Manifold-Regularized Selectable Factor Extraction for Semi-supervised Image Classification

Xin Shi[1]
shixinnk@gmail.com
Chao Zhang[1]
chzhang@cis.pku.edu.cn
Fangyun Wei[1]
weifangyun@pku.edu.cn
Hongyang Zhang[1]
hy_zh@pku.edu.cn
Yiyuan She[2]
yshe@stat.fsu.edu

[1] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, PR China
[2] Department of Statistics, Florida State University, USA

Abstract

In many vision analytics-based applications such as image classification, we confront explosive growth of high-dimensional data. Thus, many feature selection and extraction methods have been proposed to reduce the computational cost and avoid over-fitting. Recently, a novel selectable factor extraction (SFE) framework is proposed to simultaneously perform feature selection and extraction, and is theoretically and practically proved to be effective in handling high-dimensional data. The algorithm is also quite efficient and easy to implement. Although it is advantageous in several aspects, SFE is only designed for either supervised or unsupervised learning, and is not suitable when there are limited labeled samples and a large number of unlabeled samples, since the data distribution knowledge is likely to be poorly exploited. To tackle this problem, we propose a novel manifold regularized SFE (MRSFE) framework for semi-supervised image classification. In MRSFE, the local structures of the whole dataset are preserved, and the data distribution is well exploited. By integrating the label information, low rank property of the features and data distribution knowledge, the proposed MRSFE could select and extract reliable discriminative features when the labeled samples are scarce. An efficient and easy-to-implement algorithm is designed to find the solutions. Extensive experimental results on a real-world image dataset demonstrate the superiority of our method.

1 Introduction

In modern computer vision applications, we frequently confront high-dimensional data [14]. For example, in face recognition, the image pixels are usually directly utilized as the features. In natural image classification [12, 13], the local features are often clustered into a long histogram. The feature dimensions of these applications can be up to several thousands and

Pages 132.1-132.11
DOI: https://dx.doi.org/10.5244/C.29.132

even more, which often leads to the "curse of dimensionality" problem. To deal with the high dimensional problem effectively, the feature selection methods select a subset of features from the original ones to reduce the computational cost and the chance of over-fitting.

According to the number of labeled data in training set, existing feature selection algorithms fall into three groups: supervised, unsupervised and semi-supervised. Supervised feature selection (*e.g.*, [15, 16]) always requires sufficient labeled training data. But labeled training data are very expensive and time-consuming to obtain in real-world applications. Unsupervised feature selection methods (*e.g.*, [8, 19, 21]) only use the structure of graph to select features, it is not so reliable since the label information is ignored ([20]). Another large part of feature selection methods is semi-supervised feature selection, it uses both labeled and a large amount of unlabeled data to select discriminative features. For example, [22] proposed a semi-supervised feature selection method which uses labeled data to maximize the margin between different classes and uses unlabeled data to discover the geometrical structure of the data space. However, they use Laplacian criteria to help feature extraction and it may not be the best because we care more about the predict accuracy not the Laplacian score of the features. [2] proposed a convex formulation for semi-supervised multi-label feature selection based on the sparse penalty and a single convex framework. Due to the lack of graph construction, the computational cost is relatively low. However, it only focuses on the feature selection task, some important properties, such as the low rank property of original data, may not be well explored.

Recently, a selectable factor extraction (SFE)[18] framework is proposed to perform feature selection and extraction simultaneously. In this framework, both the sparsity and low rankness of the features are well explored. The framework is formulated for both supervised and unsupervised learning, and the sharp oracle inequalities is proved for various convex or nonconvex penalties. The authors have also demonstrated that SFE tends to obtain lower test error compared with rank reduction or variable selection alone empirically. Besides, the designed algorithms are quite efficient with easy implementation, and are scalable for big data computing.

In spite of these advantages, SFE is not appropriate for the scenario when the labeled samples are scarce but a large number of unlabeled samples are available. This is because a reliable solution should respect the underlying data distribution, which might be very different from the distribution of the limited labeled samples [1]. Although we can apply SFE on only the feature matrix of the large amount of unlabeled samples, the label information is ignored and thus the selected and extracted features might not be discriminative enough for classification. In addition, the designed unsupervised SEL-PCA (selectable principal component analysis)[18] only utilize the global information of all samples, and disregard the local structure, which is however, critical in image classification. This is because images are usually represented in a highly nonlinear feature space, and it is nontrivial to reveal the underlying data distribution of the images in the feature space.

Therefore, we propose the manifold-regularized selectable factor extraction (MRSFE) for semi-supervised image classification. In particular, we use a low rank penalized regression model to explore the label information. A low rank matrix of the regression coefficients, together with the $\ell_{2,1}$ or $\ell_{2,0}$ norm penalty is learned for joint feature selection and extraction. In addition, all the labeled and unlabeled samples are utilized in MRSFE to construct the data adjacency graph to approximate the underlying data manifold, which the data distribution is assumed to be supported on. The graph Laplacian is then incorporated as a regularization term to smooth the coefficients matrix. In this way, the local structures of the whole dataset are preserved, and the data distribution is well exploited. By integrating all the label infor-

mation, low rankness of the original features, as well as the data distribution knowledge, MRSFE could select or extract reliable discriminative features when the labeled samples are scarce. We design a fast and easy-to-implement algorithm to solve our optimization problem with convex or nonconvex penalties. The computational cost is very low since it just involved some small-scale SVD decompositions and thresholding operations. To evaluate the performance of our algorithm, we apply it on a challenge web image dataset, NUS-WIDE-OBJECT ([3]), and compare it with two competitive feature selection methods [2, 22]. The experimental results demonstrate that our method outperforms other compared methods in terms of both prediction accuracy and computational cost.

2 Manifold-Regularized Selectable Factor Extraction

To derive our model, we begin with the reduced rank regression (RRR) model[7]. Specifically, RRR fits sample labels by the linear combination of the design matrix. To achieve dimension reduction, RRR requires the representation matrix to be low-rank. The low rankness of the representation matrix guarantees that the data after regression lie in a low-dimension subspace. In particular, RRR formulates as

$$\min_B \|Y_L - X_L B\|_F^2, \quad \text{s.t. rank}(B) \le r, \tag{1}$$

where $B = [b_1, b_2, \cdots, b_d]^T \in \mathbb{R}^{d\times c}$ is the representation matrix, $X_L \in \mathbb{R}^{l\times d}$ is the design matrix of the labeled samples, l is the number of the labeled samples and d is the number of the features; $Y_L \in \mathbb{R}^{l\times c}$ is the response matrix where c is the number of the classes, i.e., $Y_{ij} = 1$ if and only if the ith sample belongs to the jth class. Furthermore, RRR has a closed form solution $\hat{B} = (X^TX)^{-1}X^TYV_rV_r^T$, where V_r consists of top r eigenvectors of $Y^TX(X^TX)^{-1}X^TY$. It is well known that RRR is effective for multivariate models[7], and has been widely applied in various applications, such as computer vision[6], machine learning[5] and economy[4].

Although successful in practice, the plain RRR has however certain drawbacks — the model (1) typically involves all input features of X, because the representation matrix B is dense, and we can not remove the junk features and discern the important ones, which loses interpretability in high-dimension applications. We expect to use few variables to interpret the entire model in high-dimension applications. To this end, we implement feature selection for RRR, i.e., selecting few features from high-dimension samples which keeps most of information available. So we use the $\ell_{2,1}$ regularization to impose the row sparsity on B. Namely, we have

$$\min_B \|Y_L - X_L B\|_F^2 + \alpha\|B\|_{2,1}, \quad \text{s.t. rank}(B) \le r. \tag{2}$$

where $\|B\|_{2,1} = \sum_{i=1}^d \sqrt{\sum_{j=1}^c B_{ij}^2}$, which promotes row sparsity of B, that is to say, some rows of B will be zero. Note that we select features from X_L according to the non-zero rows of the representation matrix B, i.e., if the i-th row of B is non-zero, we conclude that the ith feature of X_L(i-th column) is significant. We call (2) the selectable factor extraction (SFE) method.

In many applications where we have only a small number of labeled samples, the representation matrix B learned from model (2) is often unreliable due to the limited information the data provides. Fortunately, given that in most situations high dimensional data usually lie near low-dimension manifold, we propose to use a large number of unlabeled samples

to capture data structure (a.k.a. semi-supervised learning)[10, 11]. In response to this, we borrow the manifold regularization (MR) to help learn the manifold structure. Specifically, MR formulates as follows:

$$\begin{aligned}\frac{1}{2}\sum_{p=1}^{c}\sum_{i,j=1}^{n}(\hat{y}_{ip}-\hat{y}_{jp})^2A_{ij} &= \frac{1}{2}\sum_{i,j=1}^{n}A_{ij}(\hat{\boldsymbol{y}}_i^T\hat{\boldsymbol{y}}_i+\hat{\boldsymbol{y}}_j^T\hat{\boldsymbol{y}}_j-2\hat{\boldsymbol{y}}_i^T\hat{\boldsymbol{y}}_j)\\ &=\mathrm{tr}(\hat{Y}^T(D-A)\hat{Y})\\ &=\mathrm{tr}(\hat{Y}^TL\hat{Y})\\ &=\mathrm{tr}(B^TX_{LU}^TLX_{LU}B),\end{aligned} \tag{3}$$

where $n=l+u$ is the number of labeled and unlabeled samples, X_{LU} is the data matrix containing both the labeled and unlabeled samples together, $\hat{Y}=[\hat{\boldsymbol{y}}_1,\hat{\boldsymbol{y}}_2,\ldots,\hat{\boldsymbol{y}}_n]^T\in\mathbb{R}^{n\times c}$ is the prediction over all data, and $\hat{y}_{ip}$ is the prediction of p-th calss of i-th sample, A is the affinity matrix whose elements indicate the similarity between samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and D is a diagonal matrix such that $D_{ii}=\sum_{j=1}^{n}A_{ij}$. The matrix $L=D-A$ is also termed Laplacian matrix. MR makes the prediction look similar provided that the two samples are close in the feature space. So the geometric structure of the data samples is preserved well. As for the construction of the affinity matrix A, we use the k nearest neighbors for the elements:

$$A_{ij}=\begin{cases}1, & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are labeled data, and } y_i=y_j;\\ \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right), & \text{if } \boldsymbol{x}_i\in\mathcal{N}_k(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j\in\mathcal{N}_k(\boldsymbol{x}_i);\\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

where $\boldsymbol{x}_i\in\mathcal{N}_k(\boldsymbol{x}_j)$ indicates that $\boldsymbol{x}_i$ belongs to the k nearest neighbors of $\boldsymbol{x}_j$.

Combining model (2) with (3), we have the following optimization problem:

$$\min_B \|Y_L-X_LB\|_F^2+\alpha\|B\|_{2,1}+\beta\mathrm{tr}(B^TX_{LU}^TLX_{LU}B),\quad \text{s.t. } \mathrm{rank}(B)\le r. \tag{5}$$

To deal with the row sparsity and the rank constraint on B, we write $B=SV^T$, where $S\in\mathbb{R}^{d\times r}$ and $V\in\mathbb{R}^{c\times r}$ is an orthogonal matrix. It's easy to see that $\mathrm{rank}(B)=\mathrm{rank}(SV^T)\le r$, thus the rank constraint on B is replaced by matrix factorization. And from the orthogonality of V we could get that $\|B\|_{2,1}=\|SV^T\|_{2,1}=\|S\|_{2,1}$, the row sparse penalty can be applied to S. Thus model (5) is equivalent to

$$\min_{S,V}\|Y_L-X_LSV^T\|_F^2+\alpha\|S\|_{2,1}+\beta\mathrm{tr}(VS^TX_{LU}^TLX_{LU}SV^T),\quad \text{s.t. } V^TV=I_r, \tag{6}$$

where $I_r\in\mathbb{R}^{r\times r}$ is the identity matrix.The full-rank factorization of B enables us to tackle the sparsity regularization and the low-rank constraint separately. In other words, S selects significant features from X_L, while the orthogonal matrix V determines the subspace after dimension reduction. So we can implement feature selection and extraction simultaneously.

Considering the ℓ_1 penalty cannot handle the collinearity and may lead to inconsistent and biased estimation([23]), similar to $\ell_{2,1}$, we advocate to use non-convex constraint such as $\ell_{2,0}$ instead of widely-used $\ell_{2,1}$ penalty for S in (6)

$$\min_{S,V^TV=I_r}\|Y_L-X_LSV^T\|_F^2+\beta\mathrm{tr}(VS^TX_{LU}^TLX_{LU}SV^T),\quad \text{s.t. } \|S\|_{2,0}\le q_s \tag{7}$$

where q_s is a parameter to control the number of selected features. Using the constraint form instead of the penalty is intuitive, because we can directly control the number of features

we need. Another advantage is that the constraint constant q_s is in a large range for the optimal solution. Although the $\ell_{2,0}$ penalty is nonconvex and hard to optimization in classical methods, it is doable in our algorithm 1, and the replacement of $\ell_{2,0}$ does not affect the optimal solution. Once S is obtained from (6) or (7), we can select the significant features according to top-k index of the row-norms in descending order or nonzero rows of S. We call both of the models (6) and (7) manifold-regularized semi-supervised selectable factor extraction method(MRSFE).

3 Optimization Algorithm

We focus on model (6) and use alternating optimization method to solve this problem. Briefly, our algorithm can be divided into two sub-processes: V-optimization step and S-optimization step.

3.1 V-optimization

When matrix S is fixed, problem (6) reduces to

$$\min_{V} \|Y_L - X_L S V^T\|_F^2, \quad \text{s.t. } V^T V = I. \tag{8}$$

Note that $V^T V = I_r$, thus the optimal solution of (8) is equivalent to

$$\begin{aligned} \hat{V} &= \arg\min_{V^T V = I_r} \|Y_L - X_L S V^T\|_F^2 \\ &= \arg\max_{V^T V = I_r} \operatorname{tr}(Y_L^T X_L S V^T) \\ &= \arg\min_{V^T V = I_r} \operatorname{tr}(V^T Y_L^T X_L S) \\ &= \arg\min_{V^T V = I_r} \|V - Y_L^T X_L S\|_F^2. \end{aligned} \tag{9}$$

Namely, we have the following optimization problem:

$$\min_{V} \|V - Y_L^T X_L S\|_F^2, \quad \text{s.t. } V^T V = I. \tag{10}$$

This is the Procrustes problem [17]. To solve the problem, we can set $W = Y_L^T X_L S$ and then perform SVD of $W = U_w \Sigma_w V_w^T$, the optimal V is given by $\hat{V} = U_w V_w$. Note that $W \in \mathbb{R}^{c \times r}$, where c is the class number and r is the rank which is significantly smaller than the sample size. So the computational cost of the SVD decomposition of W is low.

3.2 S-optimization

When matrix V is fixed, problem (6) reduces to

$$\min_{S} F(S) = \min_{S} \|Y_L V - X_L S\|_F^2 + \alpha \|S\|_{2,1} + \beta \operatorname{tr}(S^T X_{LU}^T L X_{LU} S). \tag{11}$$

We now denote formulation (11) as

$$\min_{S} F(S) = f(S) + g(S), \tag{12}$$

where $f(S) = \|Y_L V - X_L S\|_F^2 + \beta \mathrm{tr}(S^T X_{LU}^T L X_{LU} S)$ and $g(S) = \alpha \|S\|_{2,1}$. Then we can define the surrogate function at a given point S^{cur}:

$$G(S, S^{cur}) = f(S^{cur}) + \langle \nabla f(S^{cur}), S - S^{cur} \rangle + \frac{L(f)}{2} \|S - S^{cur}\|_F^2 + \alpha \|S\|_{2,1}, \tag{13}$$

where

$$\nabla f(S) = 2X_L^T (X_L S - Y_L V) + 2\beta X_{LU}^T L X_{LU} S, \tag{14}$$

and $L(f)$ is the Lipschitz constant of ∇f. By simple mathematical transformation, it is easy to check that minimizing surrogate function (13) is equivalent to solving the optimization problem

$$\min_S \frac{L(f)}{2} \left\| S - S^{cur} + \frac{\nabla f(S^{cur})}{L(f)} \right\|_F^2 + \alpha \|S\|_{2,1}, \tag{15}$$

whose solution is given by $\hat{S} = \vec{\Theta}\left(S^{cur} - \frac{\nabla f(S^{cur})}{L(f)}; \frac{\alpha}{L(f)}\right)$, where $\vec{\Theta}(\cdot, \lambda)$ is a row-wise soft thresholding(group soft thresholding) operator defined as

$$\vec{\Theta}(\boldsymbol{v}; \lambda) = \boldsymbol{v}^o \Theta(\|\boldsymbol{v}\|_2; \lambda), \text{ and } \boldsymbol{v}^o = \begin{cases} \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}, & \text{if } \boldsymbol{v} \neq \mathbf{0} \\ \mathbf{0}, & \text{if } \boldsymbol{v} = \mathbf{0} \end{cases} \tag{16}$$

and $\Theta(\cdot; \lambda)$ is element-wise soft thresholding operator.

The surrogate function $G(S, S^{cur})$ is a quadratic approximation as well as an upper bound for the original objective function $F(S)$. The following theorem illustrates that minimizing the surrogate function (15) obtains the minimizer of problem $F(S)$:

Theorem 3.1. *Suppose $L_f \geq 2\lambda_{max}(X_L^T X_L + \beta X_{LU}^T L X_{LU})$. Then the sequence $\{S_k\}$ generated by problem (15) fulfills that*

$$F(S_{k+1}) \leq F(S_k), \quad \textit{for } \forall\, k. \tag{17}$$

Proof. Recall that the surrogate function at current point S_k is defined as

$$G(S, S_k) = f(S_k) + \langle \nabla f(S_k), S - S_k \rangle + \frac{L(f)}{2} \|S - S_k\|_F^2 + \alpha \|S\|_{2,1}. \tag{18}$$

Obviously,

$$G(S_{k+1}, S_k) \leq G(S_k, S_k) = F(S_k), \tag{19}$$

where S_{k+1} is the optimal solution of problem (15). From the Taylor expansion of f at S_k, we obtain that

$$\begin{aligned} f(S) &= f(S_k) + \langle \nabla f(S_k), S - S_k \rangle + \frac{1}{2}(S - S_k)^T \nabla^2 f(\xi)(S - S_k) \\ &\leq f(S_k) + \langle \nabla f(S_k), S - S_k \rangle + \frac{1}{2}(S - S_k)^T (\lambda_{\max}(\nabla^2 f(\xi)) I)(S - S_k) \\ &= f(S_k) + \langle \nabla f(S_k), S - S_k \rangle + \frac{1}{2} \lambda_{\max}(\nabla^2 f(S_k)) \|S - S_k\|_F^2 \end{aligned} \tag{20}$$

Where $\xi = \alpha S_k + (1 - \alpha) S, \alpha \in [0, 1]$. Taking $L_f \geq \lambda_{\max}(\nabla^2 f(S_k))$ and adding penalty $\|S\|_{2,1}$ to the both sides of (20), from (18) and (20), we have

$$F(S) \leq G(S, S_k), \quad \forall\, S. \tag{21}$$

Finally, let $S = S_{k+1}$, with (19) and (21), we obtain that

$$F(S_{k+1}) \leq F(S_k), \quad \forall\, k. \tag{22}$$

□

3.3 Summary of Our Algorithm

The complete algorithm of our MRSFE is summarized in Algorithm 1. Notice that our algorithm only consists of SVD decomposition of small-scale W, together with some thresholding operations which are at low cost. Moreover, we do not have to solve the inner loop exactly in Algorithm 1, which is time-consuming, rather, we set a small iteration number for the inner loop is sufficient, e.g. 5. This strategy is also termed as an inexact method, the convergence of this inexact method is guaranteed in [9]. As for the Lipschitz constant, we could simply set it as the maximum eigenvalue of the Hessian matrix of $f(S)$. The time complexity of our algorithm is at most $\mathcal{O}(cr^2) + \mathcal{O}(dr)$, which is significantly faster than state-of-the-art algorithms. The model (7) can also be solved in the framework of Algorithm 1, the only difference is that S update step is obtained by group hard thresholding not the group soft thresholding, where group hard thresholding is designed for $\ell_{2,0}$ penalty.

Algorithm 1 Manifold Regularization based Selectable Factor Extration

Input: Data matrix $X_L \in \mathbb{R}^{l\times d}, X_{LU} \in \mathbb{R}^{(l+u)\times d}$, label matrix $Y_L \in \mathbb{R}^{l\times c}$, desired rank r, parameters k, α, β and Lipschitz constant L_f.
Initialization: $S^{(0)}, V^{(0)}$ such that $(V^{(0)})^T V^{(0)} = I$.
Output: (S, V)

```
Repeat
    Compute W = Y_L^T X_L S = U_w Σ_W V_w^T, V ← U_w V_w^T.
    Repeat
        Compute ∇f(S)(Eq.(14)).
        Compute S ← Θ⃗( S − ∇f(S)/L_f ; α/L_f ).
    Until stopping criterion
Until stopping criterion
Select the significant features according to top-k index of the row-norms in descending
order or nonzero rows of S.
```

4 Experiments

In this section, we evaluate the effectiveness of our MRSFE by applying it to a challenge web image dataset, NUS-WIDE-OBJECT.

4.1 Dataset and Parameter Settings

The NUS-WIDE-OBJECT is a real world object image dataset consisting of 31 object categories and 30000 images in total. The dataset has 17927 images for training and 12703 for testing. Six different kinds of features are provided to represent each image sample, they are 500-D bag of visual words, 128-D wavelet texture, 73-D edge direction histogram,

64-D color histogram, 144-D color auto-correlogram, and 225-D block-wise color moments respectively.

In order to compare with other algorithms, we have removed images with multi labels for both training and test images, thus 14270 training images and 9683 test images are left in the dataset. In our experiments, we randomly select s={10,20,50} labeled training images from each category, namely, 310, 620 and 1550 training samples are used respectively. The remaining images in the training set are used as unlabeled samples. We utilize all 9683 single-label test samples for testing.

In particular, the following methods are compared with our algorithm:

AllFea: the baseline in which all features (1134-D) described before are concatenated and normalized.

CSFS[2]: a convex semi-supervised multi-label feature selection algorithm which can deal with large-scale datasets. The parameter μ is in the range of $\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$.

LSDF[22]: a semi-supervised feature selection method which labeled samples are used to maximize the margin between different categories, while the unlabeled samples are used to capture the manifold structure of the data. The parameter k which stands for the nearest neighbors is set in the range $\{3, 5, 8, 12, 15, 20\}$.

MRSFE: The proposed algorithm for feature selection. The parameter k is same with **LSDF**, the reduced rank parameter r is in $\{21, 23, 25, 27, 30\}$. The parameters of manifold regularization and sparse penalty are in the range of $\{10^{-3}, 5\times10^{-3}, 10^{-2}, \ldots, 1\}$.

4.2 Experimental Results and Analysis

As described above, we randomly select 10/20/50 labeled samples in each category, and both linear and non-linear SVM classifiers are applied to compare the performances of different semi-supervised feature selection methods. All of the experiments are repeated 10 times to report the average result. Figure 1 and 2 show annotation accuracy at different subset features r = {113, 226, 340, 453, 567, 680, 793, 907,1020} with linear/non-linear classifiers.

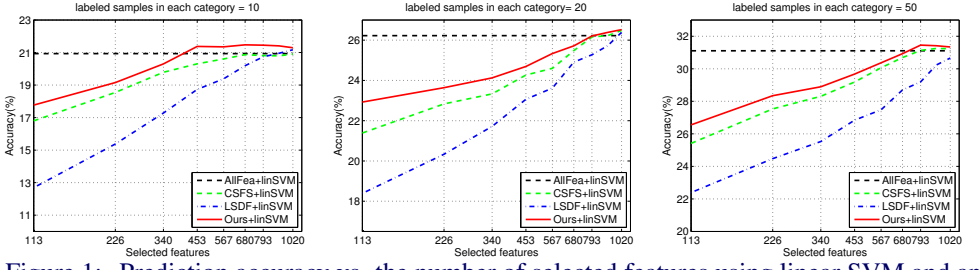

Figure 1: Prediction accuracy vs. the number of selected features using linear SVM and an increasing number of labeled samples

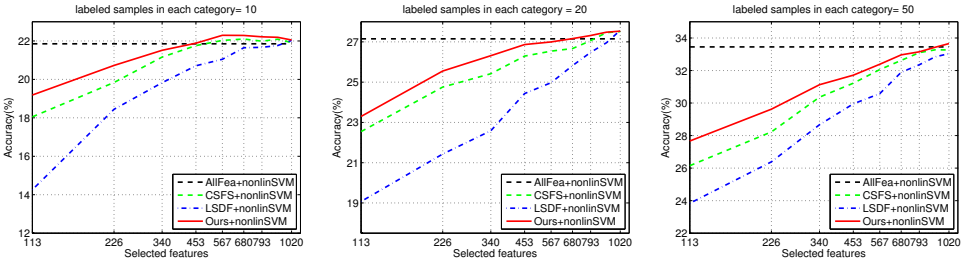

Figure 2: Prediction accuracy vs. the number of selected features using nonlinear SVM and an increasing number of labeled samples

It can be seen from the figures 1 and 2: (1) classification performance is improved with an increase of labeled samples. (2) with the increase of the selected features, the annotation accuracy gets better as more and more information is involved. (3) prediction using the subset of selected features can outperform the baseline which uses all the features in some dimensions, since feature selection methods select significant variables to predict and remove some redundant and noisy features. (4) the proposed method MRSFE outperforms the other two semi-supervised feature selection methods in almost all the cases, especially when the selected features are not sufficient, which owes to the low rank constraint of the transformation matrix.

To verify the superiority of the proposed method compared with the other semi-supervised feature selection approaches in utilizing the unlabeled data samples, we fix the number of labeled samples as 50 in each category, and vary the number of unlabeled samples to be 2000/4000/8000. The number of subset features are fixed as before and nonlinear SVM is applied to report the average result. From figure 3 we could see that: (1) the more unlabeled samples we use, the higher of average prediction accuracy we obtain. This is because the data distribution is better captured by manifold regularization if more samples are provided. (2) the proposed method MRSFE outperforms the other two methods in almost all cases, especially when the number of unlabeled samples is small. This may benefit from the low rank constraint on the transformation coefficients matrix. The constraint could help reducing the model complexity and thus the data distribution could be more effectively captured by manifold regularization.

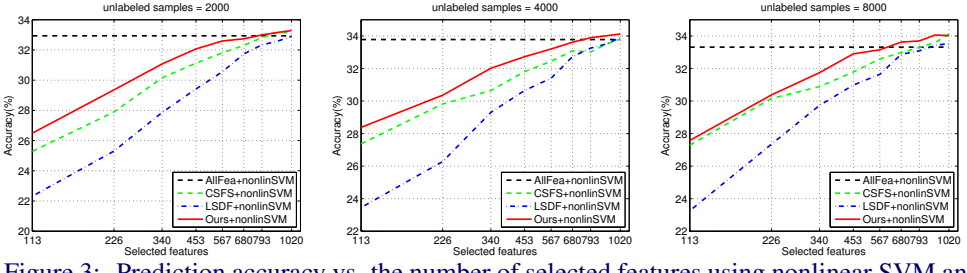

Figure 3: Prediction accuracy vs. the number of selected features using nonlinear SVM and an increasing number of unlabeled samples (#{labeled samples for each category} = 50)

We also give the time-consuming results of two semi-supervised feature selection methods and ours in Table 4.2. In this experiment, the number of labeled samples in each category is also set as 50, and the number of unlabeled samples is in $\{2000, 4000, 8000\}$. It can be seen that our method takes much less time than other methods, the reason is that our algorithm

only consists of some small scale SVD decompositions and group soft or hard thresholding operations.

unlabeled samples	2000	4000	6000	8000	10000
OURS(s)	**1.499**	**3.038**	**5.443**	**8.350**	**11.927**
CSFS(s)	3.031	6.550	11.343	14.672	20.409
LSDF(s)	2.061	4.624	9.113	13.897	20.366

Table 1: Average time on different unlabeled samples sets.

5 Conclusion

In this paper, we propose a manifold regularized selectable factor extraction method for semi-supervised feature selection problem. We use both low rank and sparse penalty to explore the intrinsic property of feature transformation matrix, and the structure of data distribution is well captured by manifold regularization. Moreover, both the convex and nonconvex penalties are applied to the selected significant features. We integrate all this information into a unified learning framework and design a fast and easy-to-implement algorithm. Experiments on a challenge web image annotation task demonstrate the superiority of the proposed method.

Acknowledgement

Xin Shi, Fangyun Wei, Hongyang Zhang and Chao Zhang are supported by National Key Basic Research Project of China (973 Program) (No.s 2015CB352303 and 2011CB302400) and National Nature Science Foundation (NSF) of China (No.s 61071156 and 61131003).

References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[2] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014.

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, page 48, 2009.

[4] John Geweke. Bayesian reduced rank regression in econometrics. *Journal of econometrics*, 75(1):121–146, 1996.

[5] Thomas Hochkirchen. Modern multivariate statistical techniques: Regression, classification, and manifold learning. *JRSS: Series A (Statistics in Society)*, 173(2):467–467, 2010.

[6] Dong Huang and Fernando De la Torre. Bilinear kernel reduced rank regression for facial expression synthesis. In *ECCV*, pages 364–377. 2010.

[7] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *JMA*, 5(2):248–264, 1975.

[8] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.

[9] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, pages 612–620, 2011.

[10] Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE TIP*, 22(2):523–536, 2013.

[11] Yong Luo, Dacheng Tao, Chang Xu, Chao Xu, Hong Liu, and Yonggang Wen. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE TNNLS*, 24(5):709–722, 2013.

[12] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Decomposition-based transfer distance metric learning for image classification. *IEEE TIP*, 23(9):3789–3801, 2014.

[13] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Multi-view matrix completion for multi-label image classification. *IEEE TIP*, 24(8):2355–2368, 2015.

[14] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE TKDE*, page DOI: 10.1109/TKDE.2015.2445757, 2015.

[15] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[16] Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. In *ICML Workshop*, pages 1813–1821, 2006.

[17] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[18] Yiyuan She. Selectable factor extraction in high dimensions. *arXiv preprint arXiv:1403.6212*, 2014.

[19] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. Unsupervised feature selection for multi-view data in social media. In *SDM*, pages 270–278, 2013.

[20] Zenglin Xu, Irwin King, MR-T Lyu, and Rong Jin. Discriminative semi-supervised feature selection via manifold regularization. *IEEE TNN*, 21(7):1033–1047, 2010.

[21] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, page 1589, 2011.

[22] Jidong Zhao, Ke Lu, and Xiaofei He. Locality sensitive semi-supervised feature selection. *Neurocomputing*, 71(10):1842–1849, 2008.

[23] Peng Zhao and Bin Yu. On model selection consistency of lasso. *JMLR*, 7:2541–2563, 2006.