selected to best preserve the local manifold structure (He, Cai, and Niyogi 2005);
4. **SPEC**: Features are selected using spectral regression (Zhao and Liu 2007);
5. **MCFS**: Features are selected based on spectral analysis and sparse regression problem (Cai, Zhang, and He 2010);
6. **UDFS**: Features are selected by a joint framework of discriminative analysis and $\ell_{2,1}$-norm minimization (Yang et al. 2011a).

With the selected features, we evaluate the performance in terms of clustering by two widely used evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI) (Cai, Zhang, and He 2010; Yang et al. 2011a). The larger ACC and NMI are, the better performance is.

There are some parameters to be set in advance. For LS, MCFS, UDFS and NDFS, we set $k = 5$ for all the datasets to specify the size of neighborhoods. For NDFS, to guarantee the orthogonality satisfied, we fix $\gamma = 10^8$ in our experiments. To fairly compare different unsupervised feature selection algorithms, we tune the parameters for all methods by a "grid-search" strategy from $\{10^{-6}, 10^{-4}, \cdots, 10^{6}\}$. The numbers of selected features are set as $\{50, 100, 150, 200, 250, 300\}$ for all the datasets. For all the algorithms, we report the the best clustering results from the optimal parameters. Different parameters may be used for different databases. In our experiments, we adopt Kmeans algorithm to cluster samples based on the selected features. The performance of Kmeans clustering depends on initialization. Following (Cai, Zhang, and He 2010; Yang et al. 2011a), we repeat the clustering 20 times with random initialization for each setup. The average results with standard deviation (std) are reported.

Results and Analysis

We summarize the clustering results of different methods on the 8 datasets in Table 2 and Table 3. From the two tables, we have the following observations. First, feature selection is necessary and effective. It can not only significantly reduce the feature number and make the algorithms more efficient, but also improve the performance. Second, the local structure of data distribution is crucial for feature selection, which is consistent with the observations in (He, Cai, and Niyogi 2005; Yang et al. 2011a). Except for MaxVar, all the other approaches consider the local structure of data distribution and yield better performance. Third, the discriminative information is crucial for unsupervised learning. MCFS, UDFS and NDFS exploit discriminative information, which results in more accurate clustering. Finally, UDFS and NDFS achieve higher ACC and NMI by evaluating features jointly than others that select features one after another or using two-step strategies. As shown in Table 2 and Table 3, NDFS achieves best performance on all datasets, which verifies that the proposed NDFS algorithm is able to select more informative features. The is mainly due to the following reasons. First, NDFS learns the pseudo class label indicators and the feature selection matrix simultaneously. It enables NDFS to select discriminative features in unsupervised learning. Second, the local structure of data and the correlation among features are explored simultaneously. Third, the $\ell_{2,1}$ regularization term is able to reduce the

[2]http://www.sheffield.ac.uk/eee/research/iel/research/face

[3]http://yann.lecun.com/exdb/mnist/

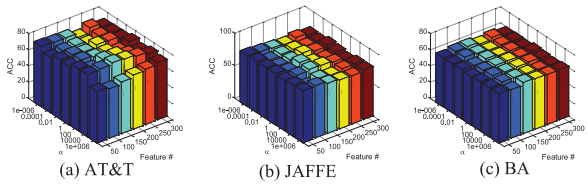
[4]http://www.cs.nyu.edu/∼roweis/data.html

Table 2: Clustering results (ACC%±std) of different feature selection algorithms on different datasets. The best results are highlighted in bold.

Dataset	UMIST	AT&T	JAFFE	Pointing4	MNIST	BA	WebKB	Lung
Baseline	41.8 ± 2.7	59.2 ± 3.4	72.5 ± 9.2	35.9 ± 2.2	52.2 ± 5.0	40.3 ± 2.0	56.7 ± 2.7	56.8 ± 3.7
MaxVar	45.8 ± 2.8	58.6 ± 3.4	67.3 ± 5.8	44.0 ± 2.8	53.3 ± 2.7	40.7 ± 1.7	54.6 ± 2.8	57.2 ± 4.1
LS	45.9 ± 2.9	60.6 ± 2.9	74.0 ± 7.6	37.1 ± 1.6	54.3 ± 4.8	42.1 ± 1.7	56.8 ± 2.9	59.5 ± 7.7
SPEC	47.9 ± 3.0	62.1 ± 3.3	76.9 ± 7.2	38.6 ± 2.2	55.6 ± 5.2	42.2 ± 2.2	61.1 ± 2.8	59.5 ± 4.0
MCFS	46.3 ± 3.6	61.0 ± 4.8	78.8 ± 9.1	46.2 ± 2.9	56.5 ± 4.1	41.5 ± 1.8	61.3 ± 2.3	60.6 ± 4.5
UDFS	48.6 ± 3.7	62.4 ± 2.8	76.7 ± 7.1	45.1 ± 2.4	56.6 ± 4.2	42.7 ± 1.8	61.7 ± 3.2	61.3 ± 4.7
NDFS	$\mathbf{51.3 \pm 3.9}$	$\mathbf{64.5 \pm 3.4}$	$\mathbf{81.2 \pm 8.1}$	$\mathbf{48.9 \pm 3.2}$	$\mathbf{58.2 \pm 3.2}$	$\mathbf{43.4 \pm 2.0}$	$\mathbf{62.4 \pm 3.0}$	$\mathbf{65.6 \pm 5.1}$

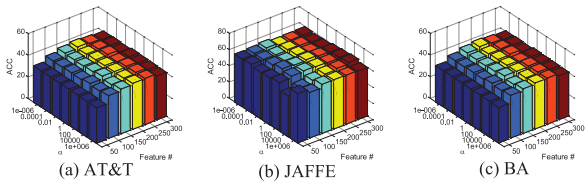
Table 3: Clustering results (NMI%±std) of different feature selection algorithms on different datasets. The best results are highlighted in bold.

Dataset	UMIST	AT&T	JAFFE	Pointing4	MNIST	BA	WebKB	Lung
Baseline	62.3 ± 2.3	79.3 ± 1.7	80.0 ± 5.7	41.7 ± 1.4	47.8 ± 2.3	56.5 ± 1.3	11.4 ± 5.0	39.4 ± 5.5
MaxVar	63.5 ± 1.5	78.5 ± 1.5	70.3 ± 4.2	50.8 ± 1.8	48.6 ± 1.1	56.9 ± 1.3	17.1 ± 1.4	38.1 ± 5.0
LS	63.9 ± 1.8	80.0 ± 1.4	79.4 ± 7.0	42.7 ± 1.2	48.6 ± 2.0	57.3 ± 0.8	10.6 ± 4.0	41.4 ± 6.0
SPEC	65.2 ± 2.0	80.2 ± 1.8	82.8 ± 3.8	40.5 ± 1.0	49.7 ± 2.0	57.9 ± 1.1	17.2 ± 3.1	33.5 ± 1.5
MCFS	66.7 ± 1.9	80.3 ± 2.5	83.4 ± 5.0	53.1 ± 1.1	50.0 ± 1.8	57.5 ± 0.8	17.6 ± 0.8	40.1 ± 3.1
UDFS	67.3 ± 3.0	80.8 ± 1.2	82.3 ± 6.5	52.4 ± 1.7	50.8 ± 1.6	58.1 ± 1.0	18.1 ± 3.3	42.8 ± 3.9
NDFS	$\mathbf{69.7 \pm 2.3}$	$\mathbf{82.2 \pm 1.6}$	$\mathbf{86.3 \pm 7.1}$	$\mathbf{56.4 \pm 1.3}$	$\mathbf{51.8 \pm 1.3}$	$\mathbf{58.8 \pm 0.8}$	$\mathbf{18.7 \pm 1.6}$	$\mathbf{45.3 \pm 2.9}$



(a) AT&T (b) JAFFE (c) BA

Figure 2: Clustering accuracy (ACC) of NDFS with different α and feature numbers while keeping $\beta = 100$.

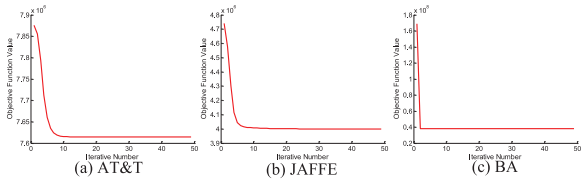


(a) AT&T (b) JAFFE (c) BA

Figure 3: Clustering accuracy (ACC) of NDFS with different β and feature numbers while keeping $\alpha = 1$.

redundant and noisy features. While both NDFS and UDFS utilize $\ell_{2,1}$ regularization term for unsupervised feature selection, we additionally impose the nonnegative constraint into the objective function, making the cluster indicators more accurate and the selected feature more informative.

Next, we study the sensitiveness of parameters and the convergence of NDFS. Due to the space limit, we only report the results in terms of ACC and objective values over AT&T, JAFFE and BA datasets. The experimental results are shown in Fig. 2 and Fig. 3. Fig. 4 shows convergence curves of NDFS. From these figures, we can see that our method is not sensitive to α and β with wide ranges, and that the proposed optimization algorithm is effective and converges quickly.



(a) AT&T (b) JAFFE (c) BA

Figure 4: Convergence curve of NDFS over AT&T, JAFFE and BA datasets.

Conclusion

In this paper, we propose a novel unsupervised feature selection approach, which jointly exploits nonnegative spectral analysis and feature selection. The cluster labels learned by spectral clustering are used to guide feature selection. The cluster indicator matrix and the feature selection matrix are iteratively learned. To select discriminative features, we impose the nonnegative constrain on the scaled cluster indicator matrix and $\ell_{2,1}$-norm minimization regularization on the feature selection matrix. Our method is able to select the discriminative features that yield better results. Extensive experiments on different real world datasets have validated the effectiveness of the proposed method.

Acknowledgments

This work was supported by 973 Program (Project No. 2010CB327905), the National Natural Science Foundation of China (Grant No. 60835002, 60903146), and by the National Science Foundation under Grants No. IIS-0917072. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *KDD*.

Craven, M.; DiPasquo, D.; Freitag, D.; McCallum, A.; Mitchell, T. M.; Nigam, K.; and Slattery, S. 1998. Learning to extract symbolic knowledge from the world wide web. In *AAAI/IAAI*.

Duda, R.; Hart, P.; and Stork, D. 2001. *Pattern Recognition (2nd Edition)*. New York, USA: John Wiley & Sons.

Gourier, N.; Hall, D.; and Crowley, J. 2004. Estimating face orientation from robust detection of salient facial features. In *ICPR Workshop on Visual Observation of Deictic Gestures*.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *JMLR* 3:1157–1182.

He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *NIPS*.

Hong, Z., and Yang, J. 1991. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition* 24(4):317–324.

Jain, A., and Zongker, D. 1997. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. on PAMI* 19:153–158.

Kong, D.; Ding, C.; and Huang, H. 2011. Robust nonnegative matrix factorization using l21-norm. In *CIKM*.

Kuhn, H., and Tucker, A. 1951. Nonlinear programming. In Berkeley Symposium on Mathematical Statistics and Probabilistics.

Lee, D., and Seung, H. 1999. Learning the parts of objects by nonnegative matrix factorization. *Nature* 401:788–791.

Lee, D., and Seung, H. 2001. Algorithms for nonnegative matrix factorization. In *NIPS*.

Liu, Y.; Jin, R.; and Yang, L. 2006. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*.

Liu, H.; Wu, X.; and Zhang, S. 2011. Feature selection using hierarchical feature clustering. In *CIKM*.

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

Masaeli, M.; Fung, G.; and Dy, J. G. 2010. From transformation-based dimensionality reduction to feature selection. In *ICML*.

Ng, A. Y.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. In *NIPS*.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*.

Rakotomamonjy, A. 2003. Variable selection using svm-based criteria. *JMLR* 3:1357–1370.

Samaria, F., and Harter, A. 1994. Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. on PAMI* 22:888–905.

Vapnik, V. 1998. *Statistical Learning Therory*. New York.

Wang, M.; Hua, X.-S.; Hong, R.; Tang, J.; Qi, G.-J.; and Song, Y. 2009a. Unified video annotation via multi-graph learning. *IEEE Trans. CSVT* 19(5):733–746.

Wang, M.; Hua, X.-S.; Tang, J.; and Hong, R. 2009b. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Trans. Multimedia* 11(3):465–476.

Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; and Zhou, X. 2011a. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*.

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

Yu, S. X., and Shi, J. 2003. Multiclass spectral clustering. In *ICCV*.

Zhao, Z., and Liu, H. 2007. Sprectral feature selection for supervised and unsupervised learning. In *ICML*.

Zhu, J.; Rosset, S.; Hastie, T.; and Tibshirani, R. 2003. 1-norm support vector machines. In *NIPS*.