$\ell_{2,1}$-Norm Regularized Discriminative Feature Selection for Unsupervised Learning

Yi Yang[1], Heng Tao Shen[1], Zhigang Ma[2], Zi Huang[1], Xiaofang Zhou[1]

[1]School of Information Technology & Electrical Engineering, The University of Queensland.
[2]Department of Information Engineering & Computer Science, University of Trento.
yangyi_zju@yahoo.com.cn, shenht@itee.uq.edu.au, ma@disi.unitn.it, {huang, zxf}@itee.uq.edu.au.

Abstract

Compared with supervised learning for feature selection, it is much more difficult to select the discriminative features in unsupervised learning due to the lack of label information. Traditional unsupervised feature selection algorithms usually select the features which best preserve the data distribution, e.g., manifold structure, of the whole feature set. Under the assumption that the class label of input data can be predicted by a linear classifier, we incorporate discriminative analysis and $\ell_{2,1}$-norm minimization into a joint framework for unsupervised feature selection. Different from existing unsupervised feature selection algorithms, our algorithm selects the most discriminative feature subset from the whole feature set in batch mode. Extensive experiment on different data types demonstrates the effectiveness of our algorithm.

Introduction

In many areas, such as computer vision, pattern recognition and biological study, data are represented by high dimensional feature vectors. Feature selection aims to select a subset of features from the high dimensional feature set for a compact and accurate data representation. It has twofold role in improving the performance for data analysis. First, the dimension of selected feature subset is much lower, making the subsequential computation on the input data more efficient. Second, the noisy features are eliminated for a better data representation, resulting in a more accurate clustering and classification result. During recent years, feature selection has attracted much research attention. Several new feature selection algorithms have been proposed with a variety of applications.

Feature selection algorithms can be roughly classified into two groups, i.e., supervised feature selection and unsupervised feature selection. Supervised feature selection algorithms, e.g., Fisher score [Duda *et al.*, 2001] , robust regression [Nie *et al.*, 2010], sparse multi-output regression [Zhao *et al.*, 2010] and trace ratio [Nie *et al.*, 2008], usually select features according to labels of the training data. Because discriminative information is enclosed in labels, supervised feature selection is usually able to select discriminative features. In unsupervised scenarios, however, there is no label information directly available, making it much more difficult to select the discriminative features. A frequently used criterion in unsupervised learning is to select the features which best preserve the data similarity or manifold structure derived from the whole feature set [He *et al.*, 2005; Zhao and Liu, 2007; Cai *et al.*, 2010]. However, discriminative information is neglected though it has been demonstrated important in data analysis [Fukunaga, 1990].

Most of the traditional supervised and unsupervised feature selection algorithms evaluate the importance of each feature individually [Duda *et al.*, 2001; He *et al.*, 2005; Zhao and Liu, 2007] and select features one by one. A limitation is that the correlation among features is neglected [Zhao *et al.*, 2010; Cai *et al.*, 2010]. More recently, researchers have applied the two-step approach, i.e., spectral regression, to supervised and unsupervised feature selection [Zhao *et al.*, 2010; Cai *et al.*, 2010]. These efforts have shown that it is a better way to evaluate the importance of the selected features jointly. In this paper, we propose a new unsupervised feature selection algorithm by simultaneously exploiting discriminative information and feature correlations. Because we utilize local discriminative information, the manifold structure is considered too. While [Zhao *et al.*, 2010; Cai *et al.*, 2010] also select features in batch mode, our algorithm is a one-step approach and it is able to select the discriminative features for unsupervised learning. We also propose an efficient algorithm to optimize the problem.

The Objective Function

In this section, we give the objective function of the proposed Unsupervised Discriminative Feature Selection (UDFS) algorithm. Later in the next section, we propose an efficient algorithm to optimize the objective function. It is worth mentioning that UDFS aims to select the most discriminative features for data representation, where manifold structure is considered, making it different from the existing unsupervised feature selection algorithms.

Denote $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ as the training set, where $x_i \in \mathbb{R}^d (1 \leq i \leq n)$ is the i-th datum and n is the total

number of training data. In this paper, I is identity matrix. For a constant m, $\mathbf{1}_m \in \mathbb{R}^m$ is a column vector with all of its elements being 1 and $H_m = I - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T \in \mathbb{R}^{m\times m}$. For an arbitrary matrix $A \in \mathbb{R}^{r\times p}$, its $\ell_{2,1}$-norm is defined as

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

Suppose the n training data $x_1, x_2, ..., x_n$ are sampled from c classes and there are n_i samples in the i-th class. We define $y_i \in \{0,1\}^{c\times 1}(1 \le i \le n)$ as the label vector of x_i. The j-th element of y_i is 1 if x_i belongs to the j-th class, and 0 otherwise. $Y = [y_1, y_2, ..., y_n]^T \in \{0,1\}^{n\times c}$ is the label matrix. The total scatter matrix S_t and between class scatter matrix S_b are defined as follows [Fukunaga, 1990].

$$S_t = \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T \quad (2)$$

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

where μ is the mean of all samples, μ_i is the mean of samples in the i-th class, n_i is the number of samples in the i-th class, $\tilde{X} = XH_n$ is the data matrix after being centered, and $G = [G_1, ..., G_n]^T = Y(Y^TY)^{-1/2}$ is the scaled label matrix. A well-known method to utilize discriminative information is to find a low dimensional subspace in which S_b is maximized while S_t is minimized [Fukunaga, 1990].

Recently, some researchers proposed two different new algorithms to exploit local discriminative information [Sugiyama, 2006; Yang *et al.*, 2010b] for classification and image clustering, demonstrating that local discriminative information is more important than global one. Inspired by this, for each data point x_i, we construct a local set $\mathcal{N}_k(x_i)$ comprising x_i and its k nearest neighbors $x_{i_1}, ..., x_{i_k}$. Denote $X_i = [x_i, x_{i_1}, ..., x_{i_k}]$ as the local data matrix. Similar to (2) and (3), the local total scatter matrix $S_t^{(i)}$ and between class scatter matrix $S_b^{(i)}$ of $\mathcal{N}_k(x_i)$ are defined as follows.

$$S_t^{(i)} = \tilde{X}_i{\tilde{X}_i}^T; \quad (4)$$

$$S_b^{(i)} = \tilde{X}_iG_{(i)}G_{(i)}^T{\tilde{X}_i}^T, \quad (5)$$

where $\tilde{X}_i = X_iH_{k+1}$ and $G_{(i)} = [G_i, G_{i_1}, ..., G_{i_k}]^T$. For the ease of representation, we define the selection matrix $S_i \in \{0,1\}^{n\times(k+1)}$ as follows.

$$(S_i)_{pq} = \begin{cases} 1 & \text{if } p = F_i\{q\}; \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where $F_i = \{i, i_1, ..., i_k\}$. In this paper, it remains unclear how to define G because we are focusing on unsupervised learning where there is no label information available. In order to make use of local discriminative information, we assume there is a linear classifier $W \in \mathbb{R}^{d\times c}$ which classifies each data point to a class, i.e., $G_i = W^Tx_i$. Note that $G_i, G_{i_1}, ..., G_{i_k}$ are selected from G, i.e., $G_{(i)} = S_i^TG$. Then we have

$$G_{(i)} = [G_i, G_{i_1}, ..., G_{i_k}]^T = S_i^TG = S_i^TX^TW. \quad (7)$$

It is worth noting that the proposed algorithm is an unsupervised one. In other words, G defined in (7) is the output of the algorithm, i.e., $G_i = W^Tx_i$, but not provided by the human supervisors. If some rows of W shrink to zero, W can be regarded as the combination coefficients for different features that best predict the class labels of the training data. Next, we give the approach which learns a discriminative W for feature selection. Inspired by [Fukunaga, 1990; Yang *et al.*, 2010b], we define the local discriminative score DS_i of x_i as

$$\begin{aligned} DS_i &= Tr\left[(S_t^{(i)} + \lambda I)^{-1}S_b^{(i)}\right] \\ &= Tr\left[G_{(i)}^T{\tilde{X}_i}^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_iG_{(i)}\right] \\ &= Tr\left[W^TXS_i{\tilde{X}_i}^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_iS_i^TX^TW\right], \end{aligned} \quad (8)$$

where λ is a parameter and λI is added to make the term $(\tilde{X}_i\tilde{X}_i^T + \lambda I)$ invertible. Clearly, a larger DS_i indicates that W has a higher discriminative ability *w.r.t.* the datum x_i. We intend to train a W corresponding to the highest discriminative scores for all the training data $x_1, ..., x_n$. Therefore we propose to minimize (9) for feature selection.

$$\sum\nolimits_{i=1}^{n}\left\{Tr[G_{(i)}^TH_{k+1}G_{(i)}] - DS_i\right\} + \gamma\|W\|_{2,1} \quad (9)$$

Considering that the data number in each local set is usually small, $G_{(i)}^TH_{k+1}G_{(i)}$ is added in (9) to avoid overfitting. The regularization term $\|W\|_{2,1}$ controls the capacity of W and also ensures that W is sparse in rows, making it particularly suitable for feature selection. Substituting DS_i in (9) by (8), the objective function of our UDSF is given by

$$\begin{aligned} &\min_{W^TW=I}\sum\nolimits_{i=1}^{n} Tr\{W^TXS_iH_{k+1}S_i^TX^TW - \\ &[W^TXS_i{\tilde{X}_i}^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_iS_i^TX^TW]\} + \gamma\|W\|_{2,1} \end{aligned} \quad (10)$$

where the orthogonal constraint is imposed to avoid arbitrary scaling and avoid the trivial solution of all zeros. Note that the first term of (10) is equivalent to the following[1]:

$$Tr\{W^TX\{\sum_{i=1}^{n}[S_i(H_{k+1} - {\tilde{X}_i}^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_i)S_i^T]\}X^TW\}$$

Meanwhile we have

$$\begin{aligned} &H_{k+1} - {\tilde{X}_i}^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_i \\ =&H_{k+1} - H_{k+1}\tilde{X}_i^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_iH_{k+1} \\ =&H_{k+1} - H_{k+1}(\tilde{X}_i^T\tilde{X}_i + \lambda I)^{-1} \\ &(\tilde{X}_i^T\tilde{X}_i + \lambda I)\tilde{X}_i^T(\tilde{X}_i\tilde{X}_i^T + \lambda I)^{-1}\tilde{X}_iH_{k+1} \\ =&H_{k+1} - H_{k+1}(\tilde{X}_i^T\tilde{X}_i + \lambda I)^{-1}\tilde{X}_i^T\tilde{X}_iH_{k+1} \\ =&H_{k+1} - H_{k+1}(\tilde{X}_i^T\tilde{X}_i + \lambda I)^{-1}(\tilde{X}_i^T\tilde{X}_i + \lambda I - \lambda I)H_{k+1} \\ =&\lambda H_{k+1}(\tilde{X}_i^T\tilde{X}_i + \lambda I)^{-1}H_{k+1} \end{aligned}$$

Therefore, the objective function of UDFS is rewritten as

$$\min_{W^TW=I} Tr(W^TMW) + \gamma\|W\|_{2,1} \quad (11)$$

[1] It can be also interpreted in regression view [Yang *et al.*, 2010a].

where

$$M = X\left[\sum_{i=1}^{n}\left(S_i H_{k+1}(\tilde{X}_i^T\tilde{X}_i + \lambda I)^{-1}H_{k+1}S_i^T\right)\right]X^T \quad (12)$$

Denote w_i as the i-th row of W, i.e., $W = [w^1, ...w^d]^T$, the objective function shown in (11) can be also written as

$$\min_{W^TW=I} Tr(W^TMW) + \gamma\sum\nolimits_{i=1}^{d}\left\|w^i\right\|_2. \quad (13)$$

We can see that many rows of the optimal W corresponding to (13) shrink to zeros [2]. Consequently, for a datum x_i, $x_i' = W^Tx_i$ is a new representation of x_i using only a small set of selected features. Alternatively, we can rank each feature $f_i|_{i=1}^d$ according to $\left\|w^i\right\|_2$ in descending order and select top ranked features.

Optimization of UDFS Algorithm

The $\ell_{2,1}$-norm minimization problem has been studied in several previous works, such as [Argyriou *et al.*, 2008; Nie *et al.*, 2010; Obozinski *et al.*, 2008; Liu *et al.*, 2009; Zhao *et al.*, 2010; Yang *et al.*, 2011]. However, it remains unclear how to directly apply the existing algorithms to optimizing our objective function, where the orthogonal constraint $W^TW = I$ is imposed. In this section, inspired by [Nie *et al.*, 2010], we give a new approach to solve the optimization problem shown in (11) for feature selection. We first describe the detailed approach of UDFS algorithm in Algorithm 1 as follows.

Algorithm 1: The UDFS algorithm.

1 **for** $i = 1$ **to** n **do**
2 $\quad B_i = (\tilde{X}_i^T\tilde{X}_i + \lambda I)^{-1}$
3 $\quad M_i = S_iH_{k+1}B_iH_{k+1}S_i^T$;
4 $M = X\left(\sum_{i=1}^{n} M_i\right)X^T$;
5 Set $t = 0$ and initialize $D_0 \in \mathbb{R}^{d\times d}$ as an identity matrix;
6 **repeat**
7 $\quad P_t = M + \gamma D_t$;
8 $\quad W_t = [p_1, ..., p_c]$ where $p_1, ..., p_c$ are the eigenvectors of P_t corresponding to the first c smallest eigenvalues;
9 $\quad$ Update the diagonal matrix D_{t+1} as

$$D_{t+1} = \begin{bmatrix} \frac{1}{2\|w_t^1\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\|w_t^d\|_2} \end{bmatrix};$$

10 $\quad t = t + 1$;
11 **until** *Convergence*;
12 Sort each feature $f_i\ |_{i=1}^d$ according to $\left\|w_t^i\right\|_2$ in descending order and select the top ranked ones.

Below, we briefly analyze Algorithm-1 proposed in this section. From line 1 to line 4, it computes M defined in (12). From line 6 to line 11, it optimizes the objective function shown in (13). Next, we verify that the proposed iterative approach, i.e., line 6 to line 11 in Algorithm 1, converges to the optimal W corresponding to (13). We begin with the following two Lemmas.

Lemma 1. For any two non-zero constants a and b, the following inequality holds [Nie *et al.*, 2010].

$$\sqrt{a} - \frac{a}{2\sqrt{b}} \le \sqrt{b} - \frac{b}{2\sqrt{b}}. \quad (14)$$

$Proof$. The detailed proof is similar as that in [Nie *et al.*, 2010]. □

Lemma 2. The following inequality holds provided that $v_t^i|_{i=1}^r$ are non-zero vectors, where r is an arbitrary number [Nie *et al.*, 2010].

$$\sum_i \left\|v_{t+1}^i\right\|_2 - \sum_i \frac{\left\|v_{t+1}^i\right\|_2^2}{2\left\|v_t^i\right\|_2} \le \sum_i \left\|v_t^i\right\|_2 - \sum_i \frac{\left\|v_t^i\right\|_2^2}{2\left\|v_t^i\right\|_2} \quad (15)$$

$Proof$. Substituting a and b in (14) by $\left\|v_{t+1}^i\right\|_2^2$ and $\left\|v_t^i\right\|_2^2$ respectively, we can see that the following inequality holds for any i.

$$\left\|v_{t+1}^i\right\|_2 - \frac{\left\|v_{t+1}^i\right\|_2^2}{2\left\|v_t^i\right\|_2} \le \left\|v_t^i\right\|_2 - \frac{\left\|v_t^i\right\|_2^2}{2\left\|v_t^i\right\|_2} \quad (16)$$

Summing (16) over i, it can be seen that (15) holds □.

Next, we show that the iterative algorithm shown in Algorithm-1 converges by the following theorem.

Theorem 1. The iterative approach in Algorithm 1 (line 6 to line 11) monotonically decreases the objective function value of $\min_{W^TW=I} Tr(W^TMW) + \gamma\sum_{i=1}^{d}\left\|w^i\right\|_2$ in each iteration[3].

$Proof$. According to the definition of W_t in line 8 of Algorithm 1, we can see that

$$W_t = \arg\min_{W^TW=I} Tr[W^T(M + \gamma D_t)W] \quad (17)$$

Therefore, we have

$$Tr[W_t^T(M + \gamma D_t)W_t] \le Tr[W_{t-1}^T(M + \gamma D_t)W_{t-1}]$$

$$\Rightarrow Tr(W_t^TMW_t) + \gamma\sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_{t-1}^i\right\|_2} \le Tr(W_{t-1}^TMW_{t-1}) + \gamma\sum_i \frac{\left\|w_{t-1}^i\right\|_2^2}{2\left\|w_{t-1}^i\right\|_2}$$

[2] Usually, many rows of the optimal W are close to zeros.

[3] When computing D_{t+1}, its diagonal element $d_{ii} = \frac{1}{2\left\|w_t^i\right\|_2}$. It is worthy noting that in practice, $\left\|w_t^i\right\|_2$ could be very close to zero but not zero. However, $\left\|w_t^i\right\|_2$ can be zero theoretically. In this case, we can follow the traditional regularization way and define $d_{ii} = \frac{1}{2\left\|w_t^i\right\|_2+\varsigma}$, where ς is a very small constant. When $\varsigma \to 0$ it is easy to see that $\frac{1}{2\left\|w_t^i\right\|_2+\varsigma}$ approximates $\frac{1}{2\left\|w_t^i\right\|_2}$.

Then we have the following inequality

$$\begin{aligned} & Tr(W_t^T M W_t) + \gamma \sum_i \left\|w_t^i\right\|_2 \\ & \qquad -\gamma(\sum_i \left\|w_t^i\right\|_2 - \sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_{t-1}^i\right\|_2}) \\ \leq\ & Tr(W_{t-1}^T M W_{t-1}) + \gamma \sum_i \left\|w_{t-1}^i\right\|_2 \\ & \qquad -\gamma(\sum_i \left\|w_{t-1}^i\right\|_2 - \sum_i \frac{\left\|w_{t-1}^i\right\|_2^2}{2\left\|w_{t-1}^i\right\|_2}) \end{aligned}$$

Meanwhile, according to Lemma 2, $\sum_i \left\|w_t^i\right\|_2 - \sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_{t-1}^i\right\|_2} \leq \sum_i \left\|w_{t-1}^i\right\|_2 - \sum_i \frac{\left\|w_{t-1}^i\right\|_2^2}{2\left\|w_{t-1}^i\right\|_2}$. Therefore, we have the following inequality:

$$\begin{aligned} & Tr(W_{t+1}^T A W_{t+1}) + \gamma \sum_i \left\|w_{t+1}^i\right\|_2 \\ & \qquad \leq Tr(W_t^T A W_t) + \gamma \sum_i \left\|w_t^i\right\|_2, \end{aligned}$$

which indicates that the objective function value of $\min_{W^T W = I} Tr(W^T M W) + \gamma \sum_{i=1}^d \left\|w^i\right\|_2$ monotonically decreases using the updating rule in Algorithm 1. □

According to Theorem 1, we can see that the iterative approach in Algorithm 1 converges to the optimal W corresponding to (13). Because k is much smaller than n, the time complexity of computing M defined in (12) is about $O(n^2)$. To optimize the objective function of UDFS, the most time consuming operation is to perform eigen-decomposition of P_t. Note that $P_t \in \mathbb{R}^{d \times d}$. The time complexity of this operation is $O(d^3)$ approximately.

Experiments

In this section, we test the performance of UDFS proposed in this paper. Following [He *et al.*, 2005; Cai *et al.*, 2010], we test the performance of the proposed algorithm in terms of clustering.

Experiment Setup

In our experiment, we have collected a diversity of 6 public datasets to compare the performance of different unsupervised feature selection algorithms. These datasets include three face image datasets, i.e., UMIST[4], FERET[5] and YALEB [Georghiades *et al.*, 2001], one gait image dataset, i.e., USF HumanID [Sarkar *et al.*, 2005], one spoken letter recognition data, i.e., Isolet1[6] and one hand written digit image dataset, i.e., USPS [Hull, 1994]. Detailed information of the six datasets is summarized in Table 1.

We compare UDFS proposed in this paper with the following unsupervised feature selection algorithms.

[4]http://images.ee.umist.ac.uk/danny/database.html

[5]http://www.frvt.org/FERET/default.htm

[6]http://www.ics.uci.edu/ mlearn/MLSummary.html

Table 1: Database Description.

Dataset	Size	# of Features	# of Classes
UMIST	575	644	20
FERET	1400	1296	200
YALEB	2414	1024	38
USF HumanID	5795	2816	122
Isolet	1560	617	26
USPS	9298	256	10

- All Features which adopts all the features for clustering. It is used as the baseline method in this paper.
- Max Variance which selects the features corresponding to the maximum variances.
- Laplacian Score [He *et al.*, 2005] which selects the features most consistent with the Gaussian Laplacian matrix.
- Feature Ranking [Zhao and Liu, 2007] which selects features using spectral regression.
- Multi-Cluster Feature Selection (MCFS) [Cai *et al.*, 2010] which selects features using spectral regression with ℓ_1-norm regularization.

For LS, MCFS and UDFS, we fix k, which specifies the size of neighborhood, at 5 for all the datasets. For LS and FR, we need to tune the bandwidth parameter for Gaussian kernel, and for MCFS and UDFS we need to tune the regularization parameter. To fairly compare different unsupervised feature selection algorithms, we tune these parameters from $\{10^{-9}, 10^{-6}, 10^{-3}, 1, 10^3, 10^6, 10^9\}$. We set the number of selected features as $\{50, 100, 150, 200, 250, 300\}$ for the first five datasets. Because the total feature number of USPS is 256, we set the number of selected features as $\{50, 80, 110, 140, 170, 200\}$ for this dataset. We report the best results of all the algorithms using different parameters. In our experiment, each feature selection algorithm is first performed to select features. Then K-means clustering algorithm is performed based on the selected features. Because the results of K-means clustering depend on initialization, it is repeated 20 times with random initializations. We report the average results with standard deviation (std).

Two evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI), are used as evaluation metrics in this paper. Denote q_i as the clustering results and p_i as the ground truth label of x_i. ACC is defined as follows.

$$ACC = \frac{\sum_{i=1}^n \delta(p_i, map(q_i))}{n} \tag{18}$$

where $\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$ otherwise, and $map(q_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the Kuhn-Munkres algorithm. A larger ACC indicates better performance. Given two variables *P* and *Q*, NMI is defined in (19).

$$NMI(P, Q) = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}}, \tag{19}$$

where $I(P, Q)$ is the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q [Strehl and

Table 2: Clustering Results (ACC% $\pm$ std) of Different Feature Selection Algorithms

	All Features	Max Variance	Laplacian Score	Feature Ranking	MCFS	UDFS
UMIST	41.9 ± 3.0	46.2 ± 2.3	46.3 ± 3.3	48.1 ± 3.7	46.5 ± 3.5	$\mathbf{49.2 \pm 3.8}$
FERET	22.0 ± 0.5	20.1 ± 0.3	22.4 ± 0.5	22.8 ± 0.5	25.1 ± 0.7	$\mathbf{26.1 \pm 0.6}$
YALEB	10.0 ± 0.6	9.6 ± 0.3	11.4 ± 0.6	13.3 ± 0.8	12.4 ± 1.0	$\mathbf{14.7 \pm 0.6}$
USF HumanID	23.1 ± 0.6	20.9 ± 0.5	18.8 ± 0.3	10.1 ± 0.1	23.2 ± 0.6	$\mathbf{24.6 \pm 0.8}$
Isolet	57.8 ± 4.0	56.6 ± 2.6	56.9 ± 2.9	57.1 ± 2.9	61.1 ± 4.4	$\mathbf{66.0 \pm 3.6}$
USPS	62.9 ± 4.3	63.4 ± 3.1	63.5 ± 3.2	63.6 ± 3.1	65.3 ± 5.4	$\mathbf{65.8 \pm 3.3}$

Table 3: Clustering Results (NMI% $\pm$ std) of Different Feature Selection Algorithms

	All Features	Max Variance	Laplacian Score	Feature Ranking	MCFS	UDFS
UMIST	62.9 ± 2.4	63.6 ± 1.8	65.1 ± 2.0	64.9 ± 2.6	65.9 ± 2.3	$\mathbf{66.3 \pm 2.0}$
FERET	62.7 ± 0.4	62.3 ± 0.4	63.2 ± 0.3	63.3 ± 0.5	64.8 ± 0.5	$\mathbf{65.6 \pm 0.4}$
YALEB	14.2 ± 0.7	13.1 ± 0.4	18.4 ± 1.0	21.3 ± 0.9	18.8 ± 1.1	$\mathbf{25.4 \pm 0.9}$
USF HumanID	50.9 ± 0.4	49.1 ± 0.4	47.5 ± 0.2	29.3 ± 0.3	50.6 ± 0.4	$\mathbf{51.6 \pm 0.5}$
Isolet	74.2 ± 1.8	73.2 ± 1.1	72.0 ± 1.1	72.5 ± 1.7	75.5 ± 1.8	$\mathbf{78.1 \pm 1.3}$
USPS	59.2 ± 1.5	59.6 ± 1.1	60.2 ± 1.3	59.6 ± 1.1	61.2 ± 1.7	$\mathbf{61.6 \pm 1.5}$

Ghosh, 2002]. Denote t_l as the number of data in the cluster $\mathcal{C}_l$ $(1 \leq l \leq c)$ according to clustering results and $\tilde{t}_h$ be the number of data in the h-th ground truth class $(1 \leq h \leq c)$. NMI is defined as follows [Strehl and Ghosh, 2002]:

$$NMI = \frac{\sum_{l=1}^{c}\sum_{h=1}^{c} t_{l,h} \log(\frac{n \cdot t_{l,h}}{t_l \tilde{t}_h})}{\sqrt{\left(\sum_{l=1}^{c} t_l \log \frac{t_l}{n}\right)\left(\sum_{h=1}^{c} \tilde{t}_h \log \frac{\tilde{t}_h}{n}\right)}}, \qquad (20)$$

where $t_{l,h}$ is the number of samples that are in the intersection between the cluster $\mathcal{C}_l$ and the h-th ground truth class. Again, a larger NMI indicates a better clustering result.

Experimental Results and Discussion

First, we compare the performance of different feature selection algorithms. The experiment results are shown in Table 2 and Table 3. We can see from the two tables that the clustering results of All Features are better than those of Max Variance. However, because the feature number is significantly reduced by performing Max Variance for feature selection, resulting in the subsequential operation, e.g., clustering, faster. Therefore, it is more efficient. The results from other feature selection algorithms are generally better than All Features and also more efficient. Except for Max Variance, all of the other feature selection algorithms are non-linear approaches. We conclude that local structure is crucial for feature selection in many applications, which is consistent with previous work on feature selection [He *et al.*, 2005]. We can also see from the two tables that MCFS gains the second best performance. Both Feature Ranking [Zhao and Liu, 2007] and MCFS [Cai *et al.*, 2010] adopt a two-step approach, i.e, spectral regression, for feature selection. The difference is that Feature Ranking analyzes features separately and selects features one after another but MCFS selects features in batch-mode. This observation validates that it is a better way to analyze data features jointly for feature selection. Finally, we observe that the UDFS algorithm proposed in this paper obtains the best performance. There are two main reasons for this. First, UDFS analyzes features jointly. Second, UDFS simultaneously utilizes discriminative information and local structure of data distribution.

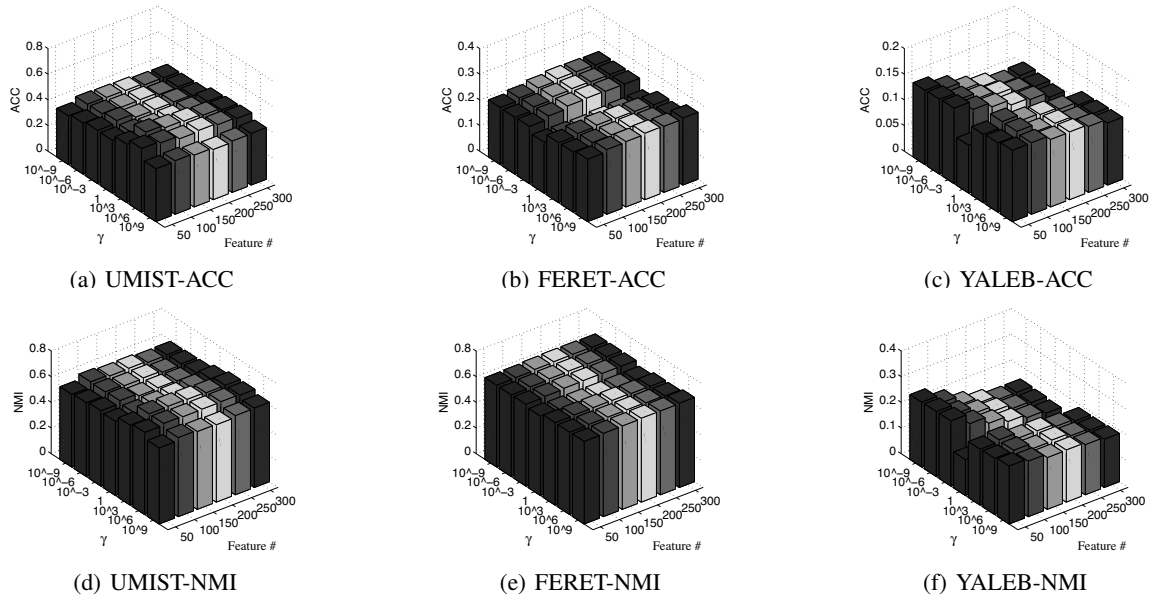
Next, we study the performance variation of UDFS with respect to the regularization parameter γ in (11) and the number of selected features. Due to the space limit, we use the three face image datasets as examples. The experimental results are shown in Fig.1. We can see from Fig.1 that the performance is not very sensitive to γ as long as it is smaller than 1. However, the performance is comparatively sensitive to the number of selected features. How to decide the number of selected features is data dependent and still an open problem.

Conclusion

While it has been shown in many previous works that discriminative information is beneficial to many applications, it is not that straightforward to utilize it in unsupervised learning due to the lack of label information. In this paper, we have proposed a new unsupervised feature selection algorithm which is able to select discriminative features in batch mode. An efficient algorithm is proposed to optimize the $\ell_{2,1}$-norm regularized minimization problem with orthogonal constraint. Different from existing algorithms which select the features which best preserve data structure of the whole feature set, UDFS proposed in this paper is able to select discriminative feature for unsupervised learning. We show that it is a better way to select discriminative features for data representation and UDFS outperforms the existing unsupervised feature selection algorithms.

Acknowledgement

This work is supported by ARC DP1094678 and partially supported by the FP7-IP GLOCAL european project.



(a) UMIST-ACC (b) FERET-ACC (c) YALEB-ACC

(d) UMIST-NMI (e) FERET-NMI (f) YALEB-NMI

Figure 1: Performance variation of UDFS *w.r.t* different parameters.

References

[Argyriou *et al.*, 2008] Andreas Argyriou, Theodoros Evgeniou, Massimiliano Pontil, Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. In *Machine Learning*, 2008.

[Cai *et al.*, 2010] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. *KDD*, 2010.

[Duda *et al.*, 2001] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (2nd Edition)*. John Wiley & Sons, New York, USA, 2001.

[Fukunaga, 1990] K. Fukunaga. *Introduction to statistical pattern recognition (2nd Edition)*. Academic Press Professional, Inc, San Diego, USA, 1990.

[Georghiades *et al.*, 2001] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, page 23(6):643660, 2001.

[He *et al.*, 2005] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *NIPS*, 2005.

[Hull, 1994] J.J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 16(5):550–554, 1994.

[Liu *et al.*, 2009] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *UAI*, 2009.

[Nie *et al.*, 2008] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, 2008.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, 2010.

[Obozinski *et al.*, 2008] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Highdimensional union support recovery in multivariate regression. In *NIPS*, 2008.

[Sarkar *et al.*, 2005] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer. The humanid gait challenge problem: data sets, performance, and analysis. *IEEE TPAMI*, pages 27(2):162–177, 2005.

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

[Sugiyama, 2006] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, 2006.

[Yang *et al.*, 2010a] Yi Yang, Feiping Nie, Shiming Xiang, Yueting Zhuang, and Wenhua Wang. Local and global regressive mapping for manifold learning with out-of-sample extrapolation. In *AAAI*, 2010.

[Yang *et al.*, 2010b] Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang. Image clustering using local discriminant models and global integration. *IEEE TIP*, pages 19(10)2761 – 2773, 2010.

[Yang *et al.*, 2011] Yang Yang, Yi Yang, Zi Huang, Heng Tao Shen, and Feiping Nie. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, pages 881–888, 2011.

[Zhao and Liu, 2007] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, 2007.

[Zhao *et al.*, 2010] Z. Zhao, L. Wang, and H. Liu. Efficient spectral feature selection with minimum redundancy. In *AAAI*, 2010.