Unsupervised Discriminative Feature Selection in a Kernel Space via $L_{2,1}$-Norm Minimization

Yang Liu
Nat'l Engineering Lab for Video Technology
Key Lab. of Machine Perception (MoE)
Sch'l of EECS
Peking University, Beijing 100871, China
liuyang.cs@pku.edu.cn

Yizhou Wang
Nat'l Engineering Lab for Video Technology
Key Lab. of Machine Perception (MoE)
Sch'l of EECS
Peking University, Beijing 100871, China
Yizhou.Wang@pku.edu.cn

Abstract

Traditional nonlinear feature selection methods map the data from an original space into a kernel space to make the data be separated more easily, then move back to the original feature space to select features. However, the performance of clustering or classification is better in the kernel space, so we are able to select the features directly in the kernel space and get the direct importance of each feature. Motivated by this idea, we propose a novel method for unsupervised feature selection directly in the kernel space. To do this, we utilize local discriminative information to find the best label for each instance with $L_{2,1}$-norm minimization, then select the most important features in the kernel space using the labels predicted. Extensive experiments demonstrate the effectiveness of our method.

1. Introduction

In recent years, feature selection has received an increased interest in the machine learning community. Given huge of features, it will take a long time for clustering and classification. However, some of the features can even be noise which will hurt the performance. With feature selection, we can remove irrelevant and redundant features for data analysis.

Based on whether the label information is available or not, feature selection problems can be classified into unsupervised feature selection and supervised feature selection. Supervised feature selection algorithms usually evaluate the importance of each feature according to the label information [6]. With label information, they can select discriminative features. However, in practice, unlabeled data is massive thus labeling them is expensive. Thus unsupervised feature selection algorithms are paid more and more attention for its unnecessary cost of labeling data manually. In unsupervised learning algorithm, selecting the features to preserve the data similarity is a common criterion [2]. However, discriminative information is always neglected though it has been demonstrated important in data analysis [4]. Yang et al. [9] tried to use discriminative information in unsupervised feature selection and it works well.

As a linear feature selection algorithm, [9] is conducted in the original input space, but can not work well on the nonlinear data. So we want to use kernel methods to improve it. Kernel methods [1][5] map the data from an original space into a kernel space to make the data be separated more easily, then move back to the original feature space to select features. As the performance of clustering or classification is better in the kernel space, we are able to select the features directly in the kernel space and get the direct importance of each feature. FSGP[3] tries this approach and works well, but it can only deal with the supervised learning problem.

In this paper, we propose a novel method for unsupervised feature selection *directly* in the kernel space. To do this, we utilize local discriminative information to find the best label for each instance with $L_{2,1}$-norm minimization, then select the most important features directly in the kernel space using the labels. Finally, we cluster the data in the kernel space with the selected feature set to evaluate our approach. We conduct extensive experiments over several datasets to prove the effectiveness of the proposed algorithm.

The paper is organized as follows. In Section 2, we briefly review the unsupervised discriminative method. In Section 3, we introduce our newly proposed kernel method. The experimental results are provided in Section 4, and a brief conclusion is presented in Section 5.

2. Unsupervised Discriminative Feature Selection

The unsupervised discriminative method UDFS [9] incorporates discriminative analysis and $L_{2,1}$-norm minimization into a joint framework for unsupervised feature selection. It handles the feature-dependency problem successfully. In the following, we briefly review UDFS, then present our algorithm to extend it in a kernel space in Section 3.

Denote $X = \{x_1, x_2, ..., x_n\}$ as the training set, where $x_i \in \mathbf{R}^d (1 \leq i \leq n)$ is the i-th datum and n is the total number of training data. $Y = [y_1, y_2, ..., y_n]^T \in \{0,1\}^{n \times c}$ is the label matrix. The total scatter matrix S_t and between class scatter matrix S_b are defined as follows [4].

$$\begin{aligned} S_t &= \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = \widetilde{X}\widetilde{X}^T \\ S_b &= \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \widetilde{X}GG^T\widetilde{X}^T \end{aligned}$$

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. $H_m = I - \frac{1}{m}1_m 1_m^T \in R^{m \times m}$ and $\widetilde{X} = XH_n$ is the data matrix after being centered. Denote $G = [G_1, ..., G_n]^T = Y(Y^TY)^{-1/2}$ as the scaled label matrix.

In the unsupervised problem, there is no label information. Hence, UDFS assumes that there is a linear classifier $W \in R^{d \times c}$ which classifies each data point into a class. So define the scaled label matrix as

$$G_i = W^T x_i \tag{1}$$

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. $L_{2,1}$-norm $\|W\|_{2,1}$ minimization [6] can achieve this.

Define the local discriminative score $DS = Tr[(S_t + \lambda I)^{-1} S_b]$. Clearly, a larger DS indicates a higher discriminative ability of W. UDFS intends to train a W corresponding to the highest discriminative scores for all the training data $x_1, ..., x_n$. Therefore the objective function is $\min_{W^TW=I}\{Tr(G^T H_n G) - DS + \gamma\|W\|_{2,1}\}$ where $\|W\|_{2,1} = \sum_{i=1}^{r}\sqrt{\sum_{j=1}^{p} w_{ij}^2}$.

Taking local discriminative information into account, we denote $X_i = [x_i, x_{i_1}, ..., x_{i_k}]$ for each data point x_i as the local data matrix. $S_i \in \{0,1\}^{n \times (k+1)}$ is the selection matrix to choose the k nearest points.

Finally, the objective function is shown as

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

where

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

Denote w^i as the i-th row of W. UDFS ranks each feature $f_i|_{i=1}^{d}$ according to $\|w^i\|_2$ in a descending order and selects top ranked features. More details can be found in [9].

3. Unsupervised Discriminative Feature Selection in a Kernel Space

3.1. Objective Function

Now we discuss how to extend UDFS into a kernel space. We want to map the data from an original space into a kernel space as $\varphi : x \longrightarrow F$, making the data be separated more easily, then perform feature selection directly in the kernel space.

Denote $\varphi(X) = [\varphi(x_1), \varphi(x_2), ..., \varphi(x_N)]$, and assuming that the transformation matrix in (2)

$$W = [\varphi(x_1), \varphi(x_2), ..., \varphi(x_N)]\widetilde{W} = \varphi(X)\widetilde{W} \tag{3}$$

As we know, $k(x_i, x_j) = \varphi(x_i)^T \cdot \varphi(x_j)$ is the inner product of data pairs, and $K = \varphi(X)^T\varphi(X)$ is the kernel Gram matrix with $K_{ij} = \varphi(x_i)^T \cdot \varphi(x_j)$. For (2), we have:

$$W^TMW = (\varphi(X)\widetilde{W})^T M(\varphi(X)\widetilde{W}) = \widetilde{W}^T\widetilde{M}\widetilde{W}$$

where $\widetilde{M} = \varphi(X)^T M \varphi(X)$.

Based on kernel matrix K, we can calculate the distance $d(i,j) = K(i,i) + K(j,j) - 2K(i,j)$ in the kernel space. Based on distance d, we can find the k-nearest instances and get $\varphi(\widetilde{X}_i)$. Then we denote $\widetilde{K}_i = \varphi(\widetilde{X}_i)^T\varphi(\widetilde{X}_i)$. So the objective function (2) can be rewritten as

$$\min_{\widetilde{W}^T\widetilde{W}=I} Tr(\widetilde{W}^T\widetilde{M}\widetilde{W}) + \gamma\|\widetilde{W}\|_{2,1} \tag{4}$$

where

$$\widetilde{M} = K[\sum_{i=1}^{n}(S_i H_{k+1}(\widetilde{K}_i + \lambda I)^{-1} H_{k+1} S_i^T)]K$$

3.2. Feature Selection

We can select features directly in the kernel space using $\widetilde{W}$. Now we get the optimal $\widetilde{W}$, and

$$G_i = W^T\varphi(x_i) = [\varphi(X)\widetilde{W}]^T\varphi(x_i) = \widetilde{W}^T\psi(x_i)$$

where $\psi(x_i) = \varphi(X)^T \varphi(x_i)$.

Based on ψ, we map the instances into a kernel space, and select features directly in the kernel space. Actually, the ith feature in ψ kernel space is the similarity between the ith instance and each of all the instances. Then we sort each feature according to the ith row of $\widetilde{W}$ in a descending order and select the top ranked ones in the ψ kernel space.

3.3. Optimization

In this section, we use the approach [6] to solve the optimization problem shown in (4). We describe the details in Algorithm 1 as follows.

```
Algorithm 1:
for $i = 1$ to $n$ do
    $B_i = (\widetilde{K}_i + \lambda I)^{-1}$
    $M_i = S_i H_{k+1} B_i H_{k+1} S_i^T$;
end
$M = K(\Sigma_{i=1}^n M_i)K$;
Set $t = 0$;
Initialized $D_0 \in R^{d \times d}$ as an identity matrix;
repeat
    $P_t = M + \gamma D_t$;
    $\widetilde{W}_t = [p_1, ..., p_c]$ where $p_1, ..., p_c$ are the
    eigenvectors of $P_t$ corresponding to the first $c$
    smallest eigenvalues;
    $D_{t+1} = \begin{pmatrix} \frac{1}{2\|\widetilde{w}_t^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|\widetilde{w}_t^d\|_2} \end{pmatrix}$;
    $t = t + 1$;
until
Sort each feature $\widetilde{f}_i|_{i=1}^d$ according to $\|\widetilde{w}_t^i\|_2$ in
descending order and select the top ranked ones in
the kernel space.
```

4. Experiments

To evaluate the proposed method, we compare it with the linear approach (UDFS[9]), and other nonlinear algorithms (Kernel PCA[8], Kernel Kmeans[7]), on 3 real-world datasets downloaded from DAT Repository[1] (the data are given in 100 predefined splits), namely, $thyroid$, $german$, $diabetis$, and 3 datasets downloaded from UCI Machine Learning Repository[2], namely, Wisconsin Diagnostic Breast Cancer ($wdbc$), Johns Hopkins University Ionosphere database (ion), SPECT heart data ($spect$). All the datasets have two classes.

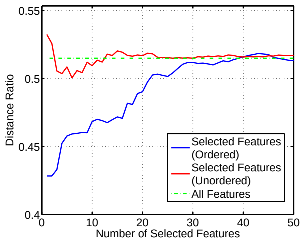

Figure 1. Distance Ratio

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. It is repeated 10 times with random initializations, and we report the average results. For calculating accuracy (ACC) of clustering, we choose the best mapping that permutes clustering labels to match the ground truth labels using the Kuhn-Munkres algorithm. A larger ACC indicates better performance.

4.1. Distance Ratio

We first use *distance ratio* [3] to show that our method uses much discriminative information. The smaller the *distance ratio* is, the easier it is for the data to be separated.

The experiment is done on the dataset ion. We apply our method to RBF kernel, with $\sigma = 1$. Figure 1 shows the relationship between the number of features and the distance ratio. The red line shows the distance ratio against the number of selected features in the dataset. After using our method, we sort the features according to the predict importance in descending order, and the blue line shows the ratio against the number of selected ordered features. Taking all the features in the kernel space into account, we draw the dashed line.

From blue line, we can see that the distance ratio is very small against the first few features. It approves that the first few ordered features contain more discriminative information than the unordered ones.

4.2. Comparison with Linear Methods

In this experiment, we compare our method with the linear approach UDFS[9] to show that it can work better on nonlinear datasets.

For UDFS and our method, we fix k, which specifies the size of neighborhood, at 5, and $\gamma = 10^{-3}$, $\lambda = 10^3$ (the values are default in the code download-

[1] http://ida.first.fhg.de/projects/bench/benchmarks.htm
[2] http://archive.ics.uci.edu/ml/

ed from Yang[9]'s website) for all the datasets. We apply our method to RBF kernel $K(x_i, x_j) = e^{\frac{-\|x_i - x_j\|^2}{2\sigma}}$, with $\sigma = const * D$, where D is the maximum distance between samples and $const$ varies in the range of [0.1, 2.0].

Here we only set the number of selected features as $\{1, 2, ..., d\}$ for all the methods, and d is the dimension of original space. Actually, the dimension of kernel space is far more than d. Hence better accuracy in the result proves the effectiveness of our method. We report the best results of all the algorithms using different parameters. Table 1 shows the accuracy for each dataset. We can see that our method achieves better performance than the linear approach UDFS.

Table 1. Accuracy of Clustering

Dataset	All Features	UDFS	Ours
thyroid	78.7	78.5	**92.8**
german	53.6	67.1	**69.6**
diabetis	68.3	68.6	**69.6**
wdbc	85.4	88.7	**90.8**
ion	71.2	71.2	**76.6**
spect	56.1	60.6	**84.2**

4.3. Comparison with Nonlinear Methods

In this experiment, we compare our method with other nonlinear approaches (Kernel PCA[8] and Kernel Kmeans[7]) to show that it can work better.

We consider all the features for Kernel-Kmeans and select features using Kernel PCA and our method. To fairly compare different nonlinear unsupervised algorithms, we use the RBF kernel with $\sigma = 1$ for all the methods, and we only tune the parameters γ and λ from $\{10^{-3}, 1, 10^3\}$ for our method.

Figure 2 shows the accuracy for each dataset. The X-axis is the number of selected features, and the Y-axis is the predictive accuracy of clutering. From these experimental results, we observe that using the first few features can indeed improve the clustering accuracy and the proposed method works better than other nonlinear approaches.

5. Conclusion

In this paper, we propose a novel method for unsupervised feature selection directly in the kernel space. It works well based on the importance of each feature in the kernel space. The experimental results demonstrated that the proposed algorithm performs well in both selecting relevant features and removing redundancy.

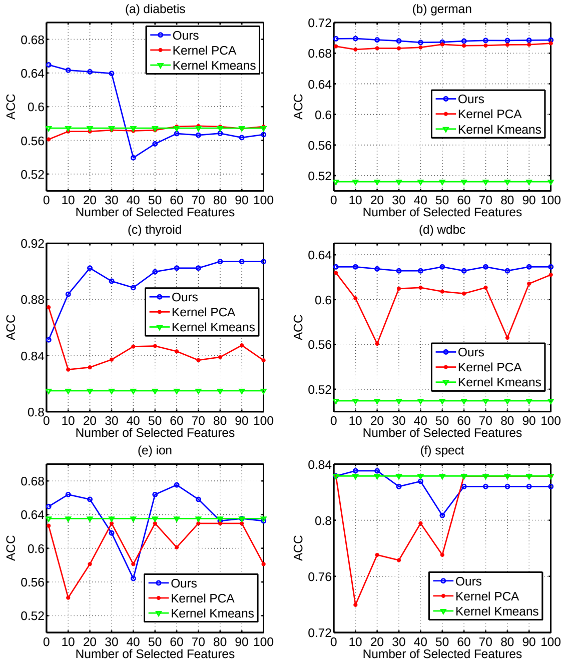

Figure 2. Accuracy of Clustering

Acknowledgment

We'd like to thank for the support from research grants National Basic Research Program of China (973 Program) 2009CB320904 and NSFC-90920012.

References

[1] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. *ICML*, 1998.

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

[3] B. Cao, D. Shen, J. T. Sun, Q. Yang, and Z. Chen. Feature selection in a kernel space. *ICML*, 2007.

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

[5] Z. Liang and T. Zhao. Feature selection for linear support vector machines. *ICPR*, 2006.

[6] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. *NIPS*, 2010.

[7] B. Schlkopf and A. Smola. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, July 1998.

[8] B. Schlkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

[9] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou. L2,1-norm regularized discriminative feature selection for unsupervised learning. *IJCAI*, 2011.