Semi-supervised Feature Analysis for Multimedia Annotation by Mining Label Correlation

Xiaojun Chang[1], Haoquan Shen[2], Sen Wang[1], Jiajun Liu[3], and Xue Li[1]

[1] The University of Queensland, QLD 4072, Australia
[2] Zhejiang University, Zhejiang, China
[3] CSIRO Brisbane

Abstract. In multimedia annotation, labeling a large amount of training data by human is both time-consuming and tedious. Therefore, to automate this process, a number of methods that leverage unlabeled training data have been proposed. Normally, a given multimedia sample is associated with multiple labels, which may have inherent correlations in real world. Classical multimedia annotation algorithms address this problem by decomposing the multi-label learning into multiple independent single-label problems, which ignores the correlations between different labels. In this paper, we combine label correlation mining and semi-supervised feature selection into a single framework. We evaluate performance of the proposed algorithm of multimedia annotation using MIML, MIRFLICKR and NUS-WIDE datasets. Mean average precision (MAP), MicroAUC and MacroAUC are used as evaluation metrics. Experimental results on the multimedia annotation task demonstrate that our method outperforms the state-of-the-art algorithms for its capability of mining label correlations and exploiting both labeled and unlabeled training data.

Keywords: Semi-supervised Learning, Multi-label Feature Selection, Multimedia Annotation.

1 Introduction

With the booming of social networks, such as Facebook and Flickr, we have witnessed a dramatical growth of multimedia data, *i.e.* image, text and video. Consequently, there are increasing demands to effectively organize and access these resources. Normally, feature vectors, which are used to represent aforementioned resources, are usually very large. However, it has been pointed out in [1] that only a subset of features carry the most discriminating information. Hence, selecting the most representative features plays an essential role in a multi-media annotation framework. Previous works [2,3,4,5] have indicated that feature selection is able to remove redundant and irrelevant information in the feature representation, thus improves subsequent analysis tasks.

Existing feature selection algorithms are designed in various ways. For example, conventional feature selection algorithms, such as Fisher Score [6], compute

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 74–85, 2014.

weights of all features, rank them accordingly and select the most discriminating features one by one. While dealing with multi-label problems, the conventional algorithms generally transform the problem into a couple of binary classification problems for each concept respectively. Hence, feature correlations and label correlations are ignored [7], which will deteriorate the subsequent annotation performance.

Another limitation is that they only use labeled training data for feature selection. Considering that there are a large number of unlabeled training data available, it is beneficial to leverage unlabeled training data for multimedia annotation. Over recent years, semi-supervised learning has been widely studied as an effective tool for saving labeling cost by using both labeled and unlabeled training data [8,9,10]. Inspired by this motivation, feature learning algorithms based on semi-supervised framework, have been also proposed to overcome the insufficiency of labeled training samples. For example, Zhao *et al.* propose an algorithm based on spectral analysis in [5]. However, similarly to Fisher Score [6], their method selects the most discriminating features one by one. Besides, correlations between labels are ignored.

Our semi-supervised feature selection algorithm integrates multi-label feature selection and semi-supervised learning into a single framework. Both labeled and unlabeled data are utilized to select features while label correlations and feature correlations are simultaneously mined.

The main contributions of this work can be summarized as follows:

1. We combine joint feature selection with sparsity and semi-supervised learning into a single framework, which can select the most discriminating features with an insufficient amount of labeled training data.
2. The correlations between different labels are taken into account to facilitate the feature selection.
3. Since the objective function is non-smooth and difficult to solve, we propose a fast iteration algorithm to obtain the global optima. Experimental results on convergence validates that the proposed algorithm converges within very few iterations.

The rest of this paper is organized as follows. In Section 2, we introduce details of the proposed algorithm. Experimental results are reported in Section 3. Finally, we conclude this paper in Section 4.

2 Proposed Framework

To mine correlations between different labels for feature selection, our algorithm is built upon a reasonable assumption that different class labels have some inherent common structures. In this section, our framework is described in details, followed by an iterative algorithm with guaranteed convergence to optimize the objective function.

2.1 Formulation of Proposed Framework

Let us define $X = \{x_1, x_2, \cdots, x_n\}$ as the training data matrix, where $x_i \in \mathbb{R}^d (1 \leq i \leq n)$ is the i-th data point and n is the total number of training data. $Y = [y_1, y_2, \cdots, y_m, y_{m+1}, \cdots, y_n]^T \in \{0,1\}^{n \times c}$ denotes the label matrix and c is the class number. $y_i \in \mathbb{R}^c$ $(1 \leq i \leq n)$ is the label vector with c classes. Y_{ij} indicates the j-th element of y_i and $Y_{ij} := 1$ if x_i is in the j-th class, and $Y_{ij} := 0$ otherwise. If x_i is not labeled, y_i is set to a vector with all zeros. Inspired by [11], we assume that there is a low-dimensional subspace shared by different labels. We aim to learn c prediction functions $\{f_t\}_{t=1}^c$. The prediction function f_t can be generalized as follows:

$$f_t(x) = v_t^T x + p_t^T Q^T x = w_t^T x, \tag{1}$$

where $w_t = v_t + Qp_t$. v and p are the weights, Q is a transformation matrix which projects features in the original space into a shared low-dimensional subspace.

Suppose there are m_t training data $\{x_i\}_{i=1}^{m_t}$ belonging to the t-th class labeled as $\{y_i\}_{i=1}^{m_t}$. A typical way to obtain the prediction function f_t is to minimize the following objective function:

$$\arg \min_{f_t, Q^T Q = I} \sum_{t=1}^{c} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} loss(f_t(x_i), y_i) + \beta \Omega(f_t)\right) \tag{2}$$

Note that to make the problem tractable we impose the constraint $Q^T Q = I$. Following the methodology in [2], we incorporate (1) into (2) and obtain the objective function as follows:

$$\min_{\{v_t, p_t\}, Q^T Q = I} \sum_{t=1}^{c} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} loss((v_t + Qp_t)^T x_i, y_i) + \beta \Omega(\{v_t, p_t\})\right) \tag{3}$$

By defining $W = V + QP$, where $V = [v_1, v_2, \cdots, v_c] \in \mathbb{R}^{d \times c}$ and $P = [p_1, p_2, \cdots, p_c] \in \mathbb{R}^{sd \times c}$ where sd is the dimension of shared lower dimensional subspace, we can rewrite the objective function as follows:

$$\min_{W, V, P, Q^T Q = I} loss(W^T X, Y) + \beta \Omega(V, P) \tag{4}$$

Note that we can implement the shared feature subspace uncovering in different ways by adopting different loss functions and regularizations. Least square loss is the most widely used in research for its stable performance and simplicity. By applying the least square loss function, the objective function arrives at:

$$\arg \min_{W, P, Q^T Q = I} \|X^T W - Y\|_F^2 + \alpha \|W\|_F^2 + \beta \|W - QP\|_F^2 \tag{5}$$

As indicated in [12], however, there are two issues worthy of further consideration. First, the least square loss function is very sensitive to outliers. Second, it is beneficial to utilize sparse feature selection models on the regularization

term for effective feature selection. Following [12,13,14], we employ $l_{2,1}$-norm to handle the two issues. We can rewrite the objective function as follows:

$$\arg \min_{W,P,Q^TQ=I} \|X^TW - Y\|_F^2 + \alpha\|W\|_{2,1} + \beta\|W - QP\|_F^2 \tag{6}$$

Meanwhile, we define a graph Laplacian as follows: First, we define an affinity matrix $A \in \mathbb{R}^{n\times n}$ whose element A_{ij} measures the similarity between x_i and x_j as:

$$A_{ij} = \begin{cases} 1, & x_i \textit{ and } x_j \textit{ are } k \textit{ nearest neighbours}; \\ 0, & \textit{otherwise}. \end{cases} \tag{7}$$

Euclidean distance is utilized to measure whether two data points x_i and x_j are k nearest neighbours in the original feature space. Then, the graph Laplacian, L, is computed according to $L = S - A$, where S is a diagonal matrix with $S_{ii} = \sum_{j=1}^{n} A_{ij}$.

Note that multimedia data have been normally shown to process a manifold structure, we adopt manifold regularization to explore it. By applying manifold regularization to the aforementioned loss function, our objective function arrives at:

$$\begin{aligned} \arg \min_{W,P,Q^TQ=I} & Tr(W^TXLX^TW) + \gamma[\alpha\|W\|_{2,1} \\ & +\beta\|W - QP\|_F^2 + \|X^TW - F\|_F^2] \end{aligned} \tag{8}$$

We define a selecting diagonal matrix U whose diagonal element $U_{ii} = \infty$, if x_i is a labeled data, and $U_{ii} = 0$ otherwise. To exploit both labeled and unlabeled training data, a label prediction matrix $F = [f_1, \cdots, f_n]^T \in \mathbb{R}^{n\times c}$ is introduced for all the training data. The label prediction of $x_i \in X$ is $f_i \in \mathbb{R}^c$. Following [5], we assume that F holds smoothness on both the ground truth of training data and on the manifold structure. Therefore, F can be obtained as follows:

$$\arg\min_F tr(F^TLF) + tr((F - Y)^TU(F - Y)). \tag{9}$$

By incorporating (9) into (6), the objective function finally becomes:

$$\begin{aligned} \arg \min_{F,W,P,Q^TQ=I} & tr(F^TLF) + tr(F - Y)^TU(F - Y) \\ & +\gamma[\alpha\|W\|_{2,1} + \beta\|W - QP\|_F^2 + \|X^TW - F\|_F^2] \end{aligned} \tag{10}$$

As indicated in [13], W is guaranteed to be sparse to perform feature selection across all data points by $\|W\|_{2,1}$ in our regularization term.

2.2 Optimization

The proposed function involves the $l_{2,1}$-norm, which is difficult to solve in a closed form. We propose to solve this problem in the following steps. By setting the derivative of (10) *w.r.t.* P equal to zero, we have

$$P = Q^TW \tag{11}$$

By denoting $W = [w^1, \ldots, w^d]^T$, the objective function becomes

$$\begin{aligned} \arg \min_{F,W,P,Q^TQ=I} & tr(F^TLF) + tr(F-Y)^TU(F-Y) \\ +\gamma[\alpha tr(W^TDW) & + \beta\|W - QP\|_F^2 + \|X^TW - F\|_F^2], \end{aligned} \tag{12}$$

where D is a matrix with its diagonal elements $D_{ii} = \frac{1}{2\|w^i\|_2}$.

Note that for any given matrix A, we have $\|A\|_F^2 = tr(A^TA)$. Substituting P in (10) by (11), we can rewrite the objective function as follows:

$$\begin{aligned} \arg \min_{F,W,Q^TQ=I} & tr(F^TLF) + tr((F-Y)^TU(F-Y)) \\ +\gamma[\alpha tr(W^TDW) & + \beta tr((W - QQ^TW)^T(W - QQ^TW)) \\ & + \|X^TW - F\|_F^2], \end{aligned} \tag{13}$$

According to the equation $(I - QQ^T)(I - QQ^T) = (I - QQ^T)$, we have:

$$\begin{aligned} \arg \min_{F,W,Q^TQ=I} & tr(F^TLF) + tr((F-Y)^TU(F-Y)) \\ +\gamma(\|X^TW - F\|_F^2 & + tr(W^T(\alpha D + \beta I - \beta QQ^T)W)) \end{aligned} \tag{14}$$

By the setting the derivative *w.r.t.* W to zero, we have:

$$W = (M - \beta QQ^T)^{-1}XF \tag{15}$$

where $M = XX^T + \alpha D + \beta I$.

Then the objective function becomes:

$$\begin{aligned} \arg \min_{F,Q^TQ=I} & tr(F^TLF) + tr((F-Y)^TU(F-Y)) \\ & +\gamma[tr(F^TF) - tr(F^TX^T(M - \beta QQ^T)^{-1}XF] \end{aligned} \tag{16}$$

By setting the derivative *w.r.t.* F to zero, we have:

$$LF + U(F - Y) + \gamma F - \gamma X^T(M - \beta QQ^T)^{-1}XF = 0 \tag{17}$$

Thus, we have

$$F = (B - \gamma X^TR^{-1}X)^{-1}UY, \tag{18}$$

where

$$B = L + U + \gamma I \tag{19}$$

$$R = M - \beta QQ^T. \tag{20}$$

Then, the objective function can be written as

$$\max_{Q^TQ=I} tr[Y^TU(B - \mu X^TR^{-1}X)^{-1}UY]. \tag{21}$$

According to the Sherman-Woodbury-Morrison matrix identity,

$$(B - \gamma X^TR^{-1}X)^{-1} = B^{-1} + \gamma B^{-1}X^T(R - \gamma XB^{-1}X^T)^{-1}XB^{-1}. \tag{22}$$

Thus, the objective function arrives at

$$\max_{Q^T Q=I} tr[Y^T U B^{-1} X^T J^{-1} X B^{-1} U Y], \tag{23}$$

where

$$J = (R - \mu X B^{-1} X^T) = (M - \beta Q Q^T - \gamma X B^{-1} X^T). \tag{24}$$

Theorem 1. *The global optimization Q^* can be obtained by solving the following ratio trace maximization problem:*

$$max_{Q^T Q=I} tr[(Q^T C Q)^{-1} Q^T D Q], \tag{25}$$

where

$$C = I - \beta(X X^T + \alpha D + \beta I - \gamma X B^{-1} X^T)^{-1} \tag{26}$$

$$D = N^{-1} X B^{-1} U Y Y^T U B^{-1} X^T N^{-1}. \tag{27}$$

Proof. See Appendix.

To obtain Q, we need to conduct eigen-decomposition of $C^{-1}D$, which is $O(d^3)$ in complexity. However, as the solution of Q requires the input of D which is obtained according to W, it is still not straightforward to obtain Q and W. So as shown in Algorithm 1, we propose an iterative approach to solve this problem.

The proposed iterative approach in Algorithm 1 can be verified to converge to the optimal W by the following theorem. Following the work in [12], we can prove the convergence of Algorithm 1.

3 Experiments

In this section, experiments are conducted on three datasets, *i.e.* MIML [16], MIRFLICKR [17] and NUS-WIDE [18] to validate performance of the proposed algorithm.

3.1 Compared Methods

To evaluate performances of the proposed method, we compare it with the following algorithms:

1. All features [All-Fea]: We directly use the original data for annotation without feature selection as a baseline.
2. Fisher Score [F-score] [6]: This is a classical method, which selects the most discriminative features by evaluating the importance of features one by one.
3. Feature Selection via Joint $l_{2,1}$-Norms Minimization [FSNM] [3]: This algorithm utilizes joint $l_{2,1}$-norm minimization on both loss function and regularization for joint feature selection.
4. Spectral Feature Selection [SPEC] [15]: It employs spectral regression to select features one by one.

Algorithm 1. The algorithm for solving the objective function

Data: The training data $X \in \mathbb{R}^{d \times n}$
The training data labels $Y \in \mathbb{R}^{n \times c}$
Parameters γ, α and β
Result:
Optimized $W \in \mathbb{R}^{d \times c}$

Compute the graph Laplacian matrix $L \in \mathbb{R}^{n \times n}$;
Compute the selection matrix $U \in \mathbb{R}^{n \times n}$;
Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d \times c}$ randomly;
repeat
  Compute the diagonal matrix D_t as:
  $D_t = \begin{bmatrix} \frac{1}{2\|w_t^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|w_t^d\|_2} \end{bmatrix}$
  Compute C according to $C = I - \beta(XX^T + \alpha D + \beta I - \gamma XB^{-1}X^T)^{-1}$.
  Compute D according to $D = N^{-1}XB^{-1}UYY^TUB^{-1}X^TN^{-1}$.
  Compute the optimal Q^* according to Theorem 1.
  Compute W according to $W = (M - \beta QQ^T)^{-1}XF$.
until *Convergence*;
Return W^*.

5. Sub-Feature Uncovering with Sparsity [SFUS] [12]: It incorporates the latest advances in a joint, sparse feature selection with multi-label learning to uncover a feature subspace which is shared among different classes.
6. Semi-supervised Feature Selection via Spectral Analysis [sSelect] [5]: It is semi-supervised feature selection approach based on spectral analysis.

3.2 Dataset Description

Three datasets, *i.e.*, MIML [16] Mflickr [17] and NUS-WIDE [18] are used in the experiments. A brief description of the three datasets is given as follows.

MIML: This image dataset consists of 2,000 natural scene images. Each image in this dataset is artificially marked with a set of labels. Over 22% of the dataset belong to more than one class. On average, each image has 1.24 class labels.

MIRFLICKR: The MIRFLICKR image dataset consists of 25 000 images collected from Flickr.com. Each image is associated with 8.94 tags. We choose 33 annotated tags in the dataset as the ground truth.

NUS-WIDE: The NUS-WIDE image dataset has 269,000 real-world images which are collected from Flickr by Lab for Media Search in the National University of Singapore. All the images have been downloaded from the website, among which 59,563 images are unlabeled. By removing unlabeled images, we use the remaining 209,347 images, along with ground-truth labels in the experiments.

Table 1. Settings of the Training Sets

Dataset	Size(n)	Labeled Training Data (m)	Number of Selected Features
MIML	$1,000$	$5 \times c$, $10 \times c$, $15 \times c$	$\{200, 240, 280, 320, 360, 400\}$
NUS-WIDE	$10,000$	$5 \times c$, $10 \times c$, $15 \times c$	$\{240, 280, 320, 360, 400, 440, 480\}$
Mflickr	$10,000$	$5 \times c$, $10 \times c$, $15 \times c$	$\{200, 240, 280, 320, 360, 400\}$

3.3 Experimental Setup

In the experiment, we randomly generate a training set for each dataset consisting of n samples, among which m samples are labeled. The detailed settings are shown in Table 1. The remaining data are used as testing data. Similar to the pipeline in [2], we randomly split the training and testing data 5 times and report average results. The libSVM [19] with RBF kernel is applied in the experiment. The optimal parameters of the SVM are determined by grid search on a tenfold crossvalidation. Following [2], the graph Laplacian, k is set as 15. Except for the SVM parameters, the regularization parameters, γ, α and β, in the objective funciton (10), are tuned in the range of $\{10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}\}$. The number of selected features can be found in Table 1.

3.4 Performance Evaluation

Tables 2, 3 and 4 present experimental results measured by MAP, MicroAUC and MacroAUC when using different numbers of labeled training data ($5 \times c$, $10 \times c$ and $15 \times c$) respectively.

Taking MAP as an example, it is observed that: 1) The proposed method is better than All-Fea which does not apply feature selection. Specifically, the proposed algorithm outperforms All-Fea by about 5.5% using $10 \times c$ labeled training data in the MIML dataset, which indicates that feature selection can contribute to annotation performance. 2) Our method has consistently better performances than the other supervised feature selection algorithms. When using $5 \times c$ labeled training data in the MIML dataset, the proposed algorithm is better than the second best supervised feature selection algorithm by 3.8%. 3) The proposed algorithm gets better performances than the compared semi-supervised feature selection algorithm, which demonstrates that mining label correlations is beneficial to multimedia annotation.

Table 2. Performance Comparison($\pm$Standard Deviation(%)) when $5 \times c$ data are labeled

Dataset	Criteria	All-Fea	F-Score	SPEC	FSNM	SFUS	sSelect	Ours
MIML	MAP	26.1 ± 0.1	26.9 ± 0.2	26.1 ± 0.2	26.1 ± 0.3	26.2 ± 0.2	28.9 ± 0.3	$\mathbf{31.4 \pm 0.1}$
	MicroAUC	54.6 ± 0.1	54.4 ± 0.2	54.6 ± 0.2	54.6 ± 0.2	54.7 ± 0.1	55.1 ± 0.2	$\mathbf{55.8 \pm 0.2}$
	MacroAUC	52.4 ± 0.3	52.6 ± 0.4	52.4 ± 0.2	52.4 ± 0.2	52.6 ± 0.3	53.1 ± 0.4	$\mathbf{54.4 \pm 0.2}$
NUS	MAP	5.8 ± 0.2	5.4 ± 0.1	5.9 ± 0.2	5.8 ± 0.3	6.0 ± 0.2	6.4 ± 0.3	$\mathbf{7.1 \pm 0.2}$
	MicroAUC	86.4 ± 0.4	86.1 ± 0.1	86.5 ± 0.3	87.2 ± 0.2	87.4 ± 0.4	87.9 ± 0.3	$\mathbf{89.1 \pm 0.2}$
	MacroAUC	64.0 ± 0.4	63.7 ± 0.2	64.2 ± 0.4	64.4 ± 0.3	64.9 ± 0.2	65.5 ± 0.3	$\mathbf{66.3 \pm 0.2}$
Mflickr	MAP	12.2 ± 0.2	12.2 ± 0.3	12.3 ± 0.2	12.3 ± 0.2	12.4 ± 0.3	13.6 ± 0.2	$\mathbf{15.8 \pm 0.1}$
	MicroAUC	75.2 ± 0.2	75.1 ± 0.3	75.4 ± 0.3	75.3 ± 0.4	75.5 ± 0.2	76.1 ± 0.3	$\mathbf{77.3 \pm 0.1}$
	MacroAUC	50.3 ± 0.3	50.3 ± 0.4	50.4 ± 0.3	50.5 ± 0.2	50.7 ± 0.4	51.3 ± 0.3	$\mathbf{52.6 \pm 0.2}$

Table 3. Performance Comparison(±Standard Deviation(%)) when $10 \times c$ data are labeled

Dataset	Criteria	All-Fea	F-Score	SPEC	FSNM	SFUS	sSelect	Ours
MIML	MAP	31.6 ± 0.3	33.0 ± 0.2	31.6 ± 0.2	31.6 ± 0.3	33.0 ± 0.1	35.2 ± 0.2	**37.1 ± 0.1**
	MicroAUC	59.3 ± 0.4	58.9 ± 0.3	59.3 ± 0.2	59.4 ± 0.3	59.8 ± 0.2	60.4 ± 0.2	**61.7 ± 0.2**
	MacroAUC	62.0 ± 0.3	61.0 ± 0.2	62.0 ± 0.2	62.0 ± 0.1	62.0 ± 0.2	62.6 ± 0.2	**63.7 ± 0.2**
NUS	MAP	6.6 ± 0.2	6.0 ± 0.1	6.5 ± 0.2	6.4 ± 0.3	7.0 ± 0.2	6.9 ± 0.3	**8.0 ± 0.2**
	MicroAUC	87.3 ± 0.3	87.2 ± 0.2	87.4 ± 0.5	87.3 ± 0.2	87.6 ± 0.3	88.2 ± 0.3	**89.5 ± 0.3**
	MacroAUC	67.5 ± 0.4	67.4 ± 0.3	67.7 ± 0.4	67.6 ± 0.2	67.9 ± 0.3	68.2 ± 0.4	**69.4 ± 0.3**
Mflickr	MAP	12.8 ± 0.3	12.6 ± 0.2	12.3 ± 0.2	12.4 ± 0.3	12.9 ± 0.2	14.2 ± 0.3	**16.1 ± 0.2**
	MicroAUC	78.1 ± 0.2	78.2 ± 0.3	78.1 ± 0.2	78.4 ± 0.3	78.4 ± 0.2	78.8 ± 0.3	**80.0 ± 0.1**
	MacroAUC	55.1 ± 0.3	55.3 ± 0.2	55.2 ± 0.4	55.4 ± 0.3	55.6 ± 0.2	56.4 ± 0.4	**57.3 ± 0.2**

Table 4. Performance Comparison(±Standard Deviation(%)) when $15 \times c$ data are labeled

Dataset	Criteria	All-Fea	F-Score	SPEC	FSNM	SFUS	sSelect	Ours
MIML	MAP	33.0 ± 0.2	34.7 ± 0.1	33.0 ± 0.2	33.5 ± 0.3	34.1 ± 0.1	35.8 ± 0.2	**37.9 ± 0.1**
	MicroAUC	63.4 ± 0.4	63.3 ± 0.3	63.5 ± 0.1	63.4 ± 0.3	63.7 ± 0.2	64.2 ± 0.3	**65.1 ± 0.2**
	MacroAUC	62.3 ± 0.3	62.5 ± 0.2	62.3 ± 0.2	62.3 ± 0.4	62.5 ± 0.2	63.1 ± 0.3	**64.2 ± 0.1**
NUS	MAP	6.9 ± 0.1	6.5 ± 0.3	6.8 ± 0.2	6.9 ± 0.2	7.3 ± 0.3	7.4 ± 0.3	**8.5 ± 0.2**
	MicroAUC	89.4 ± 0.2	89.1 ± 0.3	89.5 ± 0.2	89.8 ± 0.4	90.1 ± 0.3	90.7 ± 0.4	**91.9 ± 0.4**
	MacroAUC	69.2 ± 0.3	69.1 ± 0.2	69.3 ± 0.1	69.5 ± 0.3	69.7 ± 0.5	70.2 ± 0.3	**71.5 ± 0.5**
Mflickr	MAP	13.0 ± 0.2	12.9 ± 0.1	12.9 ± 0.1	12.8 ± 0.2	13.1 ± 0.3	14.8 ± 0.2	**16.7 ± 0.4**
	MicroAUC	79.2 ± 0.3	79.1 ± 0.3	79.2 ± 0.2	79.2 ± 0.4	79.5 ± 0.2	80.2 ± 0.4	**81.6 ± 0.2**
	MacroAUC	58.7 ± 0.4	58.5 ± 0.3	58.8 ± 0.2	59.1 ± 0.3	58.6 ± 0.3	59.9 ± 0.2	**60.4 ± 0.3**

3.5 Convergence Study

In this section, an experiment is conducted to validate that our proposed iterative algorithm monotonically decreases the objective function until convergence. $10 \times c$ labeled training data in MIML dataset are tested in this experiment. γ, α and β are fixed at 1 which is the median value of the tuned range of the parameters.

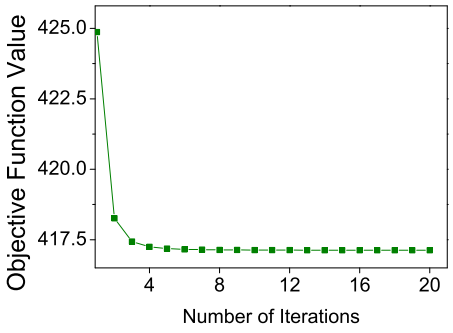

Fig. 1. Convergence

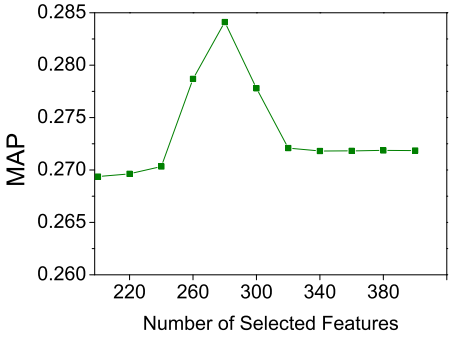

Fig. 2. Influence of selected feature number

Figure 1 shows the convergence curve of the proposed algorithm *w.r.t.* the objective function value in (10) on the MIML dataset. It is observed that the objective function values converge within 4 iterations.

3.6 Influence of Selected Features

In this section, an experiment is conducted to study how the number of selected features affect the performance of the proposed algorithm. Following the above experiment, we still use the same 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 annotation is quite small. 2) When the number of selected features rises to 280, MAP increases from 0.269 to 0.284. 3) When we select 280 features, MAP arrives at the peak level. 4) MAP keeps stable when we increase the number of selected features from 320 to full features. From this figure, feature selection benefits to the annotation performance.

3.7 Parameter Sensitivity Study

Another experiment is conducted to test the sensitivity of parameters in (10). Among different parameter combinations, the proposed algorithm gains the best performance when $\gamma = 10^1$, $\alpha = 10^4$ and $\beta = 10^2$. We show the MAP variations *w.r.t.* γ, α and β. From Figure 3, we notice that the performance of the proposed algorithm changes corresponding to different parameters. In summary, better results are obtained when α, β and γ are in the range of $[10^{-2}, \cdots, 10^2]$.

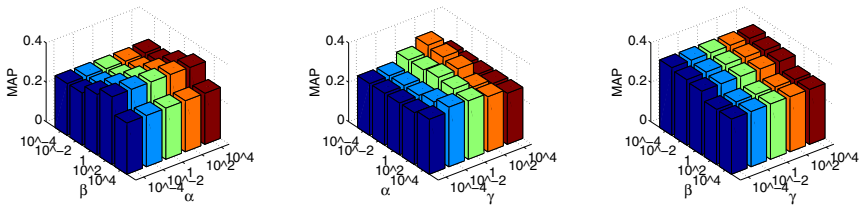

Fig. 3. The MAP variations of different parameter settings using the MIML dataset

4 Conclusion

In this paper, we have proposed a novel framework for semi-supervised feature analysis by mining label correlation. First, our method simultaneously discovers correlations between labels in a shared low-dimensional subspace to improve the annotation performance. Second, to make the classifier robust for outliers, $l_{2,1}$-norm is applied to the objective function. Third, this framework is extended into a semi-supervised scenario which exploits both labeled and unlabeled data. We evaluate the performance of a multimedia annotation task on three different datasets. Experimental results have demonstrated that the proposed algorithm consistently outperforms the other compared algorithms on all the three datasets.

Appendix

In this appendix, we prove Theorem 1.

To prove Theorem 1, we first give the following lemma and prove it.

Lemma 1. *With the same notations in the paper, we have the following equation:*

$$(R-\gamma XB^{-1}X^T)^{-1} = N^{-1} + \beta N^{-1}Q(Q^T(I-\beta N^{-1})Q)^{-1}Q^T N^{-1}, \quad (28)$$

where

$$N = M - \gamma XB^{-1}X^T. \quad (29)$$

Proof.

$$\begin{aligned}
&(R-\gamma XB^{-1}X^T)^{-1}\\
=&(M-\beta QQ^T-\gamma XB^{-1}X^T)^{-1}\\
=&N^{-1}+\beta N^{-1}Q(I-\beta Q^T N^{-1}Q)^{-1}Q^T N^{-1}\\
=&N^{-1}+\beta N^{-1}Q(Q^T(I-\beta N^{-1})Q)^{-1}Q^T N^{-1}
\end{aligned}$$

Proof of Theorem 1

Proof. From Eq. (29), we can tell that N is independent from Q. By employing Lemma 1, the objective function arrives at:

$$\max_{Q^TQ=I} tr[Y^TUB^{-1}X^TN^{-1}Q(Q^TKQ)^{-1}Q^TN^{-1}XB^{-1}UY], \quad (30)$$

where $K = I - \beta N^{-1}$. At the same time, we have:

$$N^{-1} = (M-\gamma XB^{-1}X^T)^{-1} = (XX^T+(\alpha+\beta)I-\gamma X(L+U+\gamma I)^{-1}X^T)^{-1}.$$

Thus, $K = I - \beta N^{-1} = C$. According to the property of trace operation that $tr(UV) = tr(VU)$ for any arbitrary matrices U and V, the objective function can be rewritten as:

$$\max_{Q^TQ=I} tr[Q^TN^{-1}XB^{-1}UYY^TUB^{-1}X^TN^{-1}Q(Q^TKQ)^{-1}].$$

The objective function is equivalent to:

$$\max_{Q^TQ=I} tr[(Q^TCQ)^{-1}Q^TDQ].$$

Acknowledgement. This work was partially supported by the Australian Research Council the Discovery Project DP No. 130104614 and DP No. 140100104. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Australian Research Council.

References

1. Yang, S.H., Hu, B.G.: Feature selection by nonparametric bayes error minimization. In: Proc. PAKDD, pp. 417–428 (2008)
2. Ma, Z., Nie, F., Yang, Y., Uijlings, J.R.R., Sebe, N., Hauptmann, A.G.: Discriminating joint feature analysis for multimedia data understanding. IEEE Trans. Multimedia 14(6), 1662–1672 (2012)
3. Nie, F., Henghuang, C.X., Ding, C.: Efficient and robust feature selection via joint l21-norms minimization. In: Proc. NIPS, pp. 759–768 (2007)
4. Wang, D., Yang, L., Fu, Z., Xia, J.: Prediction of thermophilic protein with pseudo amino acid composition: An approach from combined feature selection and reduction. Protein and Peptide Letters 18(7), 684–689 (2011)
5. Zhou, Z.H., Zhang, M.L.: Semi-supervised feature selection via spectral analysis. In: Proc. SIAM Int. Conf. Data Mining (2007)
6. Richard, D., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience, New York (2001)
7. Yang, Y., Wu, F., Nie, F., Shen, H.T., Zhuang, Y., Hauptmann, A.G.: Web and personal image annotation by mining label correlation with relaxed visual graph embedding. IEEE Trans. Image Process. 21(3), 1339–1351 (2012)
8. Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. In: IEEE Trans. PAMI, pp. 1553–1566 (2004)
9. Zhao, X., Li, X., Pang, C., Wang, S.: Human action recognition based on semi-supervised discriminant analysis with global constraint. Neurocomputing 105, 45–50 (2013)
10. Wang, S., Ma, Z., Yang, Y., Li, X., Pang, C., Hauptmann, A.: Semi-supervised multiple feature analysis for action recognition. IEEE Trans. Multimedia (2013)
11. Ji, S., Tang, L., Yu, S., Ye, J.: A shared-subspace learning framework for multi-label classification. ACM Trans. Knowle. Disco. Data 2(1), 8(1)–8(29) (2010)
12. Ma, Z., Nie, F., Yang, Y., Uijlings, J.R.R., Sebe, N.: Web image annotation via subspace-sparsity collaborated feature selection. IEEE Trans. Multimedia 14(4), 1021–1030 (2012)
13. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint l21-norms minimization. In: Proc. NIPS, pp. 1813–1821 (2010)
14. Yang, Y., Shen, H., Ma, Z., Huang, Z., Zhou, X.: L21-norm regularization discriminative feature selection for unsupervised learning. In: Proc. IJCAI (July 2011)
15. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proc. ICML, pp. 1151–1157 (2007)
16. Zhou, Z.H., Zhang, M.L.: Multi-instance multi-label learning with application to scene classification. In: Proc. NIPS, pp. 1609–1616 (2006)
17. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proc. MIR, pp. 39–43 (2008)
18. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: A real-world web image database from national university of singapore. In: Proc. CIVR (2009)
19. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`