Semi-supervised Multi-label Classification
A Simultaneous Large-Margin, Subspace Learning Approach

Yuhong Guo and Dale Schuurmans

[1] Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA
[2] Department of Computing Science
University of Alberta
Edmonton, AB, T6G 2E8, Canada

Abstract. Labeled data is often sparse in common learning scenarios, either because it is too time consuming or too expensive to obtain, while unlabeled data is almost always plentiful. This asymmetry is exacerbated in *multi-label* learning, where the labeling process is more complex than in the single label case. Although it is important to consider *semi-supervised* methods for multi-label learning, as it is in other learning scenarios, surprisingly, few proposals have been investigated for this particular problem. In this paper, we present a new semi-supervised multi-label learning method that combines large-margin multi-label classification with unsupervised subspace learning. We propose an algorithm that learns a subspace representation of the labeled and unlabeled inputs, while simultaneously training a supervised large-margin multi-label classifier on the labeled portion. Although joint training of these two interacting components might appear intractable, we exploit recent developments in induced matrix norm optimization to show that these two problems can be solved jointly, globally and efficiently. In particular, we develop an efficient training procedure based on subgradient search and a simple coordinate descent strategy. An experimental evaluation demonstrates that semi-supervised subspace learning can improve the performance of corresponding supervised multi-label learning methods.

Keywords: semi-supervised multi-label learning, subspace learning.

1 Introduction

In many real world data analysis problems, complex data objects such as documents, webpages, images and videos can be simultaneously assigned into multiple categories, and hence have multiple class labels. *Multi-label* classification is an important supervised learning problem that has received significant attention in the machine learning research literature. Although the earliest work on this problem simply reduced multi-label classification to a set of independent binary classification problems [1], it was quickly realized that such an approach

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 355–370, 2012.

was unsatisfactory [2] since such labels are usually not independent: most often they exhibit strong correlations. Capturing label correlations in an effective, yet tractable manner has led to a diverse set of formulations for multi-label learning, including pairwise label dependency methods [3, 4], large-margin methods [5–8], ranking based large margin methods [9–12], structure exploitation methods [13–18] and others. Of these, it has recently been observed that *large-margin* based approaches offer effective and efficient multi-label learning methods. In particular, it has been shown that a large-margin formulation based on minimizing a "calibrated separation ranking loss" demonstrates state-of-the-art performance in multi-label text categorization [8].

However, just as in any supervised learning scenario, a key bottleneck is obtaining sufficient labeled data to achieve reasonable generalization performance. In practice, one often encounters a significant amount of unlabeled data, even while labeled examples remain scarce, since labeling is an expensive and time-consuming process. *This issue is even more salient in multi-label learning*, since manually assigning multiple labels, correctly, is more challenging than assigning atomic labels. Thus, we address the challenge of exploiting significant unlabeled data to reduce the amount of labeled training data required for effective multi-label classification. Although *supervised* multi-label learning has received significant attention, *semi-supervised* multi-label learning is far from being well explored. A handful of preliminary studies have explored semi-supervised multi-label learning, using approaches such as non-negative matrix factorization [19], graph-based methods [20], and dimensionality reduction [21]. Unfortunately, these proposals rely on local optimization schemes for training, and do not offer reliable off-the-shelf procedures that protect end-users from local minima.

In this work, we propose a new approach for exploiting unlabeled data to help multi-label learning in a transductive setting, by simultaneously learning the underlying subspace feature representations of the data with a large margin multi-label classification model. Automatically discovering useful feature representations of data has been a long standing research of machine learning—from early unsupervised approaches, such as principal component analysis (PCA)—to recent supervised convex feature learning, such as multi-task feature learning [22]. Here we exploit recent results for semi-supervised convex subspace learning, which we adapt to large-margin multi-label classification. Our approach is based on two key recent ideas: (1) using calibrated separation ranking loss for large margin multi-label classification [8], and (2) using induced matrix norms to efficiently combine subspace learning with semi-supervised training [23]. By introducing a structured regularizer on the learned representation, and exploiting a particular induced matrix norm, we formulate the semi-supervised multi-label learning problem as a convex max-min optimization problem with no local maxima or minima. We then develop a specialized subgradient coordinate descent algorithm to solve the training problem efficiently, recovering a global solution.

The goal is to discover a subspace feature representation that captures discriminative structure that is not only shared across labeled and unlabeled data, but also shared across the multiple labels. Our experimental results demonstrate

that the proposed method can surpass the performance of some state-of-the-art supervised results for multi-label text categorization.

The remainder of the paper is organized as follows. After first introducing basic background concepts and notation, we review previous work on large margin multi-label classification in Section 2, with a particular emphasis on the calibrated separation ranking loss used for state-of-the-art multi-label text categorization [8]. Based on this particular multi-label classification approach, we then present a semi-supervised formulation in Section 3 that exploits implicit subspace learning through structured matrix norm regularization. An efficient global optimization algorithm is then presented in Section 4. Finally, we present an experimental evaluation in Section 5 and conclude the paper with a discussion of future research directions in Section 6.

1.1 Preliminaries: Definitions and Notation

Throughout this paper we will use capital letters to denote matrices, bold non-capital letters to denote column vectors, and regular non-capital letters to denote scalars, unless special declaration is given.

We use I_d to denote a $d \times d$ identity matrix; and use $\mathbf{1}$ to denote a column vector with all 1 entries, generally assuming its length can be inferred from context. Given a vector $\mathbf{x}$, $\|\mathbf{x}\|_2$ denotes its Euclidean norm.

Given a matrix X, $\|X\|_F^2$ denotes its Frobenius norm; the block norm $\|X\|_{p,1}$ is defined as $\|X\|_{p,1} = (\sum_i (\sum_j |X_{ij}|^p)^{\frac{1}{p}})$; and the trace norm is defined as $\|X\|_{tr} = \sum_i \sigma_i(X)$, where $\sigma_i(X)$ denotes the ith singular value of X. We use $X_{i:}$ to denote the ith row of a matrix X, use $X_{:j}$ to denote the jth column of X, and use X_{ij} to denote the entry at the ith row and jth column of X. We also need to make use of a general form of induced matrix norm given by the definition $\|X\|_{(\mathcal{Z},p)} := \max_{\mathbf{z}\in\mathcal{Z}} \|X\mathbf{z}\|_p$. It can be shown [23] that this defines a valid matrix norm for any bounded closed set $\mathcal{Z} \subset \mathbb{R}^n$ such that $\text{span}(\mathcal{Z}) = \mathbb{R}^n$ and any $1 \leq p \leq \infty$. Finally, for matrices, we use $\|X\|$ to refer to a generic norm on X, and $\|Y\|^*$ to denote its conjugate norm. The conjugate satisfies $\|Y\|^* = \max_{\|X\|\leq 1} \text{tr}(X^\top Y)$ and $\|X\|^{**} = \|X\|$, where tr denotes trace.

2 Background

Our main formulation is based on combining two key components: an effective large-margin formulation for multi-label learning, and an efficient approach for automated representation learning that avoids local optima.

2.1 Large Margin Multi-label Classification

Multi-label classification is a widely studied problem in supervised machine learning, for which large margin methods provide one of the state-of-the-art approaches. By maximizing discriminative classification margins, expressed by

different loss functions, supervised large margin learning methods are both efficient, and demonstrate good generalization performance.

In particular, [8] has recently proposed an effective method for supervised multi-label learning that exploits the dependence structure between labels in a simple yet effective manner. The basic idea is to simultaneously train a set of L predictors, one for each class label, but under a coordinated loss: the calibrated separation ranking loss captures the sum of two hinge losses, one of which is between the prediction value of the least positive labeled class and the prediction value of a threshold dummy class, and the other of which is between the prediction value of the least negative labeled class and the prediction value of the threshold dummy class.

More formally, in the supervised multi-label learning setting, one is given an input data matrix $X \in \mathbb{R}^{t \times d}$ and label indicator matrix $Y \in \{0,1\}^{t \times L}$, where L denotes the number of classes. We also assume a feature mapping function $\phi(\cdot)$ is fixed. Then, given an input instance $\mathbf{x}$, the L dimensional response vector $\mathbf{s}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top W$ is recovered using W, giving a "score" for each label. These scores will be compared to a threshold to determine which labels are to be predicted. Then the calibrated separation ranking loss is given by

$$\max_{l \in Y_{i:}} (1 + s_0(X_{i:}) - s_l(X_{i:}))_+ + \max_{\bar{l} \in \bar{Y}_{i:}} (1 + s_{\bar{l}}(X_{i:}) - s_0(X_{i:}))_+. \tag{1}$$

So, for example, given a test example $\mathbf{x}$, its classification is determined by $y_l^* = \arg\max_{y_l \in \{0,1\}} y_l(s_l(\mathbf{x}) - s_0(\mathbf{x}))$.

It is shown in [8] that minimizing this loss under standard squared regularization can be formulated as a standard convex quadratic minimization problem

$$\begin{aligned} \min_{W, \mathbf{u}, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad & \frac{\alpha}{2}(\|W\|_F^2 + \|\mathbf{u}\|_2^2) + \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} \\ \text{subject to} \quad & \boldsymbol{\xi}_i \geq 1 + X_{i:}(\mathbf{u} - W_{:l}) \quad \text{for } l \in Y_{i:}, \forall i = 1 \cdots t \\ & \boldsymbol{\eta}_i \geq 1 - X_{i:}(\mathbf{u} - W_{:\bar{l}}) \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0 \end{aligned} \tag{2}$$

where $l \in Y_{i:}$ lists through the indices of all entries of $Y_{i:}$ that contain 1 values, and $\bar{Y}$ denotes the complementary of Y, i.e., $\bar{Y} = 1 - Y$. Obviously the loss function only captures the classification relevant separation ranking that separate positive labels from negative labels for each instance, instead of the pairwise rankings among all label pairs in [9]. By conducting calibrated separation ranking, the label separation on new testing instances can be automatically determined using the trained predictors.

In [8] this approach is shown to demonstrate superior generalization and efficiency in supervised multi-label text categorization, so we make use of this loss in our semi-supervised formulation.

2.2 Unsupervised Representation Learning

To allow unlabeled data to influence the training of a multi-label classifier we consider the approach of learning a new input data representation that makes

the label correlations more apparent. We begin by adopting a recent approach to representation learning that offers a tractable way to learn a latent representation and a data reconstruction model.

Initially, consider the case where one is just given unlabeled data X. A simple goal for representation learning is to learn an $m \times d$ dictionary B of m basis vectors, and a $t \times m$ representation matrix Ψ containing new feature vectors of length m, so that X can be accurately reconstructed from $\hat{X} = \Psi B$. To measure approximation error we consider the loss function $\frac{1}{2}\|\hat{X} - X\|_F^2$. Note that the factorization $\hat{X} = \Psi B$ is invariant to reciprocal rescalings of B and Φ, so to avoid degeneracy their individual magnitudes have to be controlled. We will assume that each row $B_{i:}$ of B is constrained to belong to a bounded closed convex set $\mathcal{B} = \{\mathbf{b} : \|\mathbf{b}\|_2 \leq 1\}$. The generic training problem can be expressed

$$\min_{B \in \mathcal{B}^m} \min_{\Psi} \frac{1}{2}\|\Psi B - X\|_F^2 + \gamma \|\Psi^\top\|_{p,1} \tag{3}$$

where $\gamma \geq 0$ is a trade-off parameter. Some standard approaches to representation learning can be recovered by particular choices of p in (3). For example, a standard form of *sparse coding* can be recovered by choosing $p = 1$ [24]. Instead, choosing $p = 2$ results in a regularizer that encourages entire columns $\Psi_{:j}$ (features) to become sparse [25] while otherwise only smoothing the rows, hence implicitly reducing the dimensionality of the learned representation Ψ.

Unfortunately, the straightforward formulation (3) is not jointly convex in B and Ψ, and even recent formulations resort to local minimization strategies. However, a key observation is that the training problem can be solved globally if the number of learned features m is indirectly controlled through the use of the $\|\Psi^\top\|_{p,1}$ regularizer. As noted, for $p > 1$, such a regularizer will already naturally encourage entire columns $\Psi_{:j}$ (features) to become sparse [25]. A key result that leads to a tractable reformulation is the following identity from [23, 26].

Proposition 1. *[23, Theorem 1]:*

$$\min_{B \in \mathcal{B}^\infty} \min_{\Psi} \frac{1}{2}\|\Psi B - X\|_F^2 + \gamma \|\Psi^\top\|_{p,1} \;=\; \min_{\hat{X}} \frac{1}{2}\|\hat{X} - X\|_F^2 + \gamma \|\hat{X}\|^*_{(\mathcal{B},p^*)} \tag{4}$$

where $\min_{B \in \mathcal{B}^\infty}$ *denotes* $\min_{m \in \mathbb{N}} \min_{B \in \mathcal{B}^m}$*, and* $\|\cdot\|^*_{(\mathcal{B},p^*)}$ *is the conjugate of an induced matrix norm.*

The latter problem is convex, and for $p = 1$ or $p = 2$ can be readily solved for $\hat{X}$, after which the optimal factors B and Ψ can be readily recovered [23, 26].

3 Simultaneous Multi-label Classification and Representation Learning

Our main contribution in this paper is to combine these two components to formulate a multi-label classification and representation learning framework that uses unlabeled data to guide the learning of a multi-label classifier. Such an

approach enables semi-supervised learning, where unlabeled data assists in the otherwise supervised training of a classification model. We will investigate semi-supervised learning in a transductive setting, where a set of labeled and unlabeled data is provided, and one seeks an accurate labeling over the unlabeled portion.[1] The main advantage of the proposed formulation is that it admits an efficient global training scheme that combines multi-label classification with representation learning.

Note that global training schemes offer a significant advantage to the end-user, since they do not need to concern themselves with the inner workings of any particular solver. Rather, it is sufficient to focus on understanding the nature of the problem formulation, and the solver can be used as a black box. This separation of implementation from specification frees the end-user to focus on engineering useful features, or imposing trade-offs between training errors and regularization penalties, without having to understand the inner workings of a solver. In this paper, however, we need to show that a suitably efficient solver can exist, which we do in the next section.

To develop a combined formulation of multi-label classification and representation learning, consider the following set-up. Let $X \in \mathbb{R}^{t \times d}$ be the input feature matrix and let $X^\ell \in \mathbb{R}^{t_\ell \times d}$ be the labeled submatrix formed by the first t_ℓ rows of X, where $t_\ell + t_u = t$. Let $Y \in \{0,1\}^{t_\ell \times L}$ be the label matrix over the supervised portion.

We would like to learn a $(t_l + t_u) \times m$ representation matrix $\Psi = [\Psi_l; \Psi_u]$, an $m \times d$ basis dictionary B, and an $m \times L$ prediction model W, such that $X = [X_l; X_u]$ can be reconstructed from $\hat{X} = \Psi B$, and Y can be reconstructed from $\hat{Y} = \Psi_l W$. To accommodate the offset in the calibrated separation ranking loss (1) we consider a linear prediction function over the subspace representation $\Psi(W_{:l} - \mathbf{u})$. Then by combining (2) and (3) we reach the joint training formulation

$$\min_{\Psi, B \in \mathcal{B}^m} \min_{W, \mathbf{u}, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad \frac{\alpha}{2}(\|W\|_F^2 + \|\mathbf{u}\|_2^2) + \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}\|X - \Psi B\|_F^2 + \gamma \|\Psi^\top\|_{p,1} \quad (5)$$

$$\begin{aligned} \text{subject to} \quad & \boldsymbol{\xi}_i \geq 1 + \Psi_{i:}(\mathbf{u} - W_{:l}) \quad \text{for } l \in Y_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\eta}_i \geq 1 - \Psi_{i:}(\mathbf{u} - W_{:\bar{l}}) \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0 \end{aligned}$$

where now the multi-label predictions are made from the learned representation Ψ, which is the only component that connects the multi-label training problem to the representation learning problem. Note that for $p > 1$ the regularizer $\|\Psi^\top\|_{p,1}$ will tend to reduce the dimensionality of the learned representation Ψ, which gives an automated form of subspace learning directly coupled to the multi-label training problem. Unfortunately, (5) does not immediately offer a plausible global training algorithm that avoids local minima: although it is straightforward

[1] The approach we propose here is in principle extendible to a semi-supervised learning scenario where the test data is not available during training. However, such out-of-sample classification entails significant additional technicality that is currently left to future work.

to observe that (5) is convex in each of the components B, W, Ψ, $\mathbf{u}$, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ given the others, it is not jointly convex. Fortunately Proposition 1 can be generalized to accommodate the more general formulation given here, as we now show.

3.1 Equivalent Reformulation as a Convex Problem

Unlike staged training procedures that separate the unsupervised from the supervised phase [27], and previous work on semi-supervised dimensionality reduction that relies on alternating minimization [28], here we demonstrate a jointly convex formulation that allows all components to be trained simultaneously.

Let $M = \Psi B$ and $Z = \Psi(W - [\mathbf{u}, \cdots, \mathbf{u}])$ denote the reconstruction and response matrices respectively. To simplify the development below, we set $\mathbf{u} = 0$ and therefore let $Z = \Psi W$ and $M = \Psi B$; hence $Z \in \mathbb{R}^{t \times L}$ and $M \in \mathbb{R}^{t \times d}$. Substituting this into (5) yields

$$\begin{aligned} \min_{\Psi, B \in \mathcal{B}^m, W, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad & \min_{Z = \Psi W, M = \Psi B} \quad \frac{\alpha}{2}\|W\|_F^2 + \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}\|X - M\|_F^2 + \gamma \|\Psi^\top\|_{p,1} \\ & \text{subject to} \quad \boldsymbol{\xi}_i \geq 1 - Z_{il} \quad \text{for } l \in Y_{i:}, \forall i = 1 \cdots t_\ell \\ & \qquad\qquad \boldsymbol{\eta}_i \geq 1 + Z_{i\bar{l}} \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t_\ell \\ & \qquad\qquad \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0. \end{aligned} \tag{6}$$

To achieve compatibility with the reformulation exploited by Proposition 1 we replace the regularization penalty $\|W\|_F^2$ with a constraint on the norms of rows $W_{i:}$ in W. In particular, we constrain each $W_{i:}$ to the bounded closed set $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq \alpha\}$. This leads to a slightly modified formulation

$$\begin{aligned} \min_{\Psi, B \in \mathcal{B}^m, W \in \mathcal{W}^m, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad & \min_{Z = \Psi W, M = \Psi B} \quad \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}\|X - M\|_F^2 + \gamma \|\Psi^\top\|_{p,1} \\ & \text{subject to} \quad \boldsymbol{\xi}_i \geq 1 - Z_{il} \quad \text{for } l \in Y_{i:}, \forall i = 1 \cdots t_\ell \\ & \qquad\qquad \boldsymbol{\eta}_i \geq 1 + Z_{i\bar{l}} \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t_\ell \\ & \qquad\qquad \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0. \end{aligned} \tag{7}$$

Finally, by relaxing the feature number m and instead allowing the rank reducing regularizer $\|\Psi^\top\|_{p,1}$ to automatically choose the dimension, [23, Proposition 3] shows that the problem (7) is equivalent to

$$\begin{aligned} \min_{Z, M, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad & \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}\|X - M\|_F^2 + \gamma \|[M, Z]\|^*_{(\mathcal{U}, p^*)} \\ \text{subject to} \quad & \boldsymbol{\xi}_i \geq 1 - Z_{il} \quad \text{for } l \in Y_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\eta}_i \geq 1 + Z_{i\bar{l}} \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0 \end{aligned} \tag{8}$$

where as in Proposition 1, $\|\cdot\|_{(\mathcal{U}, p^*)}$ is a conjugate of an induced matrix norm, but now with respect to the closed bounded set $\mathcal{U} := \{[\mathbf{b}; \mathbf{w}] : \|\mathbf{b}\|_2 \leq 1 \text{ and } \|\mathbf{w}\|_2 \leq \alpha\}$.

Importantly, the problem (8) is jointly convex in Z, M, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, since norms are always convex. For $p = 1$ or $p = 2$, given an optimal pair $[M, Z]$, the underlying factors Ψ, B and W can be efficiently recovered using a procedure outlined in [23]. However, for the purposes of transduction, Z itself is sufficient for determining the label predictions on the unlabeled data, so this extra recovery procedure can be bypassed.

3.2 Tractable Special Case

Even though the problem (8) is convex, it is not guaranteed to be tractable, since not every induced matrix norm is tractable to compute [29]. However, the important special cases of $p = 1$ and $p = 2$ both allow the induced norm $\|\cdot\|_{(\mathcal{U},p^*)}$ to be efficiently evaluated. In particular, for $p = 2$, [23] establishes the following useful characterization.

Proposition 2. *[23, Lemma 5]:* $\|[M, Z]\|^*_{(\mathcal{U},2)} = \max_{\rho \geq 0} \|D_\rho^{-1}[M, Z]^\top\|_{tr}$ *where*

$$D_\rho = \begin{bmatrix} \sqrt{1+\alpha^2\rho}\, I_d & 0 \\ 0 & \sqrt{\alpha^2 + \frac{1}{\rho}}\, I_L \end{bmatrix}. \tag{9}$$

This proposition shows that for the case $p = 2$ the conjugate induced norm can be efficiently computed: all that is required is a line search over a scalar variable $\rho \geq 0$, where for each value of ρ the inner calculation can be efficiently evaluated by computing the singular value decomposition of $[M, Z]D_\rho^{-1}$.

Below we find it more convenient to work with a re-parameterized version of the calculation.

Proposition 3. $\max_{\rho \geq 0} \|D_\rho^{-1}[M, Z]^\top\|_{tr} = \max_{0 \leq \theta \leq 1} \|E_\theta[M, Z]^\top\|_{tr}$ *where*

$$E_\theta = \begin{bmatrix} \sqrt{\theta}\, I_d & 0 \\ 0 & \frac{\sqrt{1-\theta}}{\alpha}\, I_L \end{bmatrix}. \tag{10}$$

This proposition is easy to establish by noting the relationships $D_\rho^{-1} = E_{\frac{1}{\alpha^2\rho+1}}$, $E_\theta^{-1} = D_{\frac{1-\theta}{\alpha^2\theta}}$, $\theta = \frac{1}{\alpha^2\rho+1}$, and $\rho = \frac{1-\theta}{\alpha^2\theta}$, hence optimizing with D_ρ^{-1} over the range $\rho \geq 0$ is equivalent to optimizing with E_θ over the range $0 \leq \theta \leq 1$.

Thus, we obtain the following convex optimization problem that is equivalent to (8) for the special case when $p = 2$:

$$\begin{aligned} \max_{0 \leq \theta \leq 1} \min_{Z, M, \boldsymbol{\xi}, \boldsymbol{\eta}} \quad & \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\beta}{2}\|X - M\|_F^2 + \gamma \|E_\theta[M, Z]^\top\|_{tr} \\ \text{subject to} \quad & \boldsymbol{\xi}_i \geq 1 - Z_{il} \text{ for } l \in Y_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\eta}_i \geq 1 - Z_{il} \text{ for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0. \end{aligned} \tag{11}$$

To verify that this problem has no local optima, first note that the inner problem is convex in Z, M, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ for each fixed θ. Furthermore, it can be shown that

$\|E_\theta[M,Z]^\top\|_{tr}$ is concave in θ. Since a pointwise minimum of concave functions is concave, the outer optimization in θ also has no local maxima. To solve (11), all one has to do is run a simple outer concave maximization over a scalar variable, while each inner minimization is a standard convex minimization. Essentially, the inner problem has the same complexity as the standard multi-label learning problem, which only has to be repeated a few times (say, around 10) to achieve an accurate solution.

4 Optimization Algorithm

The semi-supervised optimization problem we formulated above in (11) is a convex optimization problem but with a non-smooth trace norm. To develop an efficient optimization algorithm for it, we first derive an equivalent reformulation following a well-known variational formulation of the trace norm [22, 30]:

Proposition 4. *Let $Q \in \mathbb{R}^{t\times d}$. The trace norm of Q is equal to*

$$\|Q\|_{tr} = \frac{1}{2}\inf_{S\succeq 0} tr(Q^\top S^{-1}Q) + tr(S), \tag{12}$$

and the infimum is achieved for $S = (QQ^\top)^{1/2}$.

Following this proposition, we can reformulate (11) as the following

$$\begin{aligned}\max_{0\le\theta\le 1}\min_{Z,M,\boldsymbol{\xi},\boldsymbol{\eta}}\inf_{S\succeq 0}\ & \mathbf{1}^\top\boldsymbol{\xi} + \mathbf{1}^\top\boldsymbol{\eta} + \frac{\beta}{2}\|X-M\|_F^2 \\ & +\frac{\gamma}{2}\mathrm{tr}([M,Z]E_\theta S^{-1}E_\theta[M,Z]^\top) + \frac{\gamma}{2}\mathrm{tr}(S) \\ \text{subject to}\ & \boldsymbol{\xi}_i \ge 1 - Z_{il} \ \text{ for } l \in Y_{i:}, \forall i = 1\cdots t_\ell \\ & \boldsymbol{\eta}_i \ge 1 + Z_{i\bar{l}} \ \text{ for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1\cdots t_\ell \\ & \boldsymbol{\xi} \ge 0, \boldsymbol{\eta} \ge 0\end{aligned} \tag{13}$$

which maintains the convexity of the original formulation of (11). Although the reformulated problem remains a non-smooth max-min optimization problem, an efficient optimization procedure can still be developed. In particular, we develop a simple subgradient-based binary line search procedure, combined with a block-descent inner minimization, to solve this problem.

First, consider the max-min optimization problem in (13) as a non-smooth concave optimization problem over θ

$$\max_{0\le\theta\le 1} \quad f(\theta) \tag{14}$$

where the objective function is a non-smooth function defined by a convex minimization problem

$$\begin{aligned}f(\theta) = \min_{M,\,\{Z,\boldsymbol{\xi},\boldsymbol{\eta}\}\in\mathcal{C}}\ \inf_{S\succeq 0}\ & \mathbf{1}^\top\boldsymbol{\xi} + \mathbf{1}^\top\boldsymbol{\eta} + \frac{\beta}{2}\|X-M\|_F^2 \\ & +\frac{\gamma}{2}\mathrm{tr}([M,Z]E_\theta S^{-1}E_\theta[M,Z]^\top) + \frac{\gamma}{2}\mathrm{tr}(S).\end{aligned} \tag{15}$$

Here we use $\mathcal{C}$ to denote the feasible region defined by the linear constraints in (13) over variables $\{Z, \boldsymbol{\xi}, \boldsymbol{\eta}\}$. For such a non-smooth optimization problem, subgradient-based methods such as proximal bundle methods can be generally applied. Nevertheless, we develop a much simpler *binary line search* procedure to tackle this specific one dimensional optimization problem over θ.

4.1 Binary Line Search

The idea of binary line search is to iteratively reduce the searching region (interval) for the optimal θ value, eliminating at least half of the feasible region each time. At the beginning of the binary line search, θ is upper-bounded by $V_u = 1$ and lower-bounded by $V_\ell = 0$, which is its full feasible region, i.e., the feasible line segment. In each iteration of the binary line search, we set θ as the midpoint of its upper bound and lower bound values, $\theta = (V_u + V_\ell)/2$. We then compute the subgradient of $f(\theta)$ at this current point θ. Following Danskin's theorem, the subgradient of $f(\theta)$ can be computed as

$$\frac{\partial f}{\partial \theta} = \frac{\gamma}{2} \frac{\partial \mathrm{tr}([M^*, Z^*] E_\theta S^{*-1} E_\theta [M^*, Z^*]^\top)}{\partial \theta} \tag{16}$$

where M^*, Z^*, S^{*-1} are the optimal solution for the convex minimization problem in (15) with the given θ value. Since $f(\theta)$ is concave in θ, a positive subgradient value at θ indicates that the optimal θ^* value is larger than the current θ value, while a negative subgradient value indicates that the optimal θ^* value is smaller than the current θ value. Therefore, we increase the lower bound of θ to its current value when $\frac{\partial f}{\partial \theta} > 0$, and reduce the upper bound of θ to its current value when $\frac{\partial f}{\partial \theta} < 0$; thus ensuring the search interval is halved at each iteration.

By repeating the binary line search step, the feasible subinterval containing the optimal θ^* value can be quickly reduced at an exponential rate. When the subgradient is close to 0 or the interval between upper and lower bound values is sufficiently small, an optimal θ value can be returned. The overall binary search procedure is described in Algorithm 1.

4.2 Block-Coordinate Descent for Inner Convex Minimization

Both the computation of each subgradient value of $f(\theta)$ in the binary line search procedure and the final optimal solution recovery require solving the convex minimization problem in (15); i.e., the inner minimization problem in (13), for optimal M^*, Z^* and S^{*-1} given a fixed θ value. Although this optimization problem is convex, it is nevertheless challenging to design an efficient and scalable optimization algorithm to tackle the typically large parameter matrices M, Z, and S. For example, even for a given S, the Hessian matrix for the quadratic programming problem in M and Z can be too large to fit in memory for even a medium-sized data set with large number of input features.

Therefore, we develop a scalable block-descent optimization algorithm to solve the convex minimization problem iteratively. Specifically, in each iteration, we

conduct an optimization over each of the three sets of variables, $\{S\}$, $\{M\}$ and $\{Z, \boldsymbol{\xi}, \boldsymbol{\eta}\}$, in turn, given all other variables fixed. The optimization over each of the three sub-problems is conducted as follows.

Optimization over Z. Given fixed S and M values, the minimization problem over the remaining variables $Z, \boldsymbol{\xi}, \boldsymbol{\eta}$ forms a standard quadratic program with linear constraints; i.e.,

$$\begin{aligned} \min_{Z,\boldsymbol{\xi},\boldsymbol{\eta}} \quad & \mathbf{1}^\top \boldsymbol{\xi} + \mathbf{1}^\top \boldsymbol{\eta} + \frac{\gamma}{2}\mathrm{tr}([M,Z]Q[M,Z]^\top) \\ \text{subject to} \quad & \boldsymbol{\xi}_i \geq 1 - Z_{il} \quad \text{for } l \in Y_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\eta}_i \geq 1 + Z_{i\bar{l}} \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \forall i = 1 \cdots t_\ell \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{\eta} \geq 0 \end{aligned} \tag{17}$$

where $Q = E_\theta S^{-1} E_\theta$. Note that the linear constraints are only expressed in terms of the labeled part of Z and $\boldsymbol{\xi}$, $\boldsymbol{\eta}$. It is easy to see that

$$\begin{aligned} \mathrm{tr}([M,Z]Q[M,Z]^\top) &= \mathrm{tr}([M,Z]\begin{bmatrix} Q_{dd} & Q_{dL} \\ Q_{Ld} & Q_{LL} \end{bmatrix}[M,Z]^\top) \\ &= \mathrm{tr}(MQ_{dd}M^\top + 2ZQ_{Ld}M^\top + ZQ_{LL}Z^\top) \end{aligned} \tag{18}$$

where Q_{dd} denotes the $d \times d$ top-left submatrix of Q, Q_{LL} denotes the $L \times L$ bottom-right submatrix of Q, and $Q_{dL} = Q_{Ld}^\top$ denotes the other two submatrices of Q. Moreover, it is known that the matrix Z and matrix M can both be decomposed into two submatrices corresponding to the labeled and unlabeled data, $Z = [Z^l; Z^u]$ and $M = [M^l; M^u]$. Thus (18) can be further rewritten as

$$\begin{aligned} (18) = \mathrm{tr}(Z^\ell Q_{LL} Z^{\ell\top}) + 2\mathrm{tr}(Z^\ell Q_{Ld} M^{\ell\top}) + \mathrm{tr}(MQ_{dd}M^\top) \\ + \mathrm{tr}(Z^u Q_{LL} Z^{u\top}) + 2\mathrm{tr}(Z^u Q_{Ld} M^{u\top}) \end{aligned} \tag{19}$$

which clearly shows that the optimization over submatrices Z^ℓ and Z^u can be conducted independently.

By setting the derivative of the objective (17) (which is also the derivative of (19)) with respect to Z^u to 0, we can obtain a closed-form solution for Z^u:

$$Z^u = -M^u Q_{dL} Q_{LL}^{-1}. \tag{20}$$

Although no closed-form solution exists for Z^ℓ due to the linear constraints in (17), note that the objective in (19) actually can be further decomposed into independent terms for each row of Z^ℓ. Furthermore, the linear constraints in (17) are row-wise separable regarding $Z^\ell, \boldsymbol{\xi}, \boldsymbol{\eta}$ as well. Therefore, we can optimize each row of Z^ℓ independently by solving a small standard quadratic programming. For example, the ith row of Z, $Z_{i:}$ and $\boldsymbol{\xi}_i, \boldsymbol{\eta}_i$, can be optimized as

$$\begin{aligned} \min_{Z_{i:},\boldsymbol{\xi}_i,\boldsymbol{\eta}_i} \quad & \boldsymbol{\xi}_i + \boldsymbol{\eta}_i + \frac{\gamma}{2} Z_{i:} Q_{LL} Z_{i:}^\top + \gamma Z_{i:} Q_{Ld} M_{i,:}^\top \\ \text{subject to} \quad & \boldsymbol{\xi}_i \geq 1 - Z_{il} \quad \text{for } l \in Y_{i:}, \\ & \boldsymbol{\eta}_i \geq 1 + Z_{i\bar{l}} \quad \text{for } \bar{l} \in \bar{Y}_{i:}, \\ & \boldsymbol{\xi}_i \geq 0, \boldsymbol{\eta}_i \geq 0 \end{aligned} \tag{21}$$

Algorithm 1. Binary Line Search

Input: $X, Y, \alpha, \beta, \gamma$, a small constant $\tau > 0$.
Initialize: set $V_\ell = 0, V_u = 1$.
Repeat:
1. set $\theta = \frac{V_\ell + V_u}{2}$.
2. given the current θ, solve the inner minimization problem (15) for M^*, Z^*, S^{*-1} using block-coordinate descent method.
3. compute the subgradient $\frac{\partial f}{\partial \theta}$ according to Eq.(16).
4. **if** $\|\frac{\partial f}{\partial \theta}\| < \tau$ **then** return **end if**
5. **if** $\frac{\partial f}{\partial \theta} > 0$ **then** $V_\ell = \theta$ **else** $V_u = \theta$ **end if**

Until $(V_u - V_\ell) < \tau$

Algorithm 2. Block-Coordinate Descent Optimization

Input: $X, Y, \alpha, \beta, \gamma$, a small constant $\tau > 0$.
Initialize: set $M = X$ and randomly initialize Z.
Repeat:
1. recompute S using Eq.(24).
2. with given S, M, recompute Z^u using Eq.(20), and recompute each row $Z^\ell_{i:}$ by solving the quadratic programming in (21).
3. with given S, Z, recompute M using Eq.(23).

Until changes in M, Z is smaller than τ.

which can be solved using any standard quadratic program solver.

Optimization over M. Given fixed Z, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ and S, we optimize M as an unconstrained quadratic optimization problem

$$M = \arg\min_M \frac{\beta}{2}\|X - M\|_F^2 + \frac{\gamma}{2}\mathrm{tr}([M, Z]Q[M, Z]^\top). \tag{22}$$

Setting the derivative of the objective function with respect to M to 0 yields

$$M = (\beta X - \gamma Z Q_{Ld})(\beta I + \gamma Q_{dd})^{-1}. \tag{23}$$

Optimization over S. Given Z, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ and M, the minimization over S has a closed-form solution as suggested in the Proposition we presented above; i.e.,

$$S = (E_\theta[M, Z]^\top [M, Z] E_\theta + \epsilon_i I)^{1/2} \tag{24}$$

where $\epsilon_i > 0$ is a small value added to achieve an invertible S.

The overall block-coordinate descent procedure is given in Algorithm 2. By employing the block-coordinate descent inner convex minimization, the binary line search algorithm we developed obviously provides a scalable optimization tool for the target non-smooth convex optimization problem.

Table 1. Data set properties: $k = \#$ classes, $T = \#$ instances, $C =$ label cardinality

	Arts	Comp.	Edu.	Entain.	Health	Recr.	Ref.	Sci.	Soc.	Socie.
k	11	12	8	9	9	11	10	7	7	14
T	2525	2046	2816	2649	2932	2454	795	951	1181	4559
C	2.54	2.74	2.54	2.53	2.18	2.51	2.16	2.52	2.29	2.89

5 Experiments

To evaluate the proposed method, we conducted experiments on a set of multi-topic web page classification data sets [31]. Each data set consists of web pages collected from the yahoo.com domain. We preprocessed the data sets by first removing the largest class label (which covered more than 50% of the instances) and removing class labels that had fewer than 250 instances (for some data sets, we even used larger thresholds 300 and 400 to obtain larger label cardinalities). When the label cardinality of a data set is close to 1, the classification task is close to a standard single label multi-class task. The effectiveness of multi-label learning can be best demonstrated on data sets whose label cardinalities are reasonably large. We also removed any instances that had no labels or every label. For the input feature representation, we removed the less frequent features and converted the remaining integer features into a standard *tf-idf* encoding. The properties of the preprocessed data sets are summarized in Table 1.

We compared our proposed method (referred to as *TRANS* in the results) with four other large margin multi-label learning baselines:

- *CSRL*, the large margin multi-label learning method developed in (2) [8], based on a calibrated separation ranking loss.
- *CONS*, a variant of CSRL that replaces the regularizer $\|W\|_F^2$ with the constraint $\|W_{i:}\|_2 \leq \alpha$, as in Section 3.1.
- *dCSRL*, which first uses PCA to reduce the input dimension of the combined labeled and unlabeled data, then applies the CSRL method.
- *dCONS*, which first uses PCA to reduce the input dimension of the combined labeled and unlabeled data, then applies the CONS method.

Although numerous multi-label learning methods appear in the literature we restrict our attention to *convex* training methods to ensure that the results are repeatable independent of any particular implementation. In particular, for supervised multi-label losses we focus our comparison on CSRL, since previous work has demonstrated that this obtains state-of-the-art performance among convex supervised approaches [8]. (Note that the semi-supervised formulation presented in this paper can be easily applied to *any* convex multi-label loss [23]. However, finding tractable convex reformulations for losses specifically tailored for multi-labeled classification, such as F-measure [4], remains an open problem in the literature—e.g., the advanced formulation given in [4] still relies on an NP-hard constraint generation oracle.)

Table 2. Average transductive micro-F1 results over 10 repeats (± standard deviation)

Data set	CSRL	CONS	TRANS	dCSRL	dCONS
Arts	0.42 ±0.01	0.37 ±0.01	0.50 ±0.01	0.44 ±0.01	0.42 ±0.01
Computers	0.45 ±0.02	0.34 ±0.01	0.53 ±0.01	0.42 ±0.01	0.41 ±0.02
Education	0.56 ±0.02	0.45 ±0.01	0.61 ±0.01	0.56 ±0.01	0.49 ±0.02
Entertainment	0.61 ±0.03	0.44 ±0.02	0.63 ±0.01	0.50 ±0.02	0.46 ±0.03
Health	0.60 ±0.02	0.35 ±0.01	0.64 ±0.01	0.49 ±0.01	0.44 ±0.01
Recreation	0.48 ±0.02	0.32 ±0.05	0.53 ±0.01	0.41 ±0.01	0.39 ±0.01
Reference	0.55 ±0.02	0.41 ±0.01	0.55 ±0.01	0.36 ±0.01	0.34 ±0.01
Science	0.68 ±0.01	0.54 ±0.04	0.72 ±0.01	0.64 ±0.01	0.56 ±0.01
Social	0.63 ±0.01	0.53 ±0.06	0.67 ±0.01	0.55 ±0.01	0.48 ±0.01
Society	0.31 ±0.02	0.29 ±0.01	0.43 ±0.01	0.34 ±0.01	0.31 ±0.01

Table 3. Average transductive macro-F1 results on 10 repeats (± standard deviation)

Data set	CSRL	CONS	TRANS	dCSRL	dCONS
Arts	0.37 ±0.01	0.34 ±0.02	0.47 ±0.01	0.43 ±0.01	0.41 ±0.02
Computers	0.39 ±0.02	0.28 ±0.01	0.48 ±0.02	0.40 ±0.01	0.40 ±0.02
Education	0.48 ±0.02	0.39 ±0.01	0.54 ±0.01	0.54 ±0.01	0.47 ±0.02
Entertainment	0.50 ±0.05	0.34 ±0.01	0.53 ±0.03	0.46 ±0.01	0.42 ±0.03
Health	0.52 ±0.02	0.31 ±0.01	0.57 ±0.01	0.47 ±0.01	0.42 ±0.01
Recreation	0.37 ±0.02	0.28 ±0.02	0.45 ±0.01	0.38 ±0.01	0.36 ±0.01
Reference	0.34 ±0.01	0.32 ±0.01	0.42 ±0.01	0.32 ±0.01	0.30 ±0.01
Science	0.62 ±0.02	0.47 ±0.04	0.67 ±0.01	0.61 ±0.01	0.54 ±0.01
Social	0.53 ±0.02	0.44 ±0.05	0.58 ±0.02	0.51 ±0.01	0.45 ±0.01
Society	0.19 ±0.02	0.24 ±0.01	0.34 ±0.01	0.32 ±0.01	0.29 ±0.01

To also provide a comparison to *semi-supervised* methods, along the lines of [19–21], we furthermore include the latter two competitors, which use the unlabeled and labeled data to first learn a low dimensional representation for the input data. Note that the dimensionality reduction in this case is independent of the target labels. The goal of these experiments therefore is to isolate the consequences of using unlabeled data for subspace identification, and using label information in choosing such subspaces.

In these experiments we simply set the regularization parameters for TRANS to $\alpha = 0.01$, $\beta = 100$ and $\gamma = 50$, and set the regularization parameter for CSRL and CONS to $\alpha = 0.01$. The target dimensionality was set to 50 for the dimensionality reduction methods dCSRL and dCONS. The performance of each method is evaluated using the macro-F1 and micro-F1 measures [32]. We randomly selected 200 instances from each data set to be the labeled part, and another 1000 instances to be the unlabeled part. The process is repeated 10 times to generate 10 random partitions. The average performance and standard deviations of the five methods are reported in Table 2 and Table 3 respectively.

One can see from these results that the unlabeled data generally provides an improvement in generalization accuracy over the baseline supervised methods. First, using dimensionality reduction as a preprocessing step only gave mixed

benefits for the CSRL and CONS methods, although CONS generally benefited more. Interestingly, the proposed TRANS method, which learns a low dimensional subspace that also depends on the training labels, obtains a systematic and noticeable improvement over the other methods. TRANS significantly improves the macro-F1 measure over all comparison methods in every case, while achieving the same result for micro-F1 measure in every case except the "Reference" data set. The run times of TRANS and CSRL on 1200 data points (200 labeled and 1000 unlabeled) were approximately 10m for TRANS and 1m for CSRL, respectively, using simple Matlab implementations.

6 Conclusions

We have proposed a new method for semi-supervised multi-label classification that combines a state-of-the-art large margin multi-label learning approach with a current representation learning method. A key aspect of this approach is that it allows an efficient global training procedure. Experimental results show that the semi-supervised combination can outperform corresponding supervised and simple semi-supervised learning methods in a transductive setting.

There remains several important directions for future work. The current formulation is transductive; an out-of-sample extension of our approach is possible using a proposed technique from [23]. It also remains to investigate other representation learning formulations, such as $p = 1$, to determine their impact on performance. Another interesting direction is to extend the work of [4] to incorporate a tractable convex relaxation of F-measure for training.

References

1. Joachims, T.: Text Categorization with Support Vector Machines: Learn with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, Springer, Heidelberg (1998)
2. McCallum, A.: Multi-label text classification with a mixture model trained by em. In: AAAI Workshop on Text Learning (1999)
3. Zhu, S., Ji, X., Xu, W., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Conference on Information Retrieval, SIGIR (2005)
4. Petterson, J., Caetano, T.: Submodular multi-label learning. In: Advances in Neural Information Processing Systems, NIPS (2011)
5. Kazawa, H., Izumitani, T., Taira, H., Maeda, E.: Maximal margin labeling for multi-topic text categorization. In: Neural Infor. Processing Sys., NIPS (2004)
6. Godbole, S., Sarawagi, S.: Discriminative Methods for Multi-labeled Classification. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 22–30. Springer, Heidelberg (2004)
7. Hariharan, B., Zelnik-Manor, L., Vishwanathan, S., Varma, M.: Large scale max-margin multi-label classification with priors. In: Proceedings ICML (2010)
8. Guo, Y., Schuurmans, D.: Adaptive large margin training for multilabel classification. In: Conference on Artificial Intelligence, AAAI (2011)
9. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in Neural Information Processing, NIPS (2001)

10. Schapire, R., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39(2-3), 135–168 (2000)
11. Shalev-Shwartz, S., Singer, Y.: Efficient learning of label ranking by soft projections onto polyhedra. Journal of Machine Learning Research 7, 1567–1599 (2006)
12. Fuernkranz, J., Huellermeier, E., Mencia, E., Brinker, K.: Multilabel classification via calibrated label ranking. Machine Learning 73(2) (2008)
13. Yu, K., Yu, S., Tresp, V.: Multi-label informed latent semantic indexing. In: Conference on Research and Development in Information Retrieval, SIGIR (2005)
14. Yan, R., Tesic, J., Smith, J.: Model-shared subspace boosting for multi-label classification. In: Conference on Knowledge Discovery and Data Mining, KDD (2007)
15. Zhang, M., Zhou, Z.: Multi-label dimensionality reduction via dependency maximization. In: Conference on Artificial Intelligence, AAAI (2008)
16. Rai, P., Daumé III, H.: Multi-label prediction via sparse infinite CCA. In: Advances in Neural Information Processing Systems, NIPS (2009)
17. Kong, X., Yu, P.: Multi-label feature selection for graph classification. In: Proc. of the IEEE International Conference on Data Mining, ICDM (2010)
18. Ji, S., Tang, L., Yu, S., Ye, J.: A shared-subspace learning framework for multi-label classification. ACM Trans. Knowl. Discov. Data 4(2), 1–29 (2010)
19. Liu, Y., Jin, R., Yang, L.: Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Conf. on Artificial Intelligence, AAAI (2006)
20. Chen, G., Song, Y., Wang, F., Zhang, C.: Semi-supervised multi-label learning by solving a sylvester equation. In: SIAM Conference on Data Mining, SDM (2008)
21. Qian, B., Davidson, I.: Semi-supervised dimension reduction for multi-label classification. In: Conference on Artificial Intelligence, AAAI (2010)
22. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: Advances in Neural Information Processing Systems, NIPS (2006)
23. Zhang, X., Yu, Y., White, M., Huang, R., Schuurmans, D.: Convex sparse coding, subspace learning, and semi-supervised extensions. In: Conference on Artificial Intelligence, AAAI (2011)
24. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, NIPS (2006)
25. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning 73, 243–272 (2008)
26. Bach, F., Mairal, J., Ponce, J.: Convex sparse matrix factorizations. arXiv:0812.1869v1 (2008)
27. Lee, H., Raina, R., Teichman, A., Ng, A.: Exponential family sparse coding with application to self-taught learning. In: Int. Joint Conf. Artif. Intell., IJCAI (2009)
28. Rish, I., Grabarnik, G., Cecchi, G., Pereira, F., Gordon, G.: Closed-form supervised dimensionality reduction with generalized linear models. In: Proc. ICML (2007)
29. Hendrickx, J., Olshevsky, A.: Matrix p-norms are NP-hard to approximate if $p \neq 1, 2, \infty$. SIAM J. Matrix Anal. Appl. 31(5), 2802–2812 (2010)
30. Grave, E., Obozinski, G., Bach, F.: Trace lasso: a trace norm regularization for correlated designs. In: Neural Information Processing Systems, NIPS (2011)
31. Ueda, N., Saito, K.: Parametric mixture models for multi-labeled text. In: Advances in Neural Information Processing Systems, NIPS (2002)
32. Tang, L., Rajan, S., Narayanan, V.: Large scale multi-label classification via metalabeler. In: International WWW Conference (2009)