Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Co-Regularized Ensemble for Feature Selection

Yahong Han[1,3], **Yi Yang**[2], **Xiaofang Zhou**[2]
[1]School of Computer Science and Technology, Tianjin University, China
[2]School of Information Technology & Electrical Engineering, The University of Queensland
[3]Tianjin Key Laboratory of Cognitive Computing and Application
yahong@tju.edu.cn, yee.i.yang@gmail.com, zxf@itee.uq.edu.au

Abstract

Supervised feature selection determines feature relevance by evaluating feature's correlation with the classes. Joint minimization of a classifier's loss function and an $\ell_{2,1}$-norm regularization has been shown to be effective for feature selection. However, the appropriate feature subset learned from different classifiers' loss function may be different. Less effort has been made on improving the performance of feature selection by the ensemble of different classifiers' criteria and take advantages of them. Furthermore, for the cases when only a few labeled data per class are available, over-fitting would be a potential problem and the performance of each classifier is restrained. In this paper, we add a joint $\ell_{2,1}$-norm on multiple feature selection matrices to ensemble different classifiers' loss function into a joint optimization framework. This added co-regularization term has twofold role in enhancing the effect of regularization for each criterion and uncovering common irrelevant features. The problem of over-fitting can be alleviated and thus the performance of feature selection is improved. Extensive experiment on different data types demonstrates the effectiveness of our algorithm.

1 Introduction

Feature selection aims to reduce redundancy and noise in the original feature set. It has twofold role in improving both the efficiency and accuracy of data analysis. During recent years, feature selection has attracted much research attention [Gao *et al.*, 2011; Nie *et al.*, 2010; Cai *et al.*, 2011; Zhao and Liu, 2007; Yang *et al.*, 2011]. Supervised feature selection determines feature relevance by evaluating feature's correlation with the classes. It usually yields better and more reliable performance than unsupervised feature selection because of the utilization of class labels. However, in many real world applications, labeling a large number of training data is tedious and time-consuming. Provided the number of labeled training data is small, the performance of existing feature selection algorithms is usually restrained. Therefore, it turns out to be a great challenge to design a feature selection algorithm for the cases when only a few labeled data per class are available.

Suppose we have a supervised learning problem where the number of features is very large (compared to the number of labeled training data), but there is only a small number of features that are relevant to the learning task. In such a setting, over-fitting would be a potential problem. We can apply feature selection to reduce the number of features. Alternatively, we can use the technique of regularization to control the over-fitting phenomenon. For example, the penalty term of ℓ_1-norm in Lasso [Tibshirani, 1996] is used to induce sparse model, whereas the ℓ_2-norm in ridge regression [Hoerl and Kennard, 1970] is used to discourage the coefficients from reaching large values. In order to evaluate the importance of the selected features jointly, the $\ell_{2,1}$-norm regularized feature selection algorithms [Nie *et al.*, 2010; Cai *et al.*, 2011] utilize $\ell_{2,1}$-norm to control classifier's capacity and also ensure it is sparse in rows, making it particularly suitable for feature selection.

SVM maximizes the geometric margin of training data. The hyperplane for classification is constructed by finding a subset of the most discriminative training data, i.e., the support vectors. Whereas, least square regression minimizes the sum of squared residuals between an observed value and the fitted value provided by a model and all the training data. If we add an $\ell_{2,1}$-norm to different loss functions of SVM and least square regression, e.g., $\ell_{2,1}$-norm SVM [Cai *et al.*, 2011] and $\ell_{2,1}$-norm least square regression [Nie *et al.*, 2010], the selected feature subset learned from the same training set may be different. This disagreement motivates us to ensemble different criteria and take advantages of them, the goal of which is to obtain an optimal feature subset. Minimization of a co-regularization term has been shown to be an effective way to reduce disagreement of different classifiers [Brefeld *et al.*, 2006; Sindhwani and Rosenberg, 2008]. In this paper, we propose to ensemble different feature selection algorithms by adding a joint $\ell_{2,1}$-norm of multiple feature selection matrices which correspond to different feature selection algorithms. This joint $\ell_{2,1}$-norm plays the role of co-regularization as follows. In each round of the alternating optimization algorithm developed in this paper, the updated feature selection matrices in the former rounds can be used to regularize the current optimization criterion. Thus, the joint $\ell_{2,1}$-norm can en-

hance the regularization effect for each criterion. The benefits are: On one hand, for the cases when the number of labeled training data is small, co-regularization can further alleviate over-fitting. On the other hand, common irrelevant or noisy features in different feature selection matrices should be uncovered, which results in an optimal feature subset.

Note that the co-regularized classifiers in this paper are trained on the same training set represented in the same feature space, which is different from the multi-task learning [Argyriou *et al.*, 2008; Ma *et al.*, 2012; Yang *et al.*, 2013] and the co-regularization methods in multi-view learning [Brefeld *et al.*, 2006; Sindhwani and Rosenberg, 2008]. Although previous works [Opitz, 1999; Tsymbal *et al.*, 2003] also addressed *feature selection* and *ensembles*, the goals are to improve classification by re-sampling different feature subsets, which is not for feature selection. How to ensemble different classifiers by a co-regularized $\ell_{2,1}$-norm for feature selection has not been explored in previous works.

2 The Objective Function

In this section, we give the objective function of the co-regularized Ensemble for Feature Selection (EnFS) algorithm proposed in this paper. Later in the next section, we propose an efficient algorithm to optimize the objective function. It is worth mentioning that EnFS aims to select discriminative features according to labels of the training data, where different models of supervised feature selection are integrated through a co-regularization term, making it different from existing feature selection algorithms.

Denote $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ as the training set, where $x_i \in \mathbb{R}^d (i = 1, \ldots, n)$ is the i-th datum and n is the total number of training data. Suppose the n training data $x_1, x_2, \ldots, x_n$ are sampled from c classes, we define $y_i \in \{0,1\}^{c\times 1} (i = 1, \ldots, n)$ 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. $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ is the data matrix. $Y = [y_1, y_2, \ldots, y_n] \in \{0,1\}^{c\times n}$ is the label matrix. In this paper, I is the 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 $H_m = I - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T \in \mathbb{R}^{m\times m}$ is the centering matrix. For an arbitrary matrix $A \in \mathbb{R}^{r\times p}$, its $\ell_{2,1}$-norm is defined as $||A||_{2,1} = \sum_{i=1}^{r}\sqrt{\sum_{j=1}^{p} A_{ij}^2}$. $Tr(\cdot)$ represents the trace operator.

Suppose we have t feature selection matrices $W_l \in \mathbb{R}^{d\times c} (l = 1, \ldots, t)$, which are learned respectively from t classifiers:

$$\min_{W_l} L^l(W_l^T X, Y) + \lambda_l J^l(W_l), \tag{1}$$

where $l = 1, \ldots, t$ and $\lambda_l > 0$ is the regularization parameter of the l-th classifier. Function L^l and J^l are the loss and regularization functions, respectively. In Eq. (1), regularization function J^l has twofold role in controlling the capacity of W_l to avoid over-fitting and inducing sparsity in W_l for feature selection. For example, as indicated in [Nie *et al.*, 2010; Yang *et al.*, 2011], when minimizing the $\ell_{2,1}$-norm of W_l, some rows of W_l shrink to zero, making W_l particularly suitable for feature selection. As mentioned before, when using $W_l (l = 1, \ldots, t)$ for feature selection, the selected appropriate feature subset may be different for different classifiers. That is, if we uncover the common irrelevant or noisy components in $W_l (l = 1, \ldots, t)$, we can expect a better performance of feature selection.

Denote $W = [W_1, \ldots, W_t] \in \mathbb{R}^{d\times(\sum_{l=1}^{t} c)}$ and define a co-regularization function $\Omega(W)$, the objective function of the co-regularized ensemble for feature selection is

$$\min_{W} \sum_{l=1}^{t} \Big(L^l(W_l^T X, Y) + \lambda_l J^l(W_l) \Big) + \alpha\Omega(W), \tag{2}$$

where $\alpha > 0$ is the regularization parameter. Note that, the t classifiers in Eq. (2) are trained on the same training set $\{X, Y\}$, making it different from the multi-task learning methods [Argyriou *et al.*, 2008; Ma *et al.*, 2012; Yang *et al.*, 2013]. The minimization problems in Eq. (2) can be solved using an alternating optimization method. That is, the remaining $t-1$ feature selection matrices $W_j (j = 1, \ldots, l-1, l+1, \ldots, t)$ are fixed when optimizing the l-th classifier. Thus, the corresponding fixed parts in the co-regularization function $\Omega(W)$ can be used to regularize the l-th classifier for the current iteration round. In this way, the problem of over-fitting can be further alleviated, especially for the cases when only a few labeled data per class are available. Furthermore, the minimization of the co-regularization function $\Omega(W)$ also uncovers the common irrelevant or noisy components in the t feature selection matrices $W_l (l = 1, \ldots, t)$. The optimal W may lead to a better performance of feature selection than using each $W_l (l = 1, \ldots, t)$, which are estimated respectively from each classifier.

In our EnFS algorithm, we use two representative loss functions of multi-class SVM [Crammer and Singer, 2002] and least square regression. The goal is to take advantages of them and obtain an optimal feature subset. For each classifier, we propose to select the features that are most correlated to labels and rewrite Eq. (2) as follows for feature selection.

$$\begin{aligned}\min_{W} \Big(& f^{\text{svm}}(W_1^T X, Y) + \lambda_1 ||W_1||_{2,1} \\ & + ||W_2^T X + b\mathbf{1}_n^T - Y||_F^2 + \lambda_2 ||W_2||_{2,1} \Big) \\ & + \alpha ||W||_{2,1},\end{aligned} \tag{3}$$

where $b \in \mathbb{R}^{c\times 1}$ is the bias term for the least square regression. In Eq. (3), $W = [W_1, W_2]$ and $t = 2$. Function f^{svm} is the multi-class hinge loss defined as [Crammer and Singer, 2002; Cai *et al.*, 2011]

$$f^{\text{svm}}(W_1^T X, Y) = \sum_{i=1}^{n} \big(1 - \mathbf{w}_{\mathbf{y}_i}^T x_i + \max_{m\neq \mathbf{y}_i} \mathbf{w}_m^T x_i\big)_+, \tag{4}$$

where $W_1 = [\mathbf{w}_1, \ldots, \mathbf{w}_c]$. The index $\mathbf{y}_i \in \{1, \ldots, c\}$ and $\mathbf{y}_i = j$ if $Y_{ji} = 1$. In Eq. (3), with the term $||W_1||_{2,1}$ and $||W_2||_{2,1}$, our algorithm is able to evaluate the informativeness of all features jointly for the multi-class SVM and least square regression, respectively. The co-regularization term $||W||_{2,1}$, on the other hand, enables W_1 and W_2 to have the same sparse patterns and share the common components. In the optimization of Eq. (3), the fixed part of W_1 in $||W||_{2,1}$

can be used to regularize the least square regression, and vice versa, which can help to tackle the problem of over-fitting and result in an optimal W for feature selection.

3 Optimization of EnFS Algorithm

In this section, we introduce the optimization of the objective function in Eq. (3). In fact, there are two unknown variables W_1 and W_2 to be estimated. We propose an alternating optimization algorithm to solve the optimization problem in Eq. (3).

Denote $W_l = [w_l^1, \ldots, w_l^d]$ $(l = 1, 2)$ and $W = [w^1, \ldots, w^d]$, where d is the number of features. We first fix W_2, and the optimization problem becomes:

$$\min_{W_1} f^{\text{svm}}(W_1^T X, Y) + \lambda_1 ||W_1||_{2,1} + \alpha ||W||_{2,1}, \quad (5)$$

following [Ma *et al.*, 2012], we rewrite Eq. (5) as follows:

$$\begin{aligned} &\min_{W_1} f^{\text{svm}}(W_1^T X, Y) + \lambda_1 Tr(W_1^T D_1 W_1) \\ &+\alpha\Big(Tr(W_1^T D W_1) + Tr(W_2^T D W_2)\Big), \end{aligned} \quad (6)$$

where D_1 and D are diagonal matrices with each element on the diagonal, i.e., d_{ii}^1 and d_{ii} $(i = 1, \ldots, d)$, are respectively defined as

$$d_{ii}^1 = \frac{1}{2||w_1^i||_2} \text{ and } d_{ii} = \frac{1}{2||w^i||_2}. \quad (7)$$

Omit the fixed part $Tr(W_2^T D W_2)$ in Eq. (6), we have

$$\begin{aligned} &\min_{W_1} f^{\text{svm}}(W_1^T X, Y) + \lambda_1 Tr(W_1^T D_1 W_1) \\ &+\alpha Tr(W_1^T D W_1) \\ \Rightarrow \quad &\min_{W_1} f^{\text{svm}}(W_1^T X, Y) \\ &+\lambda_1 Tr\Big(W_1^T (D_1 + \mu D) W_1\Big), \end{aligned} \quad (8)$$

where $\mu = \alpha/\lambda_1$. Denote $U = D_1 + \mu D$ we have

$$\min_{W_1} f^{\text{svm}}(W_1^T X, Y) + \lambda_1 Tr(W_1^T U W_1). \quad (9)$$

Eq. (9) is the objective function of multi-class $\ell_{2,1}$-norm SVM [Cai *et al.*, 2011], which can be efficiently solved by alternately updating W_1 and U until convergency. When U is fixed, let $W_1^* = U^{\frac{1}{2}} W$ and $X^* = U^{-\frac{1}{2}} X$, objective function in Eq. (9) is equivalent to

$$\begin{aligned} &\min_{W_1} f^{\text{svm}}(W_1^T U^{\frac{1}{2}} U^{-\frac{1}{2}} X, Y) + \lambda_1 (W_1^T U W_1) \\ \Rightarrow \quad &\min_{W_1^*} f^{\text{svm}}(W_1^{*T} X^*, Y) + \lambda_1 Tr(W_1^{*T} W_1^*), \end{aligned} \quad (10)$$

which can be solved by Crammer's algorithm [Crammer and Singer, 2002] using LIBLINEAR [Fan *et al.*, 2008]. Let $\hat{W_1}^*$ be the solution of Eq. (10), we have

$$W_1 = U^{-\frac{1}{2}} \hat{W_1}^*. \quad (11)$$

Then we fix W_1, the objective function in Eq. (3) becomes

$$\min_{W_2} ||W_2^T X + b\mathbf{1}_n^T - Y||_F^2 + \lambda_2 ||W_2||_{2,1} + \alpha ||W||_{2,1}, \quad (12)$$

Algorithm 1 Ensemble Feature Selection (EnFS)

Input: Input data $X \in \mathbb{R}^{d \times n}$ and labels $Y \in \{0, 1\}^{c \times n}$. Regularization parameters λ_1, λ_2, and α
Output: Matrix $W \in \mathbb{R}^{d \times c}$

```
Set r = 0 and initialize W_1 ∈ R^{d×c} and W_2 ∈ R^{d×c} randomly;
W = [W_1, W_2];
repeat
    Compute the diagonal matrix D_1^r, D_2^r, and D^r according to d_ii^1 = 1/(2||w_1^i||_2), d_ii^2 = 1/(2||w_2^i||_2), and d_ii = 1/(2||w^i||_2);
    repeat
        Compute W_1^* = U^{1/2} W and X^* = U^{-1/2} X;
        Solve Eq. (10) and compute W_1^*;
        Update U by U = D_1 + (α/λ_1) D;
    until Convergence
    Update W_1 by W_1^{r+1} = U^{-1/2} Ŵ_1^*;
    Update W_2 by W_2^{r+1} = (X H_n H_n^T X^T + λ_2 D_2 + αD)^{-1} X H_n Y^T;
    Update W^{r+1} = [W_1, W_2];
    r = r + 1;
until Convergence
Return W.
```

which is equivalent to

$$\begin{aligned} &\min_{W_2} ||W_2^T X + b\mathbf{1}_n^T - Y||_F^2 + \lambda_2 Tr(W_2^T D_2 W_2) \\ &+\alpha Tr(W_2^T D W_2), \end{aligned} \quad (13)$$

where D_2 is a diagonal matrix with each element d_{ii}^2 $(i = 1, \ldots, d)$ on the diagonal, which is defined as

$$d_{ii}^2 = \frac{1}{2||w_2^i||_2}. \quad (14)$$

By setting the derivative of Eq. (13) w.r.t. b to 0. we have

$$b = \frac{1}{n} Y \mathbf{1}_n - \frac{1}{n} W_2^T X \mathbf{1}_n. \quad (15)$$

Substituting Eq. (15) into Eq. (13) we obtain

$$\begin{aligned} &\min_{W_2} \Big\| W_2^T X + \big(\frac{1}{n} Y \mathbf{1}_n - \frac{1}{n} W_2^T X \mathbf{1}_n\big) \mathbf{1}_n^T - Y \Big\|_F^2 \\ &+\lambda_2 Tr(W_2^T D_2 W_2) + \alpha Tr(W_2^T D W_2) \\ \Rightarrow \quad &\min_{W_2} \Big\| W_2^T X \big(I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T\big) - Y \big(I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T\big) \Big\|_F^2 \\ &+\lambda_2 Tr(W_2^T D_2 W_2) + \alpha Tr(W_2^T D W_2) \\ \Rightarrow \quad &\min_{W_2} \Big\| W_2^T X H_n - Y H_n \Big\|_F^2 + \lambda_2 Tr(W_2^T D_2 W_2) \\ &+\alpha Tr(W_2^T D W_2). \end{aligned} \quad (16)$$

By setting the derivative of Eq. (16) w.r.t. W_2 to 0, we have

$$X H_n H_n^T X^T W_2 + \lambda_2 D_2 W_2 + \alpha D W_2 = X H_n Y^T. \quad (17)$$

Therefore, we have

$$W_2 = (X H_n H_n^T X^T + \lambda_2 D_2 + \alpha D)^{-1} X H_n Y^T. \quad (18)$$

Table 1: Dataset Description.

Dataset	Size	# of Features	# of Classes
MIML	2000	128	5
MFlickr	25000	128	33
USPS	9298	256	10
YaleB	2414	1024	38
Protein	21516	357	3
SensIT	19705	100	3

Based on the above mathematical deduction, we propose an alternating algorithm to optimize the objective function in Eq. (3), which is summarized in Algorithm 1. Once W is obtained, we sort the d features according to $||w^i||_F$ $(i = 1, \ldots, d)$ in descending order and select the top ranked ones. The convergence of solving Eq. (9) by alternately updating W_1 and U has been proved in [Cai *et al.*, 2011]. The object function in Eq. (13) is equivalent to

$$\min_{W_2} ||W_2^T X + b\mathbf{1}_n^T - Y||_F^2 + \lambda_2 Tr\Big(W_1^T(D_1 + \frac{\alpha}{\lambda_2}D)W_1\Big),$$

which is the object function of the $\ell_{2,1}$-norm regularized least square regression and its convergence has been proved in [Nie *et al.*, 2010]. Thus, when we alternately fix the values of W_1 and W_2, the optimal solutions obtained from Algorithm 1 make the value of objective functions decreased and Algorithm 1 is guaranteed to be converged.

4 Experiments

4.1 Experiment Setup

In our experiment, we have collected a diversity of 6 public datasets to compare the performance of different feature selection algorithms. These datasets include two color image datasets, i.e., MIML [Zhou and Zhang, 2007] and MFlickr [Huiskes and Lew, 2008], one hand written digit image dataset, i.e., USPS [Hull, 1994], one face image dataset, i.e., YaleB [Georghiades *et al.*, 2001], one protein gene data, i.e., Protein [Wang, 2002], and one vehicle classification data in distributed sensor networks, i.e., SensIT Vehicle [Duarte and Hen Hu, 2004]. Detailed information of the six datasets is summarized in Table 1. For MIML and MFlickr, we extract the 128 dimensional color coherence as features for each image. As there exists the multi-labeled samples in MIML and MFlickr, we remove these samples and let each sample only belong to one class.

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

- Full Features which adopts all the features for classification. It is used as baseline method in this paper.
- Fisher Score [Duda *et al.*, 2001] which depends on fully labeled training data to select features with the best discriminating ability.
- Feature Selection via Joint $\ell_{2,1}$-Norms Minimization (FSNM) [Nie *et al.*, 2010] which employs joint $\ell_{2,1}$-norm minimization on both loss function and regularization to realize feature selection across all data points.
- Sparse Multinomial Logistic Regression via Bayesian ℓ_1 Regularization (SBMLR) [Cawley *et al.*, 2007] which exploits sparsity by using a Laplace prior and is used for multi-class pattern recognition. It can also be applied to feature selection.
- Feature Selection via Spectral Analysis (FSSA) [Zhao and Liu, 2007] which is a semi-supervised feature selection method using spectral regression.
- Multi-class $\ell_{2,1}$-norm Support Vector Machine (SVM-21) [Cai *et al.*, 2011] which is an $\ell_{2,1}$-norm regularized SVM. It is one of the feature selection methods ensembled in the proposed EnFS.
- $\ell_{2,1}$-norm Least Square Regression (LSR-21) [Nie *et al.*, 2010] which is an $\ell_{2,1}$-norm regularized least square regression. It is another feature selection method ensembled in the proposed EnFS.

In this experiment, we set λ_1 and λ_2 in Eq. (3) to the same value for all the datasets. Thus, for the proposed EnFS, parameters λ (denotes the same value of λ_1 and λ_2) and α are tuned. To fairly compare different feature selection algorithms, we tune all the parameters (if any) by a "grid-search" strategy from $\{10^{-6}, 10^{-5}, \ldots, 10^5, 10^6\}$. We set the number of selected features as $\{10, 20, \ldots, 90\}$ for MIML, MFlickr, and SensIT, $\{20, 40, \ldots, 200\}$ for USPS, $\{100, 200, \ldots, 900\}$ for YaleB, and $\{30, 60, \ldots, 300\}$ for Protein. For all the algorithms, we report the best results obtained from different parameters. To investigate the performance of feature selection for the cases when only a few labeled data per class are available, we set the number of labeled data per class to 5 and randomly sample these labeled data to form the training sets. For each dataset, we repeat the sampling for 10 times and report the average results.

In our experiment, each feature selection algorithm is first performed to select features. Then three classifiers, i.e., linear multi-class SVM [Crammer and Singer, 2002], least square regression, and multi-class kNN ($k = 1$), are performed based on the selected features respectively to investigate the performance of feature selection. For the classifier of least square regression, we learn a threshold from the labeled training data to quantize the continuous label prediction scores to binary. We use Accuracy as evaluation metric in this paper.

4.2 Experimental Results and Discussions

First, we compare the performance of different feature selection algorithms. In Table 2, 3, and 4, We report the comparisons of feature selection algorithms on six datasets using multi-class SVM, least square regression, and multi-class kNN ($k = 1$) as classifiers, respectively. We can see from the three tables that the classification results of feature selection algorithms are generally better than that of Full Features, except for SBMLR. Moreover, the feature number is reduced by performing feature selection, resulting in the classification much faster. Therefore, it is more efficient. We can also see from these three tables that EnFS algorithm proposed in this paper obtains the best performance. For all these methods, the number of labeled data per class is set to 5. These results demonstrate that the proposed EnFS obtains better performance for the cases when only a few labeled data per class

Table 2: Comparisons of feature selection algorithms on six datasets in terms of Accuracy using linear multi-class SVM as classifier. For each method, the results are obtained when the number of labeled data per class is set to 5.

	Full Features	Fisher Score	FSNM	SBMLR	FSSA	SVM-21	LSR-21	EnFS
MIML	0.3100	0.3309	0.3257	0.2311	0.3449	0.3204	0.3368	**0.3531**
MFlickr	0.1022	0.1577	0.2012	0.0628	0.1912	0.1476	0.2124	**0.2634**
USPS	0.7934	0.7967	0.7972	0.6867	0.7968	0.7955	0.7978	**0.8025**
YaleB	0.7761	0.7807	0.7793	0.7783	0.7771	0.7784	0.7827	**0.7905**
Protein	0.3623	0.3799	0.3869	0.3798	0.3868	0.3764	0.3868	**0.3983**
SensIT	0.6726	0.6853	0.6921	0.6302	0.6833	0.6795	0.6880	**0.7028**

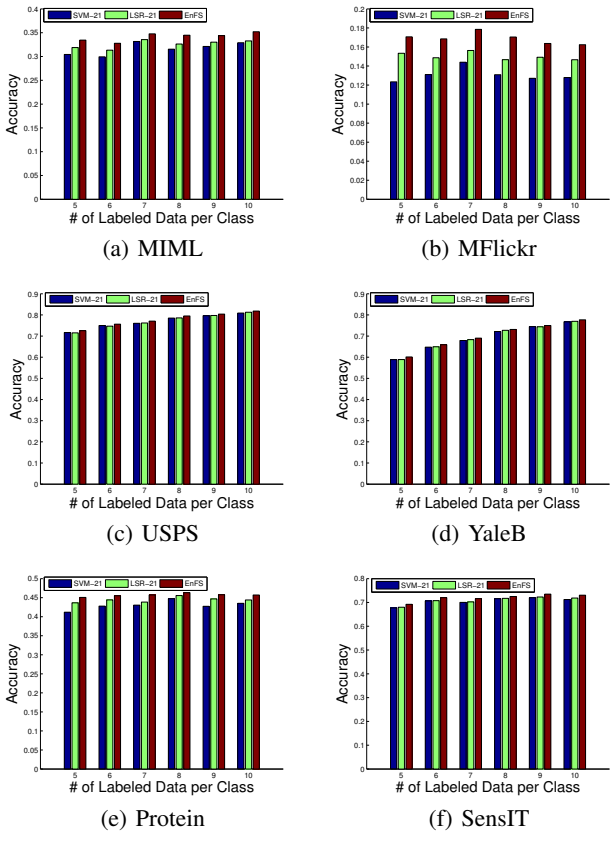
Table 3: Comparisons of feature selection algorithms on six datasets in terms of Accuracy using least square regression as classifier. For each method, the results are obtained when the number of labeled data per class is set to 5.

	Full Features	Fisher Score	FSNM	SBMLR	FSSA	SVM-21	LSR-21	EnFS
MIML	0.2474	0.2816	0.2911	0.2260	0.2741	0.2837	0.2963	**0.3144**
MFlickr	0.0481	0.0655	0.0816	0.0340	0.0657	0.0568	0.0707	**0.0859**
USPS	0.5265	0.5311	0.5391	0.4369	0.5286	0.5657	0.5602	**0.5944**
YaleB	0.6113	0.6247	0.6329	0.6155	0.6261	0.6238	0.6355	**0.6470**
Protein	0.1479	0.3126	0.3218	0.3262	0.3591	0.2099	0.3272	**0.3678**
SensIT	0.5836	0.5770	0.6127	0.5956	0.6069	0.5981	0.6096	**0.6393**

are available. Thirdly, we observe that, except for the USPS and YaleB datasets, the performance of EnFS is conspicuously better than that of SVM-21 and LSR-21. This observation validates that it is effective to add a co-regularization term $||W||_{2,1}$ (in Eq. (3)) to share the sparse patterns in W and help to alleviate over-fitting. Finally, we observe that classification using multi-class SVM and kNN achieve better performance than the least square regression. The main reason is, for classification evaluation of least square regression, the threshold learned from the small size of training data (i.e., 5 per class) leads to a bias in the quantization of continuous label prediction scores.

To further investigate the effectiveness of ensemble by the co-regularization term $||W||_{2,1}$, we compare the performance of SVM-21, LSR-21, and EnFS, when the numbers of labeled data per class are set to $\{5, 6, \ldots, 10\}$. The results are plotted in Fig. 1, which are obtained by performing kNN ($k = 1$) classifier based on the selected features. We can see from Fig. 1 that the performance of EnFS is generally better than that of SVM-21 and LSR-21 for all the numbers of labeled data per class. As the numbers of labeled data per class are small (compared to the sizes of six datasets, see Table 1), the results in Fig. 1 further validates the effectiveness of EnFS in alleviating the problem of over-fitting.

Fig. 2 demonstrates the learned matrix W of MIML dataset using algorithms SVM-21, LSR-21, and EnFS, respectively. Columns represent the classes and rows represent the 128 color coherence features. Firstly, we can observe from Fig. 2 that, when minimizing the $\ell_{2,1}$-norm of W, many rows of W shrink to zeros and results in a sparse pattern of feature selection for all classes. Secondly, it is clear in Fig. 2 that SVM-21 and LSR-21 output different sparse patterns of W by training on the same set of labeled data. This disagreement of spars patterns will result in different subsets of selected features. Finally, the results in Fig. 2 demonstrate the ensemble effect



(a) MIML (b) MFlickr

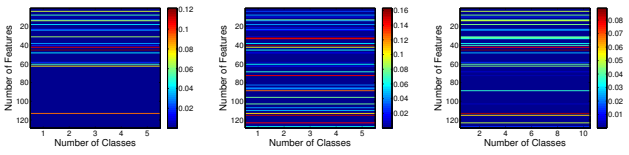
(c) USPS (d) YaleB

(e) Protein (f) SensIT

Figure 1: Classification results of SVM-21, LSR-21, and EnFS when the numbers of labeled data per class are set to 5, 6, 7, 8, 9, and 10, respectively. For all the six datasets, the results are obtained by performing kNN ($k = 1$) classifier based on the selected features.

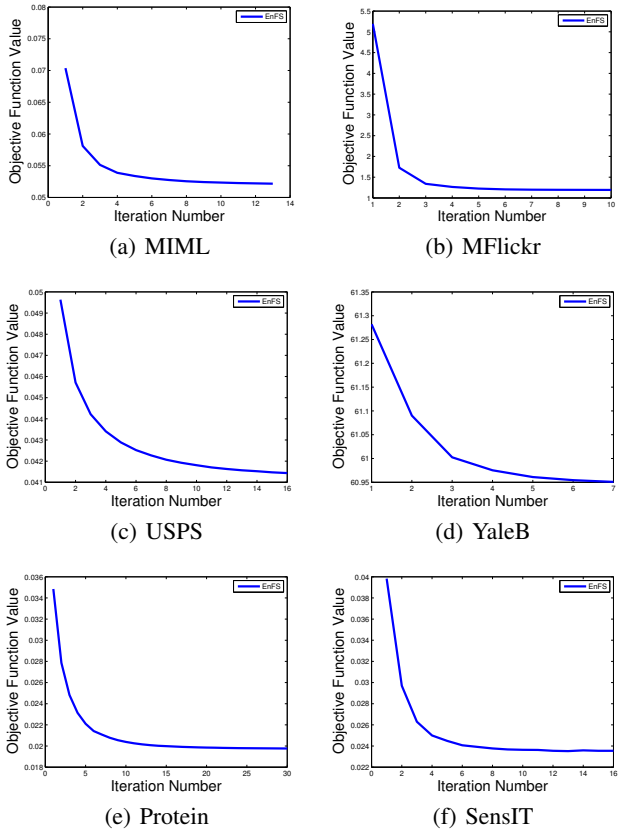
Table 4: Comparisons of feature selection algorithms on six datasets in terms of Accuracy using kNN ($k = 1$) as classifier. For each method, the results are obtained when the number of labeled data per class is set to 5.

	Full Features	Fisher Score	FSNM	SBMLR	FSSA	SVM-21	LSR-21	EnFS
MIML	0.2966	0.3124	0.2857	0.2316	0.3294	0.3043	0.3188	**0.3345**
MFlickr	0.0986	0.1261	0.1584	0.0628	0.1487	0.1234	0.1534	**0.1707**
USPS	0.6889	0.7053	0.7019	0.5253	0.6952	0.7167	0.7151	**0.7255**
YaleB	0.5324	0.5763	0.5841	0.5419	0.5699	0.5886	0.5891	**0.6012**
Protein	0.3667	0.4319	0.4396	0.4496	0.4428	0.4119	0.4364	**0.4507**
SensIT	0.6697	0.6671	0.6893	0.4907	0.6858	0.6770	0.6798	**0.6921**



(a) SVM-21 (b) LSR-21 (c) EnFS

Figure 2: The learned feature selection matrix W of MIML using algorithms SVM-21, LSR-21, and EnFS. Columns represent the classes and rows represent the 128 color coherence features. Dark blue denotes the values are close to zero.



(a) MIML (b) MFlickr

(c) USPS (d) YaleB

(e) Protein (f) SensIT

Figure 3: Convergence curves of the objective function value in Eq. (3) using Algorithm 1. The figure show that the objective function value monotomically decreases until converged by applying the proposed algorithm.

of EnFS for SVM-21 and LSR-21 clearly. For example, as the numbers of sparse row in the learned W using SVM-21 and LSR-21 are different, EnFS makes a compromise.

Next, we study the convergence of the proposed EnFS in Algorithm 1. Fig. 3 shows the convergence curves of our EnFS algorithm according to the objective function value in Eq. (3) on all the datasets. The figure shows that the objective function value monotonically decreases until converged by applying the proposed algorithm. It can be seen that our algorithm converges within a few iterations. For example, it takes no more than 10 iterations for MFlickr and YaleB and no more than 20 iterations for MIML, USPS, and SensIT. For Protein, it takes 30 iterations to converge.

5 Conclusion

While supervised feature selection has been well explored in lots of previous works, it is not straightforward to ensemble different classifiers for a better performance of feature selection, especially for the cases when the number of labeled training data is small. As over-fitting may be a potential problem in such a setting, we have proposed a method of co-regularized ensemble for feature selection. The co-regularized $\ell_{2,1}$-norm can enhance the effect of regularization and uncover common irrelevant features, so as to improve the performance of feature selection in some cases. Ensemble classifiers in the proposed EnFS are trained on the same training set represented in the same feature space, making it different from the multi-task learning methods and existing co-regularization algorithms in multi-view learning. To show the ensemble effect of the proposed general framework of EnFS, we ensemble SVM and least square regression to show the performance improvement of feature selection. It is worth noting that, when one of the ensembled classifiers uncovers the irrelevant features, classifier ensemble for feature selection not necessarily improves the performance. That is to say, performance improvement of classifier ensemble for feature selection is data dependant, which will be further explored in the future. EnFS also has an extension ability of incorporating other criteria of classification into the optimization process or define new form of regularization term.

Acknowledgments

This paper was partially supported by the NSFC (under Grant 61202166) and Doctoral Fund of Ministry of Education of China (under Grant 20120032120042).

References

[Argyriou *et al.*, 2008] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[Brefeld *et al.*, 2006] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd international conference on Machine learning*, pages 137–144, 2006.

[Cai *et al.*, 2011] X. Cai, F. Nie, H. Huang, and C. Ding. Multi-class $\ell_{2,1}$-norm support vector machine. In *IEEE 11th International Conference on Data Mining*, pages 91–100, 2011.

[Cawley *et al.*, 2007] G.C. Cawley, N.L.C. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in Neural Information Processing Systems*, 19, 2007.

[Crammer and Singer, 2002] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[Duarte and Hen Hu, 2004] M.F. Duarte and Y. Hen Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2004.

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

[Fan *et al.*, 2008] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Gao *et al.*, 2011] C. Gao, N. Wang, Q. Yu, and Z. Zhang. A feasible nonconvex relaxation approach to feature selection. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 356–361, 2011.

[Georghiades *et al.*, 2001] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[Hoerl and Kennard, 1970] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[Huiskes and Lew, 2008] M.J. Huiskes and M.S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.

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

[Ma *et al.*, 2012] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 469–478. ACM, 2012.

[Nie *et al.*, 2010] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *Advances in Neural Information Processing Systems*, 23:1813–1821, 2010.

[Opitz, 1999] D.W. Opitz. Feature selection for ensembles. In *Proceedings of the National Conference on Artificial Intelligence*, pages 379–384. AAAI Press, 1999.

[Sindhwani and Rosenberg, 2008] V. Sindhwani and D.S. Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983, 2008.

[Tibshirani, 1996] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Tsymbal *et al.*, 2003] A. Tsymbal, S. Puuronen, and D.W. Patterson. Ensemble feature selection with the simple bayesian classification. *Information Fusion*, 4(2):87–100, 2003.

[Wang, 2002] J.Y. Wang. *Application of support vector machines in bioinformatics*. PhD thesis, National Taiwan University, 2002.

[Yang *et al.*, 2011] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, and X. Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, pages 1589–1594, 2011.

[Yang *et al.*, 2013] Y. Yang, Z. Ma, A. Hauptmann, and N. Sebe. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia*, 15(3):661–669, 2013.

[Zhao and Liu, 2007] Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In *Proceedings of the 7th SIAM International Conference on Data Mining*, pages 1151–1158, 2007.

[Zhou and Zhang, 2007] Z.H. Zhou and M.L. Zhang. Multi-instance multi-label learning with application to scene classification. *Advances in Neural Information Processing Systems*, 19:1609–1616, 2007.