The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning with Incomplete Labels

Yingming Li[†], Zenglin Xu[‡], and Zhongfei Zhang[†]
[†] College of Information Science & Electronic Engineering, Zhejiang University, China
[‡]School of Computer Science and Engineering, University of Electronic Science and Technology of China
yingming@zju.edu.cn, zenglin@gmail.com, zhongfei@zju.edu.cn

Abstract

For many real-world tagging problems, training labels are usually obtained through social tagging and are notoriously incomplete. Consequently, handling data with incomplete labels has become a difficult challenge, which usually leads to a degenerated performance on label prediction. To improve the generalization performance, in this paper, we first propose the Improved Cross-View learning (referred as ICVL) model, which considers both global and local patterns of label relationship to enrich the original label set. Further, by extending the ICVL model with an outlier detection mechanism, we introduce the Improved Cross-View learning with Outlier Detection (referred as ICVL-OD) model to remove the abnormal tags resulting from label enrichment. Extensive evaluations on three benchmark datasets demonstrate that ICVL and ICVL-OD outstand with superior performances in comparison with the competing methods.

Introduction

With the rapid development of web applications such as social commerce sites and resource sharing systems, tagging has attracted much attention due to its capability of helping describe an item and allowing it to be found speedily by browsing or searching. Since the amount of data to be tagged increases explosively, it is very time-consuming and labor-intensive for manually labeling data. Consequently, automatic tagging techniques have become an effective alternative. Recent decades have witnessed the increasing applications of tagging techniques in many fields (Elisseeff and Weston 2001; Yu, Yu, and Tresp 2005; Liu, Jin, and Yang 2006; Krizhevsky, Sutskever, and Hinton 2012; Liu and Tsang 2015; Gao et al. 2016). Many different tagging approaches have been proposed based on different requirements from different areas (Zhang and Zhou 2014). However, most existing tagging methods assume that the given training labels are complete, i.e., the given training labels describe all the details of a document. In contrast, for many real-world tagging problems, training labels are usually obtained through social tagging and are notoriously incomplete, leaving typically many parts of the document content with no description at all. Therefore, handling data with incomplete training labels has become a challenge for tagging methods and has very important influence to the generalization performance of learned predictors.

Given training data with incomplete tags, a common strategy in several recent efforts is to handle the missing label problem via label space reduction (Tai and Lin 2012; Kapoor, Viswanathan, and Jain 2012). The key idea lies in reducing the dimensionality of label space via projection matrices. Consequently, predictions are first made on these reduced label spaces and then the original labels are recovered by back projection. Further, (Yu et al. 2014) takes a more direct approach by formulating the incomplete tagging problem as learning a low-rank regression matrix. However, such strategies focus on learning the local patterns of label correlation by separating labels into different clusters, where the labels within a cluster are strongly related to each other and irrelevant to the rest, and ignore the global dependency between labels. This is inappropriate since global label dependency also helps recover the useful label information.

On the other hand, certain literature, for example (Chen, Zheng, and Weinberger 2013), also considers learning with global label correlations to mitigate the influence of incomplete training label set. It is assumed that the given label set is incomplete and a label relationship matrix based on marginalized denoising autoencoder is learned to exploit the global label dependency. Consequently, the given label set is enriched with this label relationship matrix to tackle with the incomplete tagging problem. Theses proposed methods may achieve remarkably prediction performance. However, they mainly consider the global label relationship, but the correlations among labels can be complex and have local depedency.

Further, though the above enrichment with the help of label relationship would reduce the effect of incomplete label set, one cannot expect to accurately obtain a perfect correlation matrix for the given labels. For example, since the learning of label correlation usually relies on statistical modeling and requires an adequate supply of well-labeled samples, its accuracy would be affected by the limited data samples in real-world applications. Consequently, outliers (abnormal label values) would be produced in the process of label enrichment. This has adverse effect on learning robust predictors as the traditional machine learning methods are usually sensitive to outliers (Cabral et al. 2015).

To cope with the problem of incomplete tagging, it is necessary to consider both global and local patterns of label relationship. To achieve this goal, we propose to train robust predictors through exploiting global and local patterns of label correlations together. In particular, we present the Improved Cross-View learning (referred as ICVL) model, which treats training data (samples with incomplete tags) as unlabeled multi-view data and learns a cross-view agreement from two sub-tasks: 1) training a classifier to predict the complete tag set from observations, and 2) enriching the existing incomplete tag set with a label correlation matrix. This ICVL model considers both global and local patterns of label relationship by a low-rank marginalized denoising autoencoder regularization. Further, we incorporate outlier detection mechanism into the ICVL model to remove the abnormal tags resulting from label enrichment and propose an improved cross-view learning with outlier detection (ICVL-OD) model, which considers comprehensive label correlations and outlier detection simultaneously.

Respective algorithms for solving the optimization problems of ICVL and ICVL-OD are developed by using alternating optimization. We demonstrate their promise through extensive evaluations in three real datasets in comparison with the peer methods in the literature.

Related Work

Tagging methods in the literature are mainly categorized into two types: generative models and discriminative models. Most generative methods introduce a set of latent variables to learn the joint distribution of the sample features and semantic labels (Barnard et al. 2003; Blei and Jordan 2003). (Blei and Jordan 2003) proposes the correspondence Latent Dirichlet Allocation model under which labels and image features share latent variables. Discriminative methods usually reduce the multi-label problem to a set of binary classification problems. The representative techniques for this category of approaches are extensions of SVM, which have demonstrated a strong discrimination power (Qi and Han 2007; Yang, Dong, and Hua 2006; Li et al. 2013). In particular, when a document is represented as a bag of instances, and belongs to a bag of classes (Zhou and Zhang 2007; Zhou et al. 2012), the original tagging problem becomes a multi-instance and multi-label learning problem. (Zhou and Zhang 2007) solves this multi-instance and multi-label learning problem by mapping it into a single-instance and multi-label learning problem.

Further, the idea of exploiting label correlations between different labels is considered in the literature to facilitate the learning process (Jin et al. 2008; Hariharan et al. 2010; Sun, Zhang, and Zhou 2010; Huang, Zhou, and Zhou 2012; Ma et al. 2012; Xu et al. 2014). (Zhang and Zhang 2010) proposes to use a Bayesian network structure to encode the conditional dependencies of the labels so that the proposed approach is capable of modeling arbitrary order of label correlations. (Read et al. 2011) proposes a chaining method that can model label correlations while its computational complexity is maintained to be acceptable. (Chen, Zheng, and Weinberger 2013) proposes to enrich the user tags with a label correlation matrix learned from marginalized denoising autoencoder on training label set.

Label space dimension reduction methods have gradually been attracting much attention in multi-label learning with missing tags (Chen and Lin 2012; Tai and Lin 2012; Kapoor, Viswanathan, and Jain 2012). By projecting the label vector to a lower dimensional space using transformation matrices, (Kapoor, Viswanathan, and Jain 2012) makes full use of joint information within the labels to deal with datasets with incomplete labels. In order to handle large-scale data with missing labels, (Yu et al. 2014) performs label space reduction by imposing a low-rank constraint on the regression matrix. (Jing et al. 2015) proposes a semi-supervised low-rank mapping model to exploit correlations between labels and to use unlabeled samples.

Among the existing work, the closest work to ours is FastTag (Chen, Zheng, and Weinberger 2013). Through learning a label correlation matrix with a marginalized denoising autoencoder on label set, it exploits the global label dependency to enrich the user tags. We note that the difference from our work is significant as our method is to combine both merits of the label space reduction and label dependency by regularized learning with a low-rank marginalized denoising autoencoder. Moreover, we introduce outlier detection mechanism to remove the abnormal labels generated by the process of label enrichment. Consequently, our method outperforms the FastTag, which has been demonstrated in the experiments.

Learning with Incomplete Labels

In this section, we first introduce a cross-view learning method, which aims to reduce the effect of incomplete label set with label enrichment mechanism. Further, we incorporate the learning of global label dependency and local label correlations into the cross-view learning framework through a low-rank marginalized denoising autoencoder regularization. Consequently, an improved cross-view learning (ICVL) approach is developed to solve the problem of learning with missing labels.

Cross-View Learning

Given labeled training dataset $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, for supervised learning with linear regression, we obtain the following L_2 norm regularized loss

$$\mathcal{L}\left(\mathbf{W}\right) = \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2, \qquad (1)$$

where $\mathbf{W} \in \mathbb{R}^{l \times d}$ is the regression matrix.

In multi-label learning, one pervasive problem is that the given tags are usually incomplete, which means that the given labels are not complete to describe everything in an document. Such situation may exist even for a manually labeled document as human beings typically do not have the patience to give labels for all the details of an document. Thus, directly modeling with the original label set may not fully capture the relations between labels and features and lead to a decrease in the performance of the learned predictor.

To mitigate this influence, we propose to exploit label correlations to enrich the incomplete label matrix. In particular, by treating training data (samples with incomplete tags) as unlabeled multi-view data, we learn a cross-view agreement from two sub-tasks: 1) training a classifier $\mathbf{x}_i \rightarrow \mathbf{W}\mathbf{x}_i$ that predicts the complete tag set from observations, and 2) enriching the existing incomplete tag vector $\mathbf{y}_i$ with a label correlation matrix $\mathbf{y}_i \rightarrow \mathbf{B}\mathbf{y}_i$, where $\mathbf{B}$ captures the tag relationships. Correspondingly, we reformulate the criterion of linear multi-label learning in Eq.(1) as follows:

$$\mathcal{L}(\mathbf{W},\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2 \quad (2)$$

However, the above loss function in Eq.(2) has a trivial solution when $\mathbf{B} = 0$ and $\mathbf{W} = 0$, indicating that the current configuration is under-constrained. It is necessary to incorporate an additional regularization on $\mathbf{B}$ to guide to a reasonable solution.

Low-rank Marginalized Denoising Autoencoder Regularization

In this section, we propose a low-rank marginalized denoising autoencoder for estimating $\mathbf{B}$. In particular, we first model the correlation matrix $\mathbf{B}$ with a marginalized denoising autoencoder and then impose a low-rank constraint so that the learned $\mathbf{B}$ considers global and local label correlations simultaneously.

Our intention is to explore the label relationship by a reconstruction mechanism that $\mathbf{B}$ should be able to reconstruct the original tag set from its corrupted version. Let $\mathbf{y}_i \in \mathbb{R}^l$ be the i-th sample, which is under consideration now. Imagine that we first corrupt this vector by dropout distribution with probability $p \geq 0$ and then reconstruct the original $\mathbf{y}_i$ from the "corrupted version" $\tilde{\mathbf{y}}_i$ with a label relationship mapping matrix $\mathbf{B} : \mathbb{R}^l \rightarrow \mathbb{R}^l$. Here, for each tag vector $\mathbf{y}$ and dimension t, $p(\tilde{y}_t = 0) = p$ and $p(\tilde{y}_t = y_t) = 1-p$. Consequently, we learn this label relationship mapping by minimizing the squared reconstruction loss,

$$\mathcal{R}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2 \quad (3)$$

where $\mathbf{B} \in \mathbb{R}^{l \times l}$ can be considered as a label relationship matrix that predicts the presence of labels given the existing labels in $\tilde{\mathbf{y}}$.

Further, the repeated samples of $\tilde{\mathbf{y}}$ are made to reduce the variance in $\mathbf{B}$. In the limit (with infinitely corrupted versions of $\mathbf{y}$), the expected loss function under the dropout distribution can be expressed as

$$\mathcal{R}(\mathbf{B}) = \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \quad (4)$$

From the above loss function, we can see that the learning process of label relationship $\mathbf{B}_{ij}$ between label i and label j is influenced by all other labels, which helps capture a global dependency rather than a one-to-one correlation. Consequently, global label correlations are encoded into the label relationship matrix $\mathbf{B}$.

Besides the global correlations, the local patterns of label correlations naturally exist, where labels can be separated into different groups such that the labels within a group are strongly related to each other, while being irrelevant to the rest. These local patterns encourage a low-rank structure of $\mathbf{B}$, capturing the underlying local relationship among labels for the boosted generalization performance. However, this direct rank minimization problem is NP-hard in general. A widely-used convex relaxation of this problem is to regularize the target by the trace norm (nuclear norm) $\|\cdot\|_*$. Consequently, the framework of learning label relationship with a low-rank marginalized denoising autoencoder can then be formulated as

$$\begin{aligned}\mathcal{R}(\mathbf{B}) =& \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} + \eta\|\mathbf{B}\|_* \\ =& \frac{1}{2}\text{trace}\left(\mathbf{B}\mathbf{Q}\mathbf{B}^\top - 2\mathbf{P}\mathbf{B}^\top + \mathbf{Y}\mathbf{Y}^\top\right) + \eta\|\mathbf{B}\|_*\end{aligned} \quad (5)$$

where $\mathbf{P} = \sum_{i=1}^{n}\mathbf{y}_i\mathbb{E}\left[\tilde{\mathbf{y}}_i\right]^\top$, $\mathbf{Q} = \sum_{i=1}^{n}\mathbb{E}[\tilde{\mathbf{y}}_i]\mathbb{E}\left[\tilde{\mathbf{y}}_i\right]^\top + \mathbf{V}\left[\tilde{\mathbf{y}}_i\right]$, and

$$\begin{aligned}[\mathbf{Q}]_{\alpha,\beta} &= \begin{cases}\mathbf{S}_{\alpha\beta}q_\alpha q_\beta & \text{if } \alpha \neq \beta \\ \mathbf{S}_{\alpha\beta}q_\alpha & \text{if } \alpha = \beta\end{cases} \\ [\mathbf{P}]_{\alpha\beta} &= \mathbf{S}_{\alpha\beta}q_\beta\end{aligned} \quad (6)$$

where $q_\alpha = q_\beta = 1-p$, the variance matrix $\mathbf{V}\left[\tilde{\mathbf{y}}_i\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} = p(1-p)\delta(\mathbf{y}_i\mathbf{y}_i^\top)$, and $\mathbf{S} = \mathbf{Y}\mathbf{Y}^\top$ is the covariance matrix of the uncorrupted tag set. Here, $\delta(\cdot)$ denotes an operation that sets up all the entries except the diagonal to zero.

Improved Cross-View Learning

In this section we incorporate the low-rank marginalized denoising autoencoder regularization in Eq.(5) into the framework of cross-view learning in Eq.(2) to solve the problem of learning with incomplete labels. Consequently, the joint loss function can be written as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{B}}& \frac{1}{2}\sum_{i=1}^{n}||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2 \\ &+ \frac{\gamma}{2}\sum_{i=1}^{n}\mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} + \mu\|\mathbf{B}\|_*\end{aligned} \quad (7)$$

where λ, γ, and $\mu = \gamma\eta$ are regularization parameters.

Improved cross-view learning (ICVL) is referred to minimizing the above objective function. By considering training data (samples with incomplete tags) as unlabeled multi-view data, this framework can be interpreted from the following perspectives: 1) forcing a cross-view agreement between predicted vectors $\mathbf{W}\mathbf{x}$ and the enriched labels $\mathbf{B}\mathbf{y}$, and 2) learning an accurate $\mathbf{B}$ by introducing a low-rank marginalized denoising autoencoder regularization.

Optimization The loss function in Eq.(7) is jointly convex and can be solved efficiently through alternating optimization.

[**Update W**] when $\mathbf{B}$ is fixed, the problem in Eq.(7) with respect to $\mathbf{W}$ can be reformulated as

$$\min_{\mathbf{W}}\frac{1}{2}\sum_{i=1}^{n}||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2 \quad (8)$$

where $\mathbf{W}$ can be solved in a closed form:

$$\mathbf{W} = (\mathbf{B}\mathbf{Y})\,\mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}\right)^{-1} \tag{9}$$

[**Update B**] When $\mathbf{W}$ is fixed, the problem in Eq.(7) with respect to $\mathbf{B}$ can be reformulated as:

$$\begin{aligned}\min_{\mathbf{B}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \mu\|\mathbf{B}\|_* \\ + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)}\end{aligned} \tag{10}$$

The above optimization problem in Eq.(10) is convex and can be solved by various methods. In this paper, we exploit the alternating direction method of multipliers (ADMM (Boyd et al. 2011)) method to find the optimal solution of $\mathbf{B}$. By introducing an auxiliary variable $\mathbf{G} \in \mathbb{R}^{l\times l}$, the problem in Eq.(10) can be posed equivalently as a constrained optimization problem

$$\begin{aligned}\min_{\mathbf{B},\mathbf{G}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \mu\|\mathbf{G}\|_* \\ + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \\ \text{s.t.} \quad \mathbf{B} = \mathbf{G}\end{aligned} \tag{11}$$

The augmented Lagrange function of (11) is given by

$$\begin{aligned}\min_{\mathbf{B},\mathbf{G},\boldsymbol{\Upsilon}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i||^2 + \frac{\rho}{2}||\mathbf{B} - \mathbf{G}||^2 \\ +\mu\|\mathbf{G}\|_* + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \\ +\text{trace}(\boldsymbol{\Upsilon}^\top(\mathbf{B} - \mathbf{G}))\end{aligned} \tag{12}$$

where ρ, μ, and γ are regularization parameters and $\boldsymbol{\Upsilon} \in \mathbb{R}^{l\times l}$ are the Lagrange multipliers. This optimization problem can be solved with respect to each variable $(\mathbf{B}, \mathbf{G}, \boldsymbol{\Upsilon})$ by fixing other variables in an alternating manner.

When we fix $\mathbf{G}$ and $\boldsymbol{\Upsilon}$, we compute the gradient of the objective in Eq.(12) with respect to $\mathbf{B}$ and set them to 0 to obtain

$$\begin{aligned}\mathbf{B}\mathbf{Y}\mathbf{Y}^\top - \mathbf{W}\mathbf{X}\mathbf{Y}^\top + \gamma\,(\mathbf{B}\mathbf{Q} - \mathbf{P}) \\ + \rho\,(\mathbf{B} - \mathbf{G}) + \boldsymbol{\Upsilon} = 0\end{aligned} \tag{13}$$

The closed form solution can be computed as follows:

$$\begin{aligned}\mathbf{B}^{k+1} = \left(\mathbf{W}\mathbf{X}\mathbf{Y}^\top + \gamma\mathbf{P} + \rho\mathbf{G}^k - \boldsymbol{\Upsilon}^k\right) \\ \cdot\left(\mathbf{Y}\mathbf{Y}^\top + \gamma\mathbf{Q} + \rho\mathbf{I}\right)^{-1}\end{aligned} \tag{14}$$

Similarly, when we fix $\mathbf{B}$ and $\boldsymbol{\Upsilon}$, $\mathbf{G}$ can be obtained by solving the following problem:

$$\min_{\mathbf{G}} \frac{\rho}{2}||\mathbf{B} - \mathbf{G}||^2 + \text{tr}(\boldsymbol{\Upsilon}^\top(\mathbf{B} - \mathbf{G})) + \mu\|\mathbf{G}\|_* \tag{15}$$

The above optimization problem can be further transformed into the following form:

$$\min_{\mathbf{G}} \frac{1}{2}\|\mathbf{B} + \frac{\boldsymbol{\Upsilon}}{\rho} - \mathbf{G}\|^2 + \frac{\mu}{\rho}\|\mathbf{G}\|_* \tag{16}$$

The solution to $\mathbf{G}$ can be computed via the singular value thresholding (Cai, Candès, and Shen 2010):

$$\mathbf{G}^{k+1} = \mathcal{P}_{\frac{\mu}{\rho}}\left[\mathbf{B}^{k+1} + \frac{\boldsymbol{\Upsilon}^k}{\rho}\right] \tag{17}$$

where $\mathcal{P}_\theta(\mathbf{M}) = \mathbf{U}_M K_\theta(\boldsymbol{\Sigma}_\mathbf{M})\mathbf{V_M}^\top$ is the singular value thresholding operator with $\mathbf{M} = \mathbf{U_M}\boldsymbol{\Sigma}_\mathbf{M}\mathbf{V_M}^\top$ being the standard singular value decomposition of $\mathbf{M}$ and $K_\theta(\mathbf{A}_{ij}) = \text{sign}(\mathbf{A}_{ij})\max(0, |\mathbf{A}_{ij}| - \theta)$ being the soft-thresholding operator.

The multipliers $\boldsymbol{\Upsilon}$ can be updated directly by

$$\boldsymbol{\Upsilon}^{k+1} = \boldsymbol{\Upsilon}^k + \rho(\mathbf{B}^{k+1} - \mathbf{G}^{k+1}) \tag{18}$$

Based on the above analysis, the optimization algorithm of ICVL is outlined in Algorithm 1.

Algorithm 1: ICVL Algorithm

```
Input  : Training data set $\mathbf{X}$, $\mathbf{Y}$.
Output: Estimated mappings $\mathbf{W}$ and $\mathbf{B}$
Test   : Given a sample $\mathbf{x}$, $\mathbf{W}\mathbf{x}$ is used to score the
         dictionary of labels
Choose the label dropout probability $p$; obtain $\mathbf{P}$ and $\mathbf{Q}$
  with Eq.(6);
Repeat
    Optimize $\mathbf{W}$ with Eq.(9);
    Optimize $\mathbf{B}$ using the following steps;
    Repeat
        Fix $\mathbf{G}$, $\boldsymbol{\Upsilon}$ and update $\mathbf{B}$ with Eq.(14);
        Fix $\mathbf{B}$, $\boldsymbol{\Upsilon}$ and update $\mathbf{G}$ with Eq.(17);
        Fix $\mathbf{B}$, $\mathbf{G}$ and update $\boldsymbol{\Upsilon}$ with Eq.(18);
    until Convergence;
until Convergence;
```

Improved Cross-View Learning with Outlier Detection

Although the enriching mechanism $\mathbf{y} \leftarrow \mathbf{B}\mathbf{y}$ greatly reduces the effect of incomplete label set, one cannot expect to accurately obtain a perfect enrichment mapping $\mathbf{B}$ for the given labels. For example, since the learning of label correlation usually relies on statistical modeling and requires an adequate supply of well-labeled samples, its accuracy would be affected by the limited data samples in real-world applications. Consequently, outlier would be produced in the process of label enrichment. This makes it hard to learn accurate predictors as the traditional machine learning methods are usually sensitive to outliers. Thus, it is necessary to remove the outliers for building a robust model. To effectively detect the existence of the outliers and reduce their influence, we introduce an outlier detection mechanism by modeling residuals between improved label vector $\mathbf{B}\mathbf{y}$ and its estimated vector $\mathbf{Wx}$ with a parameter vector $\mathbf{r}$,

$$\mathbf{By} = \mathbf{Wx} + \mathbf{r} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{19}$$

where the vector $\mathbf{r}$ is responsible for detecting the outliers under the basic hypothesis that the i-th case is suspected to be an outlier when $\mathbf{r}_i$ is nonzero. Since we do not know which labels might be outliers, we add one residual error vector for each data point. This setup is very flexible since it allows any combination of labels to be outliers.

Further, we impose a sparse regularization constraint on this residual vector to control the capacity of the suspected data. Consequently, we can reformulate the criterion of multi-label learning in Eq.(7) as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{B},\mathbf{R}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{r}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2 + \kappa||\mathbf{R}||_1 \\ + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} + \mu\|\mathbf{B}\|_* \end{aligned} \tag{20}$$

where $\mathbf{R} \in \mathbb{R}^{l\times n} = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n]$ is the residual matrix and $||\cdot||_1$ denotes the sum of absolute values of all elements.

Improved cross-view learning with outlier detection (ICVL-OD) is referred to minimizing the above objective function. On the one hand, the *low-rank marginalized denoising autoencoder regularization* ensures that the learned enrichment mapping B encodes both global and local patterns of label correlations. On the other hand, it also considers removing outliers which are produced during the process of label enrichment to make the learned predictor robust.

Optimization The loss function in Eq.(20) is jointly convex and can be solved efficiently through alternating optimization.

[**Update** $\mathbf{W}$] when $\mathbf{B}$ and $\mathbf{R}$ are fixed, the problem in Eq.(20) with respect to $\mathbf{W}$ can be reformulated as

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{r}_i||^2 + \frac{\lambda}{2}||\mathbf{W}||_2^2 \tag{21}$$

where $\mathbf{W}$ can be solved in a closed form:

$$\mathbf{W} = (\mathbf{BY} - \mathbf{R})\,\mathbf{X}^\top \left(\mathbf{XX}^\top + \lambda\mathbf{I}\right)^{-1} \tag{22}$$

[**Update** $\mathbf{R}$] When $\mathbf{B}$ and $\mathbf{W}$ are fixed, the problem in Eq.(20) with respect to $\mathbf{R}$ can be determined by solving the following problem:

$$\min_{\mathbf{R}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{r}_i||^2 + \kappa||\mathbf{R}||_1 \tag{23}$$

where $||\cdot||_1$ denotes the sum of absolute values of all elements.

Based on L_1 norm soft thresholding operator (Herrity, Gilbert, and Tropp 2006), the solution to $\mathbf{R}$ can be directly computed as follows:

$$\mathbf{R}_{ij} = \text{sign}(\mathbf{H}_{ij}) \cdot \max(0, |\mathbf{H}|_{ij} - \kappa) \tag{24}$$

where $\mathbf{H} = \mathbf{BY} - \mathbf{WX}$.

[**Update** $\mathbf{B}$] When $\mathbf{W}$ and $\mathbf{R}$ are fixed, the problem in Eq.(20) with respect to $\mathbf{B}$ can be reformulated as:

$$\begin{aligned}\min_{\mathbf{B}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{r}_i||^2 + \mu\|\mathbf{B}\|_* \\ + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \end{aligned} \tag{25}$$

Similar to the optimization of ICVL in the previous section, ADMM is exploited to find the optimal solution of $\mathbf{B}$. By introducing an auxiliary variable $\mathbf{G} \in \mathbb{R}^{l\times l}$, the problem in Eq.(25) can be posed equivalently as a constrained optimization problem

$$\begin{aligned}\min_{\mathbf{B},\mathbf{G}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{r}_i||^2 + \mu\|\mathbf{G}\|_* \\ + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \\ \text{s.t.} \quad \mathbf{B} = \mathbf{G} \end{aligned} \tag{26}$$

The augmented Lagrange function of (26) is given by

$$\begin{aligned}\min_{\mathbf{B},\mathbf{G},\mathbf{\Upsilon}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{x}_i - \mathbf{r}_i||^2 + \frac{\rho}{2}||\mathbf{B} - \mathbf{G}||^2 \\ + \mu\|\mathbf{G}\|_* + \frac{\gamma}{2}\sum_{i=1}^{n} \mathbb{E}\left[||\mathbf{y}_i - \mathbf{B}\tilde{\mathbf{y}}_i||^2\right]_{p(\tilde{\mathbf{y}}_i|\mathbf{y}_i)} \\ + \text{trace}(\mathbf{\Upsilon}^\top(\mathbf{B} - \mathbf{G})) \end{aligned} \tag{27}$$

where ρ, μ, and γ are regularization parameters and $\mathbf{\Upsilon} \in \mathbb{R}^{l\times l}$ are the Lagrange multipliers. This optimization problem can be solved with respect to each variable $(\mathbf{B}, \mathbf{G}, \mathbf{\Upsilon})$ by fixing the other variables in an alternating manner.

When we fix $\mathbf{G}$ and $\mathbf{\Upsilon}$, we compute the gradient of the objective in Eq.(27) with respect to $\mathbf{B}$ and set them to 0 to obtain

$$\begin{aligned}\mathbf{BYY}^\top - \mathbf{WXY}^\top - \mathbf{RY}^\top + \gamma\,(\mathbf{BQ} - \mathbf{P}) \\ + \rho\,(\mathbf{B} - \mathbf{G}) + \mathbf{\Upsilon} = 0 \end{aligned} \tag{28}$$

The closed form solution can be computed as follows:

$$\begin{aligned}\mathbf{B}^{k+1} = \Big(\mathbf{WXY}^\top + \mathbf{RY}^\top + \gamma\mathbf{P} + \rho\mathbf{G}^k - \mathbf{\Upsilon}^k\Big) \\ \cdot \Big(\mathbf{YY}^\top + \gamma\mathbf{Q} + \rho\mathbf{I}\Big)^{-1} \end{aligned} \tag{29}$$

Similarly, when we fix $\mathbf{B}$ and $\mathbf{\Upsilon}$, $\mathbf{G}$ can be obtained by solving the following problem:

$$\min_{\mathbf{G}} \frac{\rho}{2}||\mathbf{B} - \mathbf{G}||^2 + \text{tr}(\mathbf{\Upsilon}^\top(\mathbf{B} - \mathbf{G})) + \mu\|\mathbf{G}\|_* \tag{30}$$

The optimization of the above problem is the same as that of Eq.(16) in the previous section.

Thus, The solution to $\mathbf{G}$ can be computed:

$$\mathbf{G}^{k+1} = \mathcal{P}_{\frac{\mu}{\rho}}\left[\mathbf{B}^{k+1} + \frac{\mathbf{\Upsilon}^k}{\rho}\right] \tag{31}$$

Algorithm 2: ICVL-OD Algorithm

```
Input  : Training data set $\mathbf{X}$, $\mathbf{Y}$.
Output: Estimated mappings $\mathbf{W}$, $\mathbf{B}$, and $\mathbf{R}$.
Test   : Given a sample $\mathbf{x}$, $\mathbf{Wx}$ is used to score the
         dictionary of labels
Choose the label dropout probability $p$; obtain $\mathbf{P}$ and $\mathbf{Q}$
  with Eq.(6);
Repeat
    Optimize $\mathbf{W}$ with Eq.(22);
    Optimize $\mathbf{R}$ with Eq.(24);
    Optimize $\mathbf{B}$ using the following steps;
    Repeat
        Fix $\mathbf{G}$, $\boldsymbol{\Upsilon}$ and update $\mathbf{B}$ with Eq.(29);
        Fix $\mathbf{B}$, $\boldsymbol{\Upsilon}$ and update $\mathbf{G}$ with Eq.(31);
        Fix $\mathbf{B}$, $\mathbf{G}$ and update $\boldsymbol{\Upsilon}$ with Eq.(32);
    until Convergence;
until Convergence;
```

The multipliers $\boldsymbol{\Upsilon}$ can be updated directly by

$$\boldsymbol{\Upsilon}^{k+1} = \boldsymbol{\Upsilon}^{k} + \rho(\mathbf{B}^{k+1} - \mathbf{G}^{k+1}) \tag{32}$$

Based on the above analysis, the optimization algorithm of ICVL-OD is outlined in Algorithm 2.

Computational Complexity For each iteration, updating $\mathbf{W}$ in Eq.(22) requires the construction of $(\mathbf{BY} - \mathbf{R})\,\mathbf{X}^\top$ and $\left(\mathbf{XX}^\top + \lambda\mathbf{I}\right)$, which will cost $\mathcal{O}(l^2n + ldn + d^2n)$. The inverse of $(\left(\mathbf{XX}^\top + \lambda\mathbf{I}\right))$ with $\mathcal{O}d^3$ is not necessary to be computed at each iteration. Updating $\mathbf{R}$ in Eq.(24) requires the construction of $(\mathbf{BY} - \mathbf{WX})$, which will cost nl^2. The soft thresholding operator of $\mathbf{H}$ will cost $\mathcal{O}(nl)$. Updating B in Eq.(29) requires the construction of $\left(\mathbf{WXY}^\top + \mathbf{RY}^\top + \gamma\mathbf{P} + \rho\mathbf{G}^k - \boldsymbol{\Upsilon}^k\right)$ and $\left(\mathbf{YY}^\top + \gamma\mathbf{Q} + \rho\mathbf{I}\right)$, which will cost $\mathcal{O}(ldn + l^2n)$. The inverse of $\left(\mathbf{YY}^\top + \gamma\mathbf{Q} + \rho\mathbf{I}\right)$ with $\mathcal{O}(l^3)$ complexity is also not necessary to be computed at each iteration. The main computation cost for updating $\mathbf{G}$ is the singular value thresholding operator and its complexity is $\mathcal{O}(d^3)$. Updating Υ in Eq.(32) cost $\mathcal{O}(l^2)$. Thus, the complexity of the optimization is $\mathcal{O}(l^2n + d^2n + d^3 + l^3 + ldn)$, where n represents the number of samples, d represents the dimension of feature vector, and l represents the dimension of label vector.

Experiments

We evaluate ICVL and ICVL-OD on three standard multi-label benchmark datasets including one audio dataset. All datasets are obtained from http://mulan.sourceforge.net/datasets-mlc.html.

Experimental Setup

In this section, we provide a detailed description of datasets, evaluation metrics, parameter setup, and baselines.

Datasets We have used three multi-label datasets, namely Enron, Bookmarks, and Birds for the experimentation purpose. Their statistics are described in Table 1.

Table 1: Statistics of the three datasets.

Dataset	Examples	Labels	Features
Enron	1702	53	1001
Bookmarks	7500	208	2150
Birds	645	19	260

Enron dataset contains email messages. It is from Enron corpus and made public during the legal investigation concerning the Enron corporation.

Bookmarks dataset is from Bibsonomy[1]. Bibsonomy is a social bookmarking and publication sharing system. Bookmarks contain metadata for bookmark items such as the URL of a web page and a description of the web page.

Birds dataset is formed by bird sounds. This audio dataset is collected in the H.J.Andrews (HJA) Long-Term Experimental Research Forest, in the Cascade mountain range of Oregon.

Evaluation Metric Three metrics, precision, recall, and F1 score, are often used to measure the performance of a tagging algorithm. Here, we also use them as our evaluation metrics. First, all the data are tagged with the five most relevant labels (i.e., labels with the highest prediction value). Second, precision (P) and recall (R) are computed for each label. The reported measurements are the average across all the labels. Further, both factors are combined in F1 score ($F1 = 2\frac{P*R}{P+R}$), which is reported separately. In all the metrics a higher value indicates a better performance.

Setup Cross-validation is used to estimate the performance of different methods. On the Enron and Birds datasets, we follow the experimental setup used in **Mulan**. Since there is no fixed split in the Bookmarks dataset in **Mulan**, we use a fixed training set of 80% of the data, and evaluate the performance of our predictions on the fixed test set of 20% of the data.

Baselines To demonstrate how ICVL-OD and ICVL improve the tagging performance in comparison with the state-of-the-art tagging methods, we compare them with the following representative tagging methods from the recent literature:

- LeastSquare (Bishop 2006).
- FastTag, a model which exploits global label dependency with marginalized denoising autoencoder regularization (Chen, Zheng, and Weinberger 2013).
- Low-rank empirical risk minimization for multi-label learning (referred to LEML) (Yu et al. 2014).
- Semi-supervised low-rank mapping learning for multi-label classification (referred to SLRM) (Jing et al. 2015).
- FastTag-OD, which incorporates the outlier detection mechanism into the FastTag method.

In particular, for SLRM, we only use its supervised version to make a fair comparison as our experimental setup is supervised.

[1] http://www.bibsonomy.org

Table 2: Comparison between the proposed algorithms and the competing models in terms of precision, recall, and F1 score on the three datasets. The Best performance in each case is indicated with bold face.

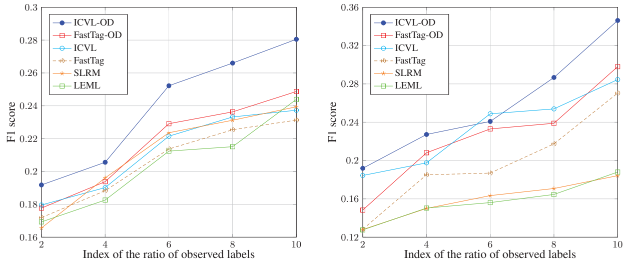
	Enron			Bookmarks			Birds		
Methods	precision	recall	F1	precision	recall	F1	precision	recall	F1
LeastSquare	0.2144	0.2396	0.2263	0.1253	0.2494	0.1669	0.1089	**0.5562**	0.1822
LEML	0.2546	0.2340	0.2439	0.1608	0.2588	0.1984	0.1134	0.5483	0.1880
SLRM	0.2456	0.2336	0.2394	0.1611	0.2566	0.1980	0.1113	0.5334	0.1842
FastTag	0.2133	0.2525	0.2313	0.1752	0.2571	0.2084	0.1952	0.4399	0.2704
ICVL	0.2020	**0.2875**	0.2373	0.1905	**0.2749**	0.2251	0.2005	0.4890	0.2845
FastTag-OD	0.2320	0.2680	0.2487	0.1876	0.2700	0.2214	0.2048	0.5461	0.2980
ICVL-OD	**0.3019**	0.2620	**0.2805**	**0.2121**	0.2567	**0.2323**	**0.2587**	0.5231	**0.3462**

Experimental Results

In Table 2, we summarize the precision, recall, and F1 score of the Enron, Bookmarks, and Birds datasets, for LeastSquare, LEML, SLRM, FastTag, FastTag-OD, ICVL, and ICVL-OD, respectively. On the task of multi-label data tagging, compared with ICVL-OD, LeastSquare mistakenly considers the incomplete training label set as the complete training label set. Although LEML formulates learning with missing labels as a general empirical risk minimization problem with a low-rank constraint, it cannot exploit the global label dependency to reduce the influence of incomplete tags. SLRM uses the trace norm regularization on regression matrix to perform label dimensional reduction, while it also ignores the global label dependency which naturally exists in the given label set. FastTag considers mining global label dependency with marginalized denoising autoencoder regularization to mitigate the influence of incomplete training label set, but it cannot take advantage of the local label correlations to further improve the generalization performance. While FastTag-OD incorporates the outlier detection mechanism into FastTag and performs better than FastTag, it also ignores the local label correlations. ICVL considers both global and local patterns of label correlations, while it ignores the effect of outliers produced from the enrichment process of label set. Consequently, from Table 2, we see that ICVL-OD performs better than LeastSquare, LEML, SLRM, FastTag, FastTag-OD, and ICVL as the F1 scores achieved by ICVL-OD are much higher than those achieved by the competing models in most cases. On one hand, this illustrates the importance of considering global label dependency and local label correlations simultaneously; on the other hand, this also demonstrates the effectiveness of outlier detection mechanism. In particular, ICVL-OD improves over the competing models with about 7% gain on F1 score in the Birds dataset.

The experiments also reveal several interesting observations:

- ICVL-OD and ICVL perform better than FastTag, LEML and SLRM. This shows that learning with global label dependency and local label correlations simultaneously may lead to a more robust tagging method.
- ICVL-OD and FastTag-OD perform better than ICVL and FastTag, respectively. This demonstrates that the incorporation of outlier detection mechanism helps improve the robustness of label-enriching models.



(a) Enron. (b) Birds.

Figure 1: The F1 score for the testing set as a function of the ratio of the observed labels for each training document on the Enron and Birds datasets. The i-th coordinate on the x-axis corresponds to the observed labels' ratio $i \times 10\%$.

Figure 1(a) and Figure 1(b) show the test F1 scores of LEML, SLRM, FastTag, FastTag-OD, ICVL, and ICVL-OD as a function of the ratio of the observed labels for each training document on Enron and Birds datasets, respectively. Random removal is exploited to control the ratios. We gradually increase the ratio of the observed labels and observe that ICVL-OD and other baselines show an increasing trend on the F1 scores with the increase of the ratio of the observed labels. Although the performance of ICVL-OD substantially drops when the observed label ratio is relatively small, ICVL-OD still achieves a comparatively better performance than the competing models in most cases. It shows the advantage of exploiting complex label correlations and outlier detection mechanism together on the task of learning with missing labels.

Conclusion

In this paper, to improve the generalization performance of incomplete tagging, in this paper, we first propose the Improved Cross-View Learning (referred as ICVL) model, which considers both global and local patterns of label relationships to enrich the original label set. Further, through extending the ICVL model with an outlier detection mechanism, the Improved Cross-View Learning with Outlier Detection (referred as ICVL-OD) model is introduced to remove the abnormal tags resulting from label enrichment.

Extensive evaluations on three benchmark datasets demonstrate that ICVL and ICVL-OD outstand with superior performances in comparison with the competing methods.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2017YFB1002400), National Natural Science Foundation of China (No. 61702448, 61672456, 61572111), the Key R&D Program of Zhejiang Province (No. 2018C03042), and the Fundamental Research Funds for the Central Universities (No. 2017QNA5008, 2017FZA5007, A03017023701012).

References

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. Springer.

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *SIGIR*.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3(1):1–122.

Cabral, R.; De la Torre, F.; Costeira, J. P.; and Bernardino, A. 2015. Matrix completion for weakly-supervised multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence* 37(1):121–135.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Chen, Y.-N., and Lin, H.-T. 2012. Feature-aware label space dimension reduction for multi-label classification. In *NIPS*, 1529–1537.

Chen, M.; Zheng, A. X.; and Weinberger, K. Q. 2013. Fast image tagging. In *ICML*, 1274–1282.

Elisseeff, A., and Weston, J. 2001. A kernel method for multi-labelled classification. In *NIPS*, 681–687.

Gao, W.; Wang, L.; Li, Y.; and Zhou, Z. 2016. Risk minimization in the presence of label noise. In *AAAI*, 1575–1581.

Hariharan, B.; Zelnik-Manor, L.; Varma, M.; and Vishwanathan, S. 2010. Large scale max-margin multi-label classification with priors. In *ICML*, 423–430.

Herrity, K. K.; Gilbert, A. C.; and Tropp, J. A. 2006. Sparse approximation via iterative thresholding. In *ICASSP*, volume 3, III–III. IEEE.

Huang, S.-J.; Zhou, Z.-H.; and Zhou, Z. 2012. Multi-label learning by exploiting label correlations locally. In *AAAI*, 949–955.

Jin, B.; Muller, B.; Zhai, C.; and Lu, X. 2008. Multi-label literature classification based on the gene ontology graph. *BMC bioinformatics* 9(1):525.

Jing, L.; Yang, L.; Yu, J.; and Ng, M. K. 2015. Semi-supervised low-rank mapping learning for multi-label classification. In *CVPR*, 1483–1491.

Kapoor, A.; Viswanathan, R.; and Jain, P. 2012. Multilabel classification using bayesian compressed sensing. In *NIPS*, 2645–2653.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Li, Y.; Tsang, I. W.; Kwok, J. T.; and Zhou, Z. 2013. Convex and scalable weakly labeled svms. *Journal of Machine Learning Research* 14(1):2151–2188.

Liu, W., and Tsang, I. W. 2015. Large margin metric learning for multi-label prediction. In *AAAI*, 2800–2806.

Liu, Y.; Jin, R.; and Yang, L. 2006. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 421–426.

Ma, H.; Chen, E.; Xu, L.; and Xiong, H. 2012. Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning. *Neurocomputing* 92:116–123.

Qi, X., and Han, Y. 2007. Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition* 40:728–741.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning* 85(3):333.

Sun, Y.; Zhang, Y.; and Zhou, Z. 2010. Multi-label learning with weak label. In *AAAI*.

Tai, F., and Lin, H.-T. 2012. Multilabel classification with principal label space transformation. *Neural Computation* 24(9):2508–2542.

Xu, L.; Wang, Z.; Shen, Z.; Wang, Y.; and Chen, E. 2014. Learning low-rank label correlations for multi-label classification with missing labels. In *ICDM*, 1067–1072. IEEE.

Yang, C.; Dong, M.; and Hua, J. 2006. Region-based image annotation using asymmetrical support vector machine-based multi-instance learning. In *CVPR*.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. S. 2014. Large-scale multi-label learning with missing labels. In *ICML*, 593–601.

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *SIGIR*, 258–265.

Zhang, M.-L., and Zhang, K. 2010. Multi-label learning by exploiting label dependency. In *KDD*, 999–1008. ACM.

Zhang, M., and Zhou, Z. 2014. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* 26(8):1819–1837.

Zhou, Z.-H., and Zhang, M.-L. 2007. Multi-instance multilabel learning with application to scene classification. In *NIPS*, 1609–1616.

Zhou, Z.; Zhang, M.; Huang, S.; and Li, Y. 2012. Multi-instance multi-label learning. *Artif. Intell.* 176(1):2291–2320.