Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

A Convex Formulation for Semi-Supervised Multi-Label Feature Selection

Xiaojun Chang[1,2], **Feiping Nie**[2], **Yi Yang**[1] **and Heng Huang**[2*]
[1]School of Information Technology & Electrical Engineering, The University of Queensland.
[2]Department of Computer Science and Engineering, University of Texas at Arlington.
cxj273@gmail.com, feipingnie@gmail.com, yi.yang@uq.edu.au, heng@uta.edu.

Abstract

Explosive growth of multimedia data has brought challenge of how to efficiently browse, retrieve and organize these data. Under this circumstance, different approaches have been proposed to facilitate multimedia analysis. Several semi-supervised feature selection algorithms have been proposed to exploit both labeled and unlabeled data. However, they are implemented based on graphs, such that they cannot handle large-scale datasets. How to conduct semi-supervised feature selection on large-scale datasets has become a challenging research problem. Moreover, existing multi-label feature selection algorithms rely on eigen-decomposition with heavy computational burden, which further prevent current feature selection algorithms from being applied for big data. In this paper, we propose a novel convex semi-supervised multi-label feature selection algorithm, which can be applied to large-scale datasets. We evaluate performance of the proposed algorithm over five benchmark datasets and compare the results with state-of-the-art supervised and semi-supervised feature selection algorithms as well as baseline using all features. The experimental results demonstrate that our proposed algorithm consistently achieve superiors performances.

Introduction

With the booming of social networks, we have witnessed a dramatic increase of multimedia data, *i.e.* video, text and images, which has brought increasing demands of how to effectively organize and retrieve these data. A straightforward way is to correlate the semantic concepts of multimedia data and labels for subsequent management tasks. Hence, it is beneficial and necessary to improve semantic concept analyzing techniques. Normally, the aforementioned resources are represented by feature vectors, the dimensions of which are very large. Previous studies have demonstrated that only a subset of the features carry the most discriminating information and appropriately designed feature selection is able to obtain higher accuracy because of its capability of removing redundant and noisy information in the feature representation.

According to the availability of labels of training data, existing feature selection algorithms fall into two groups: supervised feature and semi-supervised feature selection (Xu and Jin 2010) (Kong and Yu 2010b). Supervised feature selection, *i.e.* Fisher score (Richard, Hart, and Stork 2001), only adopt labeled training data for feature selection. Higher accuracy and more reliable performance can be obtained with sufficient labeled training data. However, labeled training data are expensive and time-consuming to obtain in real-world applications (Luo et al. 2013). Inspired by the progress of semi-supervised learning, researchers have introduced semi-supervised learning to the field of feature selection. For example, Zhao *et al.* propose a semi-supervised feature selection algorithm based on spectral analysis in (Zhao and Liu 2007). However, these classical algorithms are only designed for single label dataset. To address multi-label problem, they decompose the multi-label learning to multiple independent single-label problem, which fails to take into consideration correlations between different labels (Ma et al. 2012b).

To tackle the multi-label feature selection problem, Ma *et al.* (Ma et al. 2012b) propose a feature selection technique which uncovers a feature subspace that is shared among multiple different classes. Their experiments validate that performance can be improved by mining correlations among multiple labels. Nevertheless, they design their approach in a supervised way. Another limitation is that their algorithm can not be applied for large-scale multimedia analysis because their solution involves an eigen-decomposition operation.

In order to solve the aforementioned problems, we propose a convex semi-supervised multi-label feature selection algorithm for large-scale multimedia analysis. Both labeled and unlabeled data are utilized to select features while correlations among different features are simultaneously taken into consideration. We name our algorithm Convex Semi-supervised multi-label Feature Selection (CSFS).

Taking image annotation as an example, the main steps of our method are as follows: We first represent all training and testing data with different types of features, followed by initializing labels of unlabled data to zero. Then, we conduct sparse feature selection and label prediction by minimizing

*To whom all correspondence should be addressed. This work was partially supported by US NSF IIS-1117965, IIS-1302675, IIS-1344152, Australian Research Council Project DE130101311.

the least square loss function. Afterwards, we only preserve the unlabeled training data with higher confidence and use them as new training data in the next step. Finally, we apply the obtained sparse coefficients for feature selection.

The main contributions of our work are:

1. Joint feature selection with sparsity and semi-supervised learning are combined into a single framework, which can select the most informative features with limited number of labeled training samples.
2. Different from traditional graph based semi-supervised algorithms, the computation cost of our algorithm is relatively low since it does not require the graph construction. Hence, it can be readily applied to large-scale datasets. Another novelty is that the proposed formulation is convex.
3. We propose a fast iterative algorithm to solve the non-smooth objective function. Different from existing multi-label feature selection algorithms, which involve with eigen-decomposition, the proposed algorithm only needs to solve several linear equation systems.
4. To evaluate performance of our algorithm, we apply it on several large-scale databases. The experimental results indicate that our algorithm consistently outperforms other compared algorithms on all the databases.

Related Work

Feature Selection

Existing feature selection algorithms are designed in various ways. According to whether the label information of training datfishera are available, feature selection algorithms fall into two categories: supervised and unsupervised feature selection. Supervised feature selection algorithms, *i.e.* Fisher Score (Richard, Hart, and Stork 2001) and ReliefF (Kenji and Rendell 1992), usually gain better and more reliable performances with sufficient labeled training data. However, they have two main limitations. First, they ignore the correlation between different features since they evaluate the features one by one. Second, labeled data are very expensive to obtain in the real-world applications.

Researchers have also proposed sparsity-based feature selection, which can mine correlations among different features (Tan, Wang, and Tsang 2010). Among these approaches, $l_{2,1}$-norm regularization has shown to be an effective model for sparse-based feature selection (Nie et al. 2010a; Cai et al. 2011; Wang et al. 2011; Cai, Nie, and Huang 2013).

Semi-Supervised Learning

Graph Laplacian based semi-supervised learning has gained increasing interest for its efficiency and simplicity. Nie *et al.* propose a manifold learning framework based on graph Laplacian and conduct extensive experiments to show its advantage over other state-of-art semi-supervised learning algorithms (Nie et al. 2010b). In (Ma et al. 2012a), Ma *et al.* propose a semi-supervised feature selection algorithm built upon manifold learning. Although their algorithms have shown good performances even with insufficient amount of labeled training data, they can not be readily applied to large-scale dataset since building graph Laplacian matrix on large-scale dataset is very time-consuming and unrealistic.

Multi-Label Classification

Although multi-label classification has attracted much research attention in recent years, very few research efforts have been made on multi-label feature selection (Kong and Yu 2010a) (Agrawal et al. 2013) (Wu, Yuan, and Zhuang 2010). Meanwhile, researchers have theoretically and empirically demonstrate that taking correlations between different labels into consideration can facilitate feature selection. For example, Ma *et al.* integrate shared subspace uncovering and joint sparse-based feature selection to mine the correlations among multiple labels in (Ma et al. 2012b). Nevertheless, they implement their approach in a supervised way.

The Proposed Framework

In this section, we first describe in detail the proposed algorithm. Then an efficient iterative algorithm is proposed to solve the objective function.

Problem Formulation

Let us denote $X = \{x_1, \ \ldots, \ x_n\}$ as the training sample matrix, where $x_i \in \mathbb{R}^d$ is the i-th data point and n is the total number of training samples. $Y = \{y_1, \ \ldots, \ y_{n_L}\}^T \in \{0, \ 1\}^{n_L \times c}$ is label matrix, c is the number of labels and n_L is the number of labeled training samples. $y_i \in \mathbb{R}^c$ is the label vector of the i-th sample. Y_{ij} is the j-th element of Y_i. $Y_{ij} := 1$ if x_i is associated with the j-th class and $Y_{ij} := 0$ otherwise. We denote a predicted label matrix $F = \begin{bmatrix} F_l \\ F_u \end{bmatrix} \in \mathbb{R}^{n \times c}$. For all the labeled training samples, $F_l = Y_l$, where F_l is predicted label matrix for labeled training data and F_u is predicted label matrix for unlabeled training data. For all the unlabeled training samples, the label vectors are set to zeros. We can generalize our algorithm as the following objective function:

$$\min_{f, F_l = Y_l} \sum_{i=1}^{n} loss(f(x_i), f_i) + \mu \Omega(f), \tag{1}$$

where $loss(\cdot)$ is a loss function and $\Omega(f)$ is the regularization term with μ as its parameter.

We can implement the semi-supervised multi-label feature selection in various ways with different loss functions and regularizations. Least square regression has been widely used in many applications for its efficiency and simplicity. By applying the least square loss function, the objective function is then defined as:

$$\min_{W, F, \mathbf{b}, F_l = Y_l} \sum_{i=1}^{n} s_i \|W^T x_i + \mathbf{b} - f_i\|_2^2 + \mu \|W\|_F^2, \tag{2}$$

where $\mathbf{1}$ denotes a column vector with all its elements being 1 and s_i is the score of one training data point. Empirically, the score of labeled training data is larger than unlabeled training data. In order to conduct effective feature selection,

it is beneficial to exert the sparse feature selection models on the regularization term. Nie *et al.* claim that $l_{2,1}$-norm based regularization is able to exert the sparse feature selection in (Nie et al. 2010a). By utilizing $l_{2,1}$-norm, our objective function arrives at:

$$\min_{W,F,\mathbf{b},F_l=Y_l} \sum_{i=1}^{n} s_i \|W^T x_i + \mathbf{b} - f_i\|_2^2 + \mu \|W\|_{2,1}. \quad (3)$$

$$s.t.\ 0 \le f_i \le 1$$

The most important part of this framework is the constraint above. Without this constraint, the solution will be trivial. It is worthwhile noticing that by adding another constraint $y_i^T \mathbf{1} = 1$ to our objective function, the proposed framework can be readily applied to semi-supervised single-label feature selection.

Optimization

Since the objective function is non-smooth and difficult to solve, we propose to solve it as follows.

First, by denoting S as a matrix with its diagonal elements $S_{ii} = s_i$, we write the objective function shown in (3) as follows.

$$\min_{W,F,\mathbf{b},F_l=Y_l} Tr((X^T W + \mathbf{1}b^T - F)^T S (X^T W + \mathbf{1}b^T - F)) + \mu \|W\|_{2,1}, \quad (4)$$

For simplicity, we refer to the objective function in Eq. (4) as $g(F, W, b^T)$. First, we prove that the optimization problem in Eq. (4) is jointly convex with respect to F, W and b^T.

Theorem 1. *Denote* $S, M \in \mathbb{R}^{m\times m}, F \in \mathbb{R}^{m\times c}, W \in \mathbb{R}^{f\times c}, b \in \mathbb{R}^{c\times 1}$. $g(W, F, b^T) = Tr((X^T W + \mathbf{1}b^T - F)^T S (X^T W + \mathbf{1}b^T - F)) + \mu\|W\|_{2,1}$ *is jointly convex with respect to* W, F *and* b^T.

Proof. We can write $g(W, F, b^T)$ in matrix form as:

$$g(F, W, b^T) = Tr \begin{bmatrix} W \\ F \\ b^T \end{bmatrix}^T P \begin{bmatrix} W \\ F \\ b^T \end{bmatrix} + \mu \|W\|_{2,1},$$

where

$$P = \begin{bmatrix} XSX^T & -XS & XS\mathbf{1} \\ -SX^T & S & -S\mathbf{1} \\ \mathbf{1}^T SX^T & -\mathbf{1}^T S & \mathbf{1}^T S\mathbf{1} \end{bmatrix} + \mu \|W\|_{2,1}.$$

Thus in order to prove that $g(W, F, b^T)$ is jointly convex with respect to W, F, b^T, we only need to prove that $Tr \begin{bmatrix} W \\ F \\ b^T \end{bmatrix}^T P \begin{bmatrix} W \\ F \\ b^T \end{bmatrix}$ is positive semi-definite.

For arbitrary vector $z = \left[z_1^T, z_2^T, z_3\right]^T \in \mathbf{R}^{m+f+1}$, where $z_1 \in \mathbf{R}^{m\times 1}$, $z_2 \in \mathbf{R}^{f\times 1}$ and z_3 is a scalar, we have

$$\begin{aligned} & Tr \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}^T P \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = Tr \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}^T \begin{bmatrix} XSX^T & -XS & XS\mathbf{1} \\ -SX^T & S & -S\mathbf{1} \\ \mathbf{1}^T SX^T & -\mathbf{1}^T S & \mathbf{1}^T S\mathbf{1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\ & = Tr(X^T z_1 + z_2 \mathbf{1} - z_3)^T S (X^T z_1 + z_2 \mathbf{1} - z_3) \\ & \ge 0 \end{aligned}$$

So P is positive semi-definite. Thus $Tr(X^T W + \mathbf{1}b^T - F)^T S (X^T W + \mathbf{1}b^T - F)$ is a convex function. $\|W\|_{2,1}$ is convex, the sum of two convex functions is also convex. □

By setting the derivative of (4) *w.r.t.* b to 0, we have:

$$b = \frac{1}{m} F^T S\mathbf{1} - \frac{1}{m} W^T XS\mathbf{1}, \quad (5)$$

where $m = \mathbf{1}^T S\mathbf{1}$.

Substituting (5) into (4) we have

$$\begin{aligned} & \min_{W,F} Tr(((I - \frac{1}{m}\mathbf{1}\mathbf{1}^T S)X^T W - (I - \frac{1}{m}\mathbf{1}\mathbf{1}^T S)F)^T S \\ & ((I - \frac{1}{m}\mathbf{1}\mathbf{1}^T S)X^T W - (I - \frac{1}{m}\mathbf{1}\mathbf{1}^T F)F) + \mu\|W\|_{2,1}, \end{aligned} \quad (6)$$

where I is an identity matrix. By denoting $H = I - \frac{1}{m}\mathbf{1}\mathbf{1}^T S$ as a centering matrix, we can rewrite (6) as follows:

$$\min_{W,F} Tr((HX^T W - HF)^T S (HX^T W - HF)) + \mu\|W\|_{2,1} \quad (7)$$

By setting the derivative of (7) *w.r.t.* W to zero, we obtain:

$$XHSHX^T W + \mu DW = XHSHF, \quad (8)$$

where D is a diagonal matrix which is defined as

$$D = \begin{bmatrix} \frac{1}{2\|w^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|w^d\|_2} \end{bmatrix} \quad (9)$$

Since D is related to W, it is difficult to solve this problem. Hence, we propose an iterative method to solve it. We can obtain D with randomly initialized W. Then, we have:

$$W = (XHSHX^T + \mu D)^{-1} XHSHF. \quad (10)$$

After obtain W and b, we can compute $\widetilde{F} = X^T W + 1b^T$. In order to minimize the objective function, we adjust the labels of unlabeled training data as follows:

$$F_{ij} = \begin{cases} 0, & if\ \widetilde{F}_{ij} \le 0 \\ \widetilde{F}_{ij}, & if\ 0 \le \widetilde{F}_{ij} \le 1 \\ 1, & if \widetilde{F}_{ij} \ge 1 \end{cases} \quad (11)$$

Base on the above mathematical deduction, we propose an efficient iterative algorithm to optimize the objective function (3).

After we obtain the final solution for F, we select the unlabeled training data with high confidence and assign corresponding labels to them. By adding the selected unlabeled training data into the original labeled training data, we get new constructed labeled training data.

Algorithm 1: Optimization Algorithm for CSFS

Data: Training data $X_i|_{i=1}^n \in \mathbb{R}^{d\times n}$
Training data labels $Y_l|_{l=1}^{n_L} \in \mathbb{R}^{n\times c}$
Parameters μ
Result:
Feature Selection Matrix $W \in \mathbb{R}^{d\times c}$
Global Optimized Predicted Label $F_i|_{i=1}^n \in \mathbb{R}^{n\times c}$

Compute training data weighting matrix S ;
Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d\times c}$;
repeat
  Compute the diagonal matrix D_t according to (9) ;
  Compute W_{t+1} according to $W_{t+1} = (XHSHX^T + \mu D_t)^{-1}XHSHY$;
  Compute b_{t+1} according to $b_{t+1} = \frac{1}{m}F^T S\mathbf{1} - \frac{1}{m}W^T XS\mathbf{1}$;
  Compute $\widetilde{F}_{t+1}$ according to $\widetilde{F}_{t+1} = X^T W + \mathbf{1}b^T$;
  Adjust F according to Eq. (11) ;
until *Convergence*;
Return W^* and F^*.

Convergence analysis

The proposed iterative approach in Algorithm 1 can be verified to converge by the following theorem.

Theorem 2. *The iterative approach monotonically decreases the objective function value in each iteration until convergence.*

Proof. Suppose after the t-th iteration, we obtain W^t, b^t and F^t. In the next iteration, we fix F as F^t and solve for W^{t+1}. According to Algorithm 1, it can be inferred that

$$\begin{aligned} W^{t+1} = \arg\min & Tr((X^T W + \mathbf{1}b^T - F)^T S \\ & (X^T W + \mathbf{1}b - F)) + \mu Tr(W^T DW) \end{aligned} \tag{12}$$

The same as (Nie et al. 2010a), we obtain:

$$\begin{aligned} & Tr((X^T W^{t+1} + \mathbf{1}(b^{t+1})^T - F^t)^T S(X^T W^{t+1} + \mathbf{1}(b^{t+1})^T \\ & - F^t)) + \mu\|W^{t+1}\|_{2,1} \\ & \le Tr((X^T W^t + \mathbf{1}(b^t)^T - F^t)^T S(X^T W^t + \mathbf{1}(b^t)^T \\ & - F^t)) + \mu\|W^t\|_{2,1} \end{aligned} \tag{13}$$

In the same manner, when we fix W as W^t and b as b^t, we have:

$$\begin{aligned} & Tr((X^T W^t + \mathbf{1}(b^t)^T - F^{t+1})^T S(X^T W^t + \mathbf{1}(b^{t+1})^T \\ & - F^{t+1})) + \mu\|W^t\|_{2,1} \\ & \le Tr((X^T W^t + \mathbf{1}(b^{t+1})^T - F^t)^T S(X^T W^t + \mathbf{1}(b^t)^T \\ & - F^t)) + \mu\|W^t\|_{2,1} \end{aligned} \tag{14}$$

By integrating Eq. (13) and Eq. (14), we arrive at:

$$\begin{aligned} & Tr((X^T W^{t+1} + \mathbf{1}(b^{t+1})^T - F^{t+1})^T S(X^T W^{t+1} + \mathbf{1}(b^{t+1})^T \\ & - F^{t+1})) + \mu\|W^{t+1}\|_{2,1} \\ & \le Tr((X^T W^t + \mathbf{1}(b^t)^T - F^t)^T S(X^T W^t + \mathbf{1}(b^t)^T \\ & - F^t)) + \mu\|W^t\|_{2,1} \end{aligned} \tag{15}$$

Eq. (15) demonstrates that the objective function value decreases after each iteration. Thus, Theorem 2 has been proved.

□

Experiments

In this section, we conduct several experiments on large scale datasets to validate the performance of our algorithm. First we compare our algorithm with other feature selection algorithms, followed by studying the performance *w.r.t.* parameter sensitivity and the convergence of Algorithm 1.

Experiment Setup

To evaluate performance of the proposed algorithm, we apply this algorithm to three different applications. Five datasets are adopted in the experiment, including NUS WIDE, MSRA, MRMI. We compare its performance with the following algorithms:

1. All Features [All-Fea]: The original data with no feature selection has been used as a baseline in this experiment.
2. Fisher Score [F-score] (Richard, Hart, and Stork 2001): This is a classical feature selection algorithm. It conducts feature selection by evaluating the importance of features one by one.
3. Feature Selection via Joint $l_{2,1}$-Norms Minimization [FSNM] (Nie et al. 2010a): Joint $l_{2,1}$-norm minimization is used on both loss function and regularization term for feature selection.
4. Spectral Feature Selection [SPEC] (Zhao and Liu 2007): Spectral regression is employed to select features one by one.
5. Sub-Feature Uncovering with Sparsity [SFUS] (Ma et al. 2012b): This algorithm incorporates joint sparse feature selection with multi-label learning to uncover shared feature subspace.
6. Locality sensitive semi-supervised feature selection [LSDF] (Zhao, Lu, and He 2008): This is a semi-supervised feature selection approach based on within-class and between-class graph construction.
7. Noise insensitive trace ratio criterion for feature selection [TRCFS] (Liu et al. 2013): This is a recent semi-supervised feature selection algorithm based on noise insensitive trace ratio criterion.
8. Structural Feature Selection with Sparsity (Ma et al. 2012a) [SFSS]: This semi-supervised feature selection algorithms incorporates joint feature selection and semi-supervised learning into a single framework. Correlations between different features have been taken into consideration.

Table 1: SETTINGS OF THE TRAINING SETS

Dataset	Size(n)	Labeled Training Data (m)	Number of Selected Features
MIML	$1,000$	$1 \times c, 3 \times c, 5 \times c$	$\{200, 240, 280, 320, 360, 400\}$
NUS-WIDE	$10,000$	$1 \times c, 3 \times c, 5 \times c$	$\{240, 280, 320, 360, 400, 440, 480\}$
Mflickr	$10,000$	$1 \times c, 3 \times c, 5 \times c$	$\{200, 240, 280, 320, 360, 400\}$
YEAST	$1,500$	$1 \times c, 3 \times c, 5 \times c$	$\{50, 60, 70, 80, 90, 100\}$
SCENE	$1,000$	$1 \times c, 3 \times c, 5 \times c$	$\{170, 190, 210, 230, 250, 270, 290\}$

We tune all the parameters (if any) in the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$ for each algorithm and the best results are reported. In the experiments, we randomly generate a training set for each dataset consisting n samples, among which $m\%$ samples are labeled. Similarly to the pipeline in (Ma et al. 2012a), we randomly split the training and testing data 5 times and report average results. The libSVM (Chang and Lin 2011) with RBF kernel is applied in the experiment. The optimal parameters of the SVM are determined by grid search on a tenfold cross-validation. Mean Average Precision (MAP) is used to evaluate the performances.

Dataset Description

We utilize three datasets, *i.e.* MIML Mflickr and NUS-WIDE in the experiments. We give a brief description of the three datasets as follows.

MIML (Zhou and Zhang 2006): The MIML dataset consists of 2,000 natural scene images. Each image in the dataset is artificially marked with several labels. More than 22% of the dataset belong to more than one class. On average, 1.24 class labels are assigned to each image.

MIRFLICKR (Huiskes and Lew 2008): This image dataset has 25,000 images which are collected from Flickr.com. Each image in this dataset is associated with 8.94 tags. 33 annotated tags are chosen from the dataset as the ground truth.

NUS-WIDE (Chua et al. 2009): The NUS-WIDE image dataset consists of 269,000 real-world images collected from Flickr by Lab for Media Search in the National University of Singapore. We download all the images from the website, among which 59,263 images are unlabeled. By removing the unlabeled images, we use remaining 209,347 images, along with the ground-truth labels in the experiment.

YEAST (Elisseeff and Weston 2002): The yeast dataset contains micro-array expression data and phylogenetic profiles with 1500 genes in the training set and 917 in the testing set. Each gene is associated with a bunch of functional classes whose maximum size may be potentially more than 190.

SCENE (Boutell et al. 2004): This dataset consists of 2,000 natural scene images, where each image is manually associated with a set of labels. On average, about 1.24 class labels are assigned to each image.

Performance Evaluation

We present the experimental results measured by MAP in Tables 2-4 when different numbers of labeled training data are used respectively.

From the experimental results, we observe that (1) All the feature selection methods generally get better performance than All-Fea which does not conduct feature selection. This observation indicates that feature selection contributes to improvement of annotation performance. (2) The proposed algorithm consistently outperform the other supervised feature selection algorithms. Hence, we can conclude that utilizing both labeled and unlabeled training data can boost annotation performance. (3) Compared with other semi-supervised feature selection algorithms, our method still gets better performances. The advantage is especially visible when there are only few training data are labeled. Semi-supervised approaches are designed for the cases when only limited number of training data are labeled. Thus we can safely conclude that our method is better than LSDF, TRCFS and SFSS.

Convergence Study

In this section, we conduct experiments to demonstrate that the proposed iterative algorithm monotonically decrease the objective function value until convergence. MIML dataset is utilized in the experiment with $10 \times c$ labeled training data. We fix the parameter μ at 1 which is the median value of the tuned range of the parameters.

We show the convergence curve of the proposed algorithm *w.r.t.* the objective function value in Eq. (3) on the MIML dataset. From this curve, we can observe that the objective function value converge within very few iterations, which is very efficient.

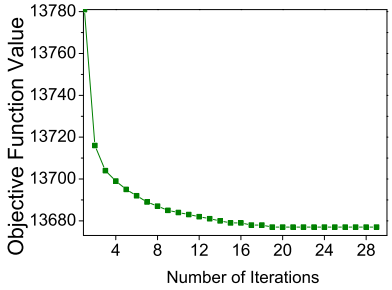

Figure 1: Convergence curve of the objective function value in (3) using Algorithm 1.

Table 2: Performance Comparison(±Standard Deviation(%)) when $1 \times c$ data are labeled.

Dataset	All-Fea	F-Score	SPEC	FSNM	SFUS	LSDF	TRCFS	SFSS	CSFS
MIML	23.9 ± 0.5	23.8 ± 0.3	24.0 ± 0.4	24.2 ± 0.3	24.3 ± 0.3	26.4 ± 0.2	26.9 ± 0.3	27.4 ± 0.3	$\mathbf{28.5 \pm 0.2}$
NUS-WIDE	4.6 ± 0.4	4.5 ± 0.3	4.5 ± 0.3	4.7 ± 0.1	4.9 ± 0.5	4.8 ± 0.3	5.0 ± 0.4	5.2 ± 0.3	$\mathbf{5.6 \pm 0.4}$
Mflickr	9.6 ± 0.5	9.4 ± 0.3	9.8 ± 0.4	9.9 ± 0.3	10.3 ± 0.2	10.6 ± 0.2	11.2 ± 0.4	11.6 ± 0.3	$\mathbf{11.9 \pm 0.1}$
YEAST	31.2 ± 0.3	32.5 ± 0.2	31.4 ± 0.3	31.2 ± 0.1	32.8 ± 0.2	31.6 ± 0.2	33.2 ± 0.3	33.9 ± 0.3	$\mathbf{35.1 \pm 0.2}$
SCENE	15.2 ± 0.4	16.8 ± 0.5	17.6 ± 0.3	15.4 ± 0.1	18.9 ± 0.2	19.3 ± 0.4	19.6 ± 0.3	21.2 ± 0.3	$\mathbf{23.5 \pm 0.4}$

Table 3: Performance Comparison(±Standard Deviation(%)) when $3 \times c$ data are labeled.

Dataset	All-Fea	F-Score	SPEC	FSNM	SFUS	LSDF	TRCFS	SFSS	CSFS
MIML	26.6 ± 0.3	27.0 ± 0.2	26.8 ± 0.2	26.9 ± 0.3	27.3 ± 0.2	27.1 ± 0.2	27.4 ± 0.4	27.8 ± 0.3	$\mathbf{29.1 \pm 0.4}$
NUS-WIDE	5.8 ± 0.4	5.6 ± 0.3	5.5 ± 0.4	5.9 ± 0.3	6.2 ± 0.4	6.1 ± 0.3	6.3 ± 0.4	6.5 ± 0.3	$\mathbf{6.8 \pm 0.3}$
Mflickr	10.8 ± 0.3	10.6 ± 0.5	10.7 ± 0.2	10.9 ± 0.1	11.4 ± 0.4	11.8 ± 0.3	12.0 ± 0.3	12.3 ± 0.3	$\mathbf{12.7 \pm 0.2}$
YEAST	32.2 ± 0.3	32.4 ± 0.4	32.9 ± 0.2	33.7 ± 0.3	34.2 ± 0.3	32.3 ± 0.2	32.9 ± 0.3	33.2 ± 0.4	$\mathbf{34.4 \pm 0.1}$
SCENE	47.2 ± 0.4	49.2 ± 0.5	49.3 ± 0.4	52.3 ± 0.5	53.4 ± 0.3	53.9 ± 0.4	54.4 ± 0.2	54.9 ± 0.3	$\mathbf{56.1 \pm 0.3}$

Table 4: Performance Comparison(±Standard Deviation(%)) when $5 \times c$ data are labeled.

Dataset	All-Fea	F-Score	SPEC	FSNM	SFUS	LSDF	TRCFS	SFSS	CSFS
MIML	28.2 ± 0.4	29.1 ± 0.2	28.3 ± 0.4	28.4 ± 0.5	28.7 ± 0.3	29.1 ± 0.2	29.4 ± 0.3	29.9 ± 0.3	$\mathbf{31.5 \pm 0.4}$
NUS-WIDE	6.5 ± 0.5	6.3 ± 0.2	6.4 ± 0.2	6.8 ± 0.5	6.9 ± 0.3	6.4 ± 0.4	7.1 ± 0.5	7.3 ± 0.3	$\mathbf{7.5 \pm 0.5}$
Mflickr	11.3 ± 0.5	10.9 ± 0.3	11.0 ± 0.4	11.2 ± 0.3	11.9 ± 0.4	12.2 ± 0.3	12.4 ± 0.2	12.7 ± 0.3	$\mathbf{13.4 \pm 0.2}$
YEAST	34.2 ± 0.6	34.6 ± 0.4	35.5 ± 0.5	35.6 ± 0.4	36.7 ± 0.3	34.4 ± 0.5	34.5 ± 0.3	36.2 ± 0.3	$\mathbf{37.3 \pm 0.4}$
SCENE	55.1 ± 0.5	55.4 ± 0.4	55.2 ± 0.3	55.3 ± 0.4	56.1 ± 0.5	55.8 ± 0.2	56.2 ± 0.3	56.4 ± 0.3	$\mathbf{56.9 \pm 0.5}$

Influence of Selected Features

In this section, an experiment is conducted to learn influence of selected features. Following the above experiment, we still use the same experimental setting.

Figure 2 shows MAP varies *w.r.t.* the number of selected features. We can observe that: 1) When the number of selected features is relatively small, MAP of classification is quite small. 2) When we increase the number of selected features to 280, MAP rises from 0.274 to 0.297. 3) When the first 280 features are selected, MAP arrives at the peak level. 4) When the number of selected features increase from 340 to full features, the classification performance keeps stable. Based on the above observations, we can conclude that feature selection benefits to the classification performance.

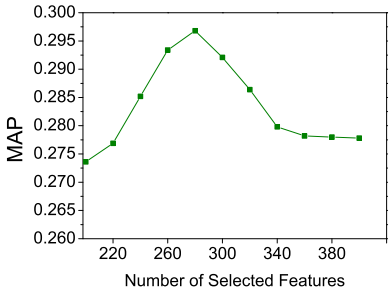

Figure 2: Influence of selected feature number

Conclusion

In this paper, a novel convex framework for semi-supervised multi-label feature selection for large-scale multi-media analysis. First, different from traditional graph based semi-supervised algorithms, the proposed algorithm does not require graph construction and eigen-decomposition. Therefore, the computational cost is comparably low and the algorithm can be readily applied to large-scale dataset. Second, we apply $l_{2,1}$-norm regularization to the objective function to make the classifier robust for outliers. Third, we propose an efficient approach with guaranteed convergence to solve the objective function. It is worthwhile mentioning that the proposed framework can be readily applied to semi-supervised single-label problem by adding another constraint. Extensive experiments demonstrate that the proposed algorithm consistently outperforms state-of-the-art related algorithms on all the used datasets.

References

Agrawal, R.; Gupta, A.; Prabhu, Y.; and Varma, M. 2013. Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages. In *Proc. WWW*, 13–24.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern Recogn.* 37(9):1757–1771.

Cai, X.; Nie, F.; Huang, H.; and Ding, C. H. Q. 2011. Multi-class l2, 1-norm support vector machine. In *ICDM*, 91–100.

Cai, X.; Nie, F.; and Huang, H. 2013. Exact top-k feature selection via l2,0-norm constraint. *23rd International Joint Conference on Artificial Intelligence (IJCAI)* 1240–1246.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: A library for

support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: A real-world web image database from national university of singapore. In *Proc. CIVR*.

Elisseeff, A., and Weston, J. 2002. A kernel method for multi-labelled classification. In *Proc. NIPS*.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proc. MIR*, 39–43.

Kenji, K., and Rendell, L. A. 1992. The feature selection problem: Traditional methods and a new algorithm. In *Proc. ICML*, 129–134.

Kong, X., and Yu, P. S. 2010a. Multi-label feature selection for graph classification. In *Proc. ICDM*, 274–283.

Kong, X., and Yu, P. S. 2010b. Semi-supervised feature selection for graph classification. In *Proc. SIGKDD*, 793–802.

Liu, Y.; Nie, F.; Wu, J.; and Chen, L. 2013. Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. *Neurocomputing* 105:12–18.

Luo, Y.; Tao, D.; Xu, C.; Li, D.; and Xu, C. 2013. Vector-valued multi-view semi-supervised learning for multi-label image classification. In *Proc. AAAI*.

Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J. R. R.; Sebe, N.; and Hauptmann, A. G. 2012a. Discriminating joint feature analysis for multimedia data understanding. *IEEE Trans. Multimedia* 14(6):1662–1672.

Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J. R. R.; and Sebe, N. 2012b. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Trans. Multimedia* 14(4):1021–1030.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010a. Efficient and robust feature selection via joint l21-norms minimization. In *Proc. NIPS*, 759–768.

Nie, F.; Xu, D.; Tsang, I. W.-H.; and Zhang, C. 2010b. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Trans. Image Process.* 19(7):1921–1932.

Richard, D.; Hart, P. E.; and Stork, D. G. 2001. *Pattern Classification*. New York: Wiley-Interscience.

Tan, M.; Wang, L.; and Tsang, I. W. 2010. Learning sparse svm for feature selection on very high dimensional datasets. In *Proc. ICML*, 1047–1054.

Wang, H.; Nie, F.; Huang, H.; Risacher, S. L.; Ding, C.; Saykin, A. J.; Shen, L.; and ADNI. 2011. A new sparse multi-task regression and feature selection method to identify brain imaging predictors for memory performance. *ICCV 2011: IEEE Conference on Computer Vision* 557–562.

Wu, F.; Yuan, Y.; and Zhuang, Y. 2010. Heterogeneous feature selection by group lasso with logistic regression. In *ACM Multimedia*, 983–986.

Xu, Zenglin, I. K. M.-T. L., and Jin, R. 2010. Discriminative semi-supervised feature selection via manifold regularization. *IEEE Trans. Neural Networks* 21(7):1033–1047.

Zhao, Z., and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *Proc. ICML*, 1151–1157.

Zhao, J.; Lu, K.; and He, X. 2008. Locality sensitive semi-supervised feature selection. *Neurocomputing* 71(10):1842–1849.

Zhou, Z.-H., and Zhang, M.-L. 2006. Multi-instance multi-label learning with application to scene classification. In *Proc. NIPS*, 1609–1616.