Learning from Semi-Supervised Weak-Label Data[*]

Hao-Chen Dong and **Yu-Feng Li** and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210023, China
{donghc, liyf, zhouzh}@lamda.nju.edu.cn

Abstract

Multi-label learning deals with data objects associated with multiple labels simultaneously. Previous studies typically assume that for each instance, the full set of relevant labels associated with each training instance is given. In many applications such as image annotation, however, it's usually difficult to get the full label set for each instance and only a *partial* or even *empty* set of relevant labels is available. We call this kind of problem as 'semi-supervised weak-label learning' problem. In this work we propose the SSWL (Semi-Supervised Weak-Label) method to address this problem. Both instance similarity and label similarity are considered for the complement of missing labels. Ensemble of multiple models are utilized to improve the robustness when label information is insufficient. We formulate the objective as a bi-convex optimization problem with an efficient block coordinate descent algorithm. Experiments validate the effectiveness of SSWL.

Introduction

Conventional supervised learning often assumes that each instance is associated with a single label. However, in many real-world tasks, one instance usually has more than one labels. For example, in text categorization, a document on Olympic Game belongs to *business* and *sport* simultaneously; in image annotation, an image on the scene of Paris is associated with *tower* and *sky* simultaneously. Conventional supervised learning based on *one label per instance* is out of its capability to cope with this problem, and *multi-label learning* (Zhang and Zhou 2014) that deals with instances associated with a set of labels has received much attention.

In previous multi-label studies, a basic assumption for training data is that all the relevant labels of every training instance are known. However, in many applications, such assumption is hard to hold because obtaining all relevant labels is difficult, and generally only a *partial* or even *empty* label set can be *observed*. For example, suppose a training image is related to the concepts *car*, *road*, *people* and *building*. In real cases, the user may only tag *car*, *road* for the training image while missing the label *people* and *building*. What is worse, the training image may not be selected to tag for users due to the limited resources and thus the observed relevant label of the image is even an *empty* set.

Label incompleteness significantly influences the performance of multi-label learning (Zhou 2017). To alleviate it, there are some previous work. *Weak-label learning* focuses on the issue of *partial* relevant label set. Sun et al., (2010) presented the WELL approach based on the assumption that instance similarities are determined by a group of low-rank similarity matrixes. Bucak et al., (2011) presented the MLR-GL approach with the use of group lasso to regularize the training errors. *Semi-supervised multi-label learning* tries to handle the issue of *empty* relevant label set, where transductive multi-label learning methods (Liu, Jin, and Yang 2006; Chen et al. 2008; Guo and Schuurmans 2012; Kong, Ng, and Zhou 2013) that assume testing instances are from unlabeled instances, whereas pure semi-supervised multi-label learning methods (Zhao and Guo 2015; Zhan and Zhang 2017) try to make multi-label prediction for any unseen instance.

It is evident that neither weak-label learning nor semi-supervised multi-label learning can tackle the problem concerned in the paper. For example, weak-label learning ignores the use of many unlabeled instances that could be very useful; semi-supervised multi-label learning assumes that all the relevant labels are available for labeled instances, which is not the case in our situation. Note that the data scenario studies in the paper is quite different from previous multi-label studies. We call this kind of multi-label problem as *semi-supervised weak-label learning*. We illustrate the differences between the learning scenario in the paper and previous multi-label learning frameworks in Figure 1.

In this paper, we study semi-supervised weak-label learning problem and propose the SSWL (Semi-Supervised Weak-Label) method. Our basic assumption is that both the instance and label similarity are helpful for the complement of missing labels. Moreover, ensemble of multiple models usually performs more robust than a single model, when the label information is insufficient. Specifically, we first construct a regularization term based on smoothness assumption, that is, similar instances should have similar concept compositions within their label sets, which requires the final prediction to be coherent to the smoothness of instance and label similarity simultaneously. We build models for labeled and unlabeled instances respectively, and then we leverage the diverse models via the co-regularization frame-

[*] This research was supported by the National Natural Science Foundation of China (61333014, 61772262).

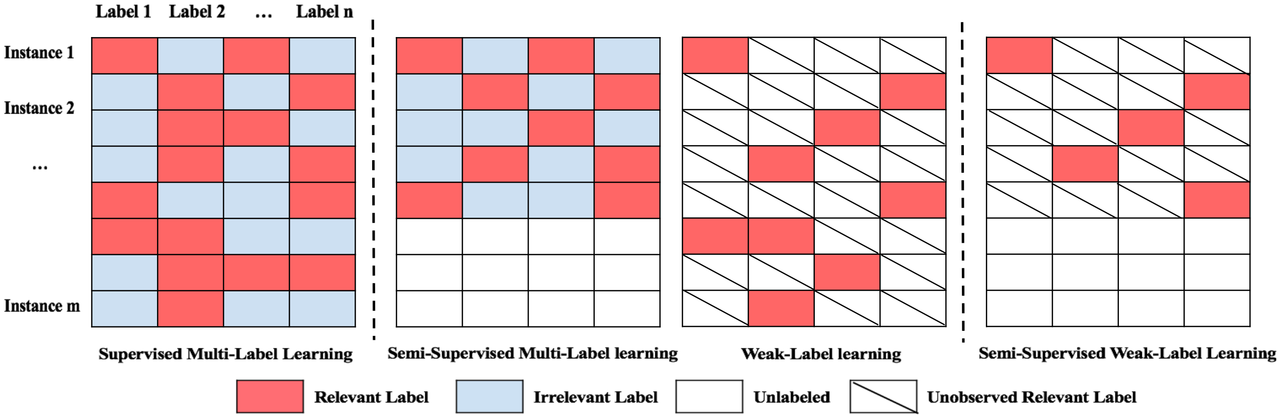
Label 1
Label 2
...
Label n
Instance 1
Instance 2
...
Instance m
Supervised Multi-Label Learning
Semi-Supervised Multi-Label learning
Weak-Label learning
Semi-Supervised Weak-Label Learning
Relevant Label
Irrelevant Label
Unlabeled
Unobserved Relevant Label

Figure 1: Four multi-label learning settings

work (Sindhwani, Niyogi, and Belkin 2005). We formulate the problem as a bi-convex formulation and provide an efficient block coordinate descent solution. The effectiveness of the proposed method is validated in experiments.

The rest of this paper is organized as follows. We start by a brief review of related work. Then we formulate the problem and present the proposed approach. Experimental results are reported, followed by the conclusion of this work.

Related Work

This line of weak-label learning research is raised in the past few years. Sun et al., (2010) and Bucak et al., (2011) are two early studies on this direction. Sun et al., (2010) presented the WELL approach based on the assumption that instance similarities are determined by a group of low-rank similarity matrixes. Bucak et al., (2011) presented the MLR-GL approach with the use of group lasso to regularize the training errors. Recently, many learning methods have tried to conquer weak-label problems. Examples include methods based on label co-occurrence (Wu, Jin, and Jain 2013; Zhu, Yan, and Ma 2010), sparse reconstructions (Lin et al. 2013), low-rank matrix completion (Xu, Jin, and Zhou 2013), etc. The weak-label problem also occurs in other learning scenarios, such as multi-instance multi-label learning (Yang, Jiang, and Zhou 2013). However, weak-label learning methods are not sufficient to tackle well the semi-supervised weak-label data, because they neglect the exploitation of a large amount of unlabeled instances that is known to be very useful.

Semi-supervised multi-label learning falls into two categories. One is transductive multi-label learning that assumes testing instances are from unlabeled instances. Examples include methods (Liu, Jin, and Yang 2006; Chen et al. 2008; Kong, Ng, and Zhou 2013; Wang, Tu, and Tsotsos 2013). Specifically, Liu et al., (2006) assumed that the similarity in the label space is closely related to that in the feature space, and thus employed the similarity in feature space to guide the learning of missing label assignments, which leads to a constrained nonnegative matrix factorization optimization. Chen et al., (Chen et al. 2008) constructed instance and label graph respectively and showed that the labels of unlabeled data finally can be obtained by solving a *Sylvester Equation*. Kong et al., (2013) formulated the transductive multi-label learning as an optimization problem of estimating label concept compositions. The other one is pure semi-supervised multi-label learning that could make multi-label prediction for any unseen instance. Examples include methods (Zhao and Guo 2015; Zhan and Zhang 2017). Specifically, Zhao et al., (2015) aimed to improve multi-label prediction performance by integrating label correlation and multi-label prediction in a mutually beneficial manner. Zhan et al., (2017) proposed an inductive co-training style method to address this problem. They generated two classification models by dichotomizing the feature space with diversity maximization to handle multi-label data, and then pairwise ranking predictions on unlabeled data was iteratively communicated for model refinement. Nevertheless, although semi-supervised multi-label learning have taken the incompleteness of relevant labels into account, it still assumes that full relevant labels are available for labeled instances and such an assumption does not hold in semi-supervised weak-label data.

The Proposed Method

Problem Statement and Notations

In the original supervised multi-label setting, we are given a training data set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$. The instance $\mathbf{x}_i$ is represented as a d-dimensional real value vector. The label $\mathbf{y}_i$ can be represented as an n-dimensional binary label vector, with 1 indicating that the instance belongs to the concept corresponding to the dimension and -1 otherwise. All the labels consist of the label space $\mathcal{Y} = \{1, -1\}^n$. In other words, we have an instance matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]'$ each row for one instance, and a full label matrix $\mathbf{Y} \in \{1, -1\}^{m \times n}$ where $Y_{ij} = 1$ means the i-th instance has the j-th label, while $Y_{ij} = -1$ means the i-th instance doesn't have the j-th label. We let $\mathcal{S}_i = \{j | Y_{ij} = 1, j = 1, \ldots, n\}$ denote the full set of relevant labels for instance $\mathbf{x}_i$, $\forall i = 1, \ldots, m$.

In the semi-supervised weak-label learning setting, we

Table 1: Summary of Some Basic Notations

Notations	Meaning
m	number of instances
n	number of labels
d	number of features
$\mathbf{x} \in \mathbb{R}^d$	instance feature vector
$\mathbf{X} \in \mathbb{R}^{m \times d}$	instance feature matrix
$\mathcal{Y} = \{+1, -1\}^n$	label space
$\mathbf{y} \in \mathcal{Y}$	label vector
$\mathbf{Y} \in \{1, -1\}^{m \times n}$	label matrix
$\mathbf{C} \in \{0, 1\}^{m \times n}$	label occurrence matrix
$\widehat{\mathbf{Y}} \in \{1, -1\}^{m \times n}$	predictive label matrix

have the same instance matrix $\mathbf{X}$. However, the full label matrix $\mathbf{Y}$ is not available and instead we are only given a label occurrence matrix $\mathbf{C} \in \{0, 1\}^{m \times n}$ where $C_{ij} = 1$ means the i-th instance has the j-th label (the same as the case of $Y_{ij} = 1$), while when $C_{ij} = 0$, the underlying label Y_{ij} has two possible values. One is that $Y_{ij} = 1$, which means the i-th instance has the j-th label but it is unobserved, and the other is $Y_{ij} = -1$, meaning that the i-th instance does not has the j-th label, which is also unknown. Moreover, in semi-supervised weak-label learning, there is no further constraint about the number of observed relevant labels for each instance. Specifically, let $\hat{\mathcal{S}}_i = \{j | C_{ij} = 1, j = 1, \ldots, n\}$ denote the *observed* set of relevant labels for instance $\mathbf{x}_i$, $\forall i = 1, \ldots, m$. We then have $\hat{\mathcal{S}}_i$ is only a subset of full relevant label set $\mathcal{S}_i$, or even $\hat{\mathcal{S}}_i$ is even an empty set. Our goal is to learn a predictive label matrix $\widehat{\mathbf{Y}} \in \{1, -1\}^{m \times n}$ from $\{\mathbf{X}, \mathbf{C}\}$ to approximate $\mathbf{Y}$.

Problem Formulation

A direct strategy to deal with semi-supervised weak-label setting is to decompose the task into n independent binary classification problems, each for one label. For each label, a number of existing binary semi-supervised learning algorithms, such as label propagation (Zhu and Goldberg 2009) and semi-supervised SVMs (Chapelle, Schölkopf, and Zien 2006), can be employed. Zhu et al., (2005) gave an example illustrating the importance of considering label correlation. Such a strategy, however, ignores label correlation that could be very useful and often leads to suboptimal problem. In order to take label correlation into account, we in this paper propose to use both the instance and label similarity for the complementation of missing relevant labels. Specifically, we introduce a regularization term based on smoothness assumption, that is, similar instances should have similar concept compositions within their label sets, which requires the final prediction to be coherent to the smoothness of instance and label similarity simultaneously.

Formally, let G_I be a weighted neighborhood graph on labeled and unlabeled instances. Each vertex in G_I corresponds an instance $\mathbf{x}_i$, and an edge between $\mathbf{x}_i$ and $\mathbf{x}_p$ means, $\mathbf{x}_i$ is a k nearest neighbor of $\mathbf{x}_p$ or $\mathbf{x}_p$ is a k nearest neighbor of $\mathbf{x}_i$. We define a sparse $m \times m$ matrix $\mathbf{S}$ (Kong, Ng, and Zhou 2013), indicating the similarities among neighboring instances:

$$S_{ip} = \begin{cases} \frac{1}{z_i}\exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_p\|_2^2}{2\sigma^2}), & \text{if } p \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases}$$

where $\mathcal{N}_i$ is the instance set of i-th instance's k nearest neighbors. $z_i = \sum_{p \in \mathcal{N}_i} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_p\|_2^2}{2\sigma^2})$, thus $\sum_{p \in \mathcal{N}_i} S_{ip} = 1$. In order to reduce computational cost of k nearest neighbor search among labeled and unlabeled instances, we use kd-tree (Weber, Schek, and Blott 1998) to efficiently search for approximate k nearest neighbors for each instance and use multi-label dimensionality reduction approach to reduce the impact of the curse of dimensionality (Zhang and Zhou 2010).

In semi-supervised weak-label learning, the observed relevant label sets of instances are incomplete. We can not directly compute the label similarity matrix $\mathbf{L}$ like instance similarity, and thus we need to learn it. In the sequel for simplicity of discussion, we first assume that label similarity matrix $\mathbf{L}$ is given.

To estimate the predictive label matrix $\widehat{\mathbf{Y}}$, there are two main approaches from the perspective of smoothness assumption. Firstly, from the perspective of the instance similarity, the relevant label set of an instance can be derived by that of its nearest neighbors, i.e., $\hat{Y}_{ij} \approx \sum_{p \in \mathcal{N}_i} S_{ip}\hat{Y}_{pj}$. Secondly, from the perspective of the similarity of label, the assignment of one certain label on training instances, can be derived by the assignments of its adjacent labels, i.e., $\hat{Y}_{ij} \approx \sum_{q \in \hat{\mathcal{N}}_j} \hat{Y}_{iq}L_{qj}$, where $\hat{\mathcal{N}}_j$ is the label set of j-th label's k nearest neighbors. Obviously, the predictive label matrix is not only related to the instance similarity but also the label similarity. This motivates us to characterize both the smoothness of the instance similarity and the label similarity, that is,

$$\hat{Y}_{ij} \approx \sum_{p \in \mathcal{N}_i} \sum_{q \in \hat{\mathcal{N}}_j} S_{ip}\hat{Y}_{pq}L_{qj} \qquad (1)$$

Consequently we obtain a new regularization term:

$$\begin{aligned} \Omega(\widehat{\mathbf{Y}}, \mathbf{S}, \mathbf{L}) &= \sum_{ij}(\hat{Y}_{ij} - \sum_{p \in \mathcal{N}_i} \sum_{q \in \hat{\mathcal{N}}_j} S_{ip}\hat{Y}_{pq}L_{qj})^2 \\ &= \|\widehat{\mathbf{Y}} - \mathbf{S}\widehat{\mathbf{Y}}\mathbf{L}\|_F^2 \qquad (2) \end{aligned}$$

where $\|\mathbf{M}\|_F^2 = tr(\mathbf{M}\mathbf{M}^{'})$ and $tr(\cdot)$ is the trace of a matrix.

With the new regularization term, we aim to learn a promising predictive label $\widehat{\mathbf{Y}}$ in semi-supervised weak-label setting. Inspired by (Zhou 2012), we employ ensemble learning which is known to be more robust than a single model, especially when label information is insufficient. Specifically, we first build two models with the new regularization term for labeled and unlabeled instances respectively, and then leverage the diverse models by the co-regularization framework (Sindhwani, Niyogi, and Belkin 2005) to derive a robust predictive result.

Formally, let $\mathbf{X}\mathbf{W}$ and $\mathbf{X}\bar{\mathbf{W}}$ denote two linear multi-label models, where $\mathbf{W}, \bar{\mathbf{W}} \in \mathbb{R}^{d \times n}$ are the coefficient matrixes.

The first model $\mathbf{XW}$ is initialized to predict the observed relevant labels, i.e., the elements with $C_{ij} = 1$, whose objective is formulated as $\|(\mathbf{XW})\circ\mathbf{C}-\mathbf{C}\|_F^2$ where $\circ$ is the Hadamard product (the entrywise product). The second model is initialized to predict the uncertain elements in the label occurrence matrix $\mathbf{C}$, i.e., the elements with $C_{ij} = 0$, whose objective is formulated as $\|(\mathbf{X}\bar{\mathbf{W}})\circ(\mathbf{E}-\mathbf{C})+(\mathbf{E}-\mathbf{C})\|_F^2$ where $\mathbf{E}_{m,n}$ is the all-one matrix. It is obvious that these two models are diverse but not strong enough, we then leverage them via the promising co-regularization framework to derive a robust predictive result. The idea is to enforce two models become consistent on the prediction of the uncertain elements in the label occurrence matrix $\mathbf{C}$, whose objective is cast as $\|(\mathbf{X}(\mathbf{W}-\bar{\mathbf{W}}))\circ(\mathbf{E}-\mathbf{C})\|_F^2$. Summarizing the above consideration, we then derive our objective which is to find $\mathbf{W}$, $\bar{\mathbf{W}}$ and the label similarity matrix $\mathbf{L}$ such that the following objective is minimized,

$$\begin{aligned}\min_{\mathbf{W},\bar{\mathbf{W}},\mathbf{L}} \quad & \|(\mathbf{XW})\circ\mathbf{C}-\mathbf{C}\|_F^2+\alpha\Omega(\mathbf{U},\mathbf{S},\mathbf{L})+ \\ & \beta\|(\mathbf{X}(\mathbf{W}-\bar{\mathbf{W}}))\circ(\mathbf{E}-\mathbf{C})\|_F^2+ \\ & \zeta\|(\mathbf{X}\bar{\mathbf{W}})\circ(\mathbf{E}-\mathbf{C})+(\mathbf{E}-\mathbf{C})\|_F^2 \\ \text{s.t.} \quad & \mathbf{U}=(\mathbf{XW})\circ\mathbf{C}+(\mathbf{X}\bar{\mathbf{W}})\circ(\mathbf{E}-\mathbf{C})\end{aligned} \tag{3}$$

where α, β, ζ are the parameters. $\mathbf{U}=(\mathbf{XW})\circ\mathbf{C}+(\mathbf{X}\bar{\mathbf{W}})\circ(\mathbf{E}-\mathbf{C})$ is the integrated prediction of two models. Eq. 3 on one side considers the smoothness on both instance and label similarity. On the other side it absorbs ensemble learning to derive robust results. It is worth noting that classical label propagation techniques can be realized as a special case of our proposal. Specifically, when setting $\beta = \zeta = 0$ and $\mathbf{L}$ be the identity matrix, our proposal is equivalent to n classical label propagation forms, each for one label. Another advantage of our proposal is that it does not restrict the testing instances to be picked from unlabeled ones, and is able to make prediction for any unseen instance.

```
Algorithm 1: SSWL Method
Input  : X: $m \times d$ instance matrix
         C: $m \times n$ label occurrence matrix
         S: $m \times m$ similarity matrix of instances
Output: W and $\bar{W}$: $d \times n$ coefficient matrixes
Initialize W, $\bar{W}$, L;
while not converged do
    Fix $\bar{W}$ and L, update W by Eq.5;
    Fix W and L, update $\bar{W}$ by Eq.7;
    Fix W and $\bar{W}$, update L by Eq.9;
end
```

Block Coordinate Descend Algorithm

The objective function in Eq.3 involves $\mathbf{W}$, $\bar{\mathbf{W}}$ and $\mathbf{L}$, and it is not easy to optimize with respect to all the variables simultaneously. Fortunately, Eq.3 is a bi-convex function (Gorski, Pfeuffer, and Klamroth 2007) which means if we fix $\mathbf{W}$ and $\bar{\mathbf{W}}$, the optimization of $\mathbf{L}$ is convex, and alternatively, when fixing $\mathbf{L}$ and $\mathbf{W}$(or $\bar{\mathbf{W}}$), the optimization of $\bar{\mathbf{W}}$(or $\mathbf{W}$) is also convex. In this case, here we extend an efficient block coordinate descend algorithm (Tseng 2001). Specifically, we first optimize the objective function with respect to $\mathbf{W}$ when $\bar{\mathbf{W}}$ and $\mathbf{L}$ are fixed, then optimize variable $\bar{\mathbf{W}}$ when $\mathbf{W}$ and $\mathbf{L}$ are fixed, and finally optimize variable $\mathbf{L}$ when the first two variables are fixed. These three subroutines are repeated until convergence. Algorithm 1 summarizes the pseudo-code of our proposal. More specifically, we first introduce some notations.

$$\begin{aligned}\mathbf{H} &= \mathbf{I}\otimes\mathbf{X} \\ \mathbf{O} &= (\mathbf{I}-\mathbf{L}^{'}\otimes\mathbf{S}) \\ \mathbf{P} &= \mathbf{H}^{'}\mathbf{diag}(\mathbf{vec}(\mathbf{C})) \\ \mathbf{Q} &= \mathbf{H}^{'}\mathbf{diag}(\mathbf{vec}(\mathbf{E}-\mathbf{C}))\end{aligned}$$

Here **vec**$(\mathbf{M})$ is the vectorization of matrix $\mathbf{M}$, **diag**$(\mathbf{v})$ is a diagonal matrix with vector **v** as its diagonal elements and $\otimes$ is the Kronecker product. Moreover, since the updating subroutines of variables $\{\mathbf{W},\bar{\mathbf{W}},\mathbf{L}\}$ involve the solving of linear equations, we introduce a theorem from (Horn and Johnson 1991). By using this theorem, we can easily transform the complex linear matrix equation encountered in the updating subroutines to the normal linear equations.

Theorem 1. *(Horn and Johnson 1991) Suppose a matrix $\hat{\mathbf{X}}$ satisfies an equation, $\sum_{i=1}^{b}\mathbf{A}_i\hat{\mathbf{X}}\mathbf{B}_i = \boldsymbol{V}$, where $\{\mathbf{A}_i\}_{i=1}^{b}$, $\{\mathbf{B}_i\}_{i=1}^{b}$ and $\boldsymbol{V}$ are known. To obtain the solution $\hat{\mathbf{X}}$, one could solve the following equivalent problem instead, $(\sum_{i=1}^{b}\mathbf{B}_i^{'}\otimes\mathbf{A}_i)\boldsymbol{vec}(\hat{\mathbf{X}}) = \boldsymbol{vec}(\boldsymbol{V})$, which is a normal linear equation.*

Update W with Fixed $\bar{\mathbf{W}}$ and L

When $\bar{\mathbf{W}}$ and $\mathbf{L}$ are fixed, we have the following equation for $\mathbf{W}$ by setting the derivative of Eq.3 w.r.t $\mathbf{W}$ to zero,

$$\begin{aligned}&\alpha\mathbf{X}^{'}((\mathbf{U}+\mathbf{S}^{'}\mathbf{SULL}^{'}-\mathbf{SUL}-\mathbf{S}^{'}\mathbf{UL}^{'})\circ\mathbf{C}) \\ &+\beta\mathbf{X}^{'}((\mathbf{X}(\mathbf{W}-\bar{\mathbf{W}}))\circ(\mathbf{E}-\mathbf{C})) \\ &+\mathbf{X}^{'}(\mathbf{R}-\mathbf{C})=\mathbf{0}\end{aligned} \tag{4}$$

where $\mathbf{R} = (\mathbf{XW})\circ\mathbf{C}$. According to Theorem 1, we can rewrite Eq.4 as,

$$\begin{aligned}&(\mathbf{PH}+\beta\mathbf{QH}+\alpha\mathbf{PO}^{'}\mathbf{OP}^{'})\mathbf{vec}(\mathbf{W}) \\ &=\mathbf{H}^{'}\mathbf{vec}(\mathbf{C})+(\beta\mathbf{QH}-\alpha\mathbf{PO}^{'}\mathbf{OQ}^{'})\mathbf{vec}(\bar{\mathbf{W}})\end{aligned} \tag{5}$$

which is a simple and normal linear equation, and we employ the conjugate gradient algorithm (Møller 1993) which is known as a computationally efficient algorithm for solving linear equations.

Update $\bar{\mathbf{W}}$ with Fixed W and L

When $\mathbf{W}$ and $\mathbf{L}$ are fixed, similar to the case in the update of $\mathbf{W}$, we have the following equation for $\bar{\mathbf{W}}$ by setting the derivative w.r.t. $\bar{\mathbf{W}}$ to zero,

$$\begin{aligned}&\alpha\mathbf{X}^{'}((\mathbf{U}+\mathbf{S}^{'}\mathbf{SULL}^{'}-\mathbf{SUL}-\mathbf{S}^{'}\mathbf{UL}^{'})\circ(\mathbf{E}-\mathbf{C})) \\ &+\beta\mathbf{X}^{'}((\mathbf{X}(\bar{\mathbf{W}}-\mathbf{W}))\circ(\mathbf{E}-\mathbf{C})) \\ &+\zeta\mathbf{X}^{'}(\mathbf{T}+\mathbf{E}-\mathbf{C})=\mathbf{0}\end{aligned} \tag{6}$$

where $\mathbf{T} = (\mathbf{X}\bar{\mathbf{W}})\circ(\mathbf{E}-\mathbf{C})$. We rewrite Eq.6 as the following one with theorem 1,

$$\begin{aligned}&((\zeta+\beta)\mathbf{QH}+\alpha\mathbf{QO}^{'}\mathbf{OQ}^{'})\mathbf{vec}(\bar{\mathbf{W}}) \qquad (7)\\ &=-\zeta\mathbf{H}^{'}\mathbf{vec}(\mathbf{E}-\mathbf{C})+(\beta\mathbf{QH}-\alpha\mathbf{QO}^{'}\mathbf{OP}^{'})\mathbf{vec}(\mathbf{W})\end{aligned}$$

The efficient conjugate gradient algorithm is also employed for solving the above linear equations.

Update L with Fixed W and $\bar{\mathbf{W}}$

When $\mathbf{W}$ and $\bar{\mathbf{W}}$ are fixed, by setting the derivative of Eq.3 w.r.t. $\mathbf{L}$ to zero, we have the following equation for $\mathbf{L}$, i.e.,

$$(\mathbf{SU})^{'}(\mathbf{SU})\mathbf{L} = (\mathbf{SU})^{'}\mathbf{U} \qquad (8)$$

We have the following closed-form solution for $\mathbf{L}$ which is updated efficiently,

$$\mathbf{L} = \mathbf{Z}^{+}(\mathbf{SU})^{'}\mathbf{U} \qquad (9)$$

where $\mathbf{Z} = (\mathbf{SU})^{'}(\mathbf{SU})$ and $\mathbf{Z}^{+}$ indicates the pseudo inverse matrix of $\mathbf{Z}$. We use the optimized solution of Eq.3 as our final $\mathbf{L}$. After getting the coefficient matrix $\mathbf{W}$, we need to discretize the predictive label matrix. If $[\mathbf{XW}]_{ij} > 0$, we set $\widehat{Y}_{ij} = 1$, otherwise we set $\widehat{Y}_{ij} = -1$.

Experiments

In this section, we first give the experimental setup and then show the evaluation of our proposal compared to several state-of-the-art algorithms on a number of real-world tasks.

Experimental Setup

The proposed approach is compared with a number of methods, including the state-of-the-art weak-label learning method MLR-GL (Bucak, Jin, and Jain 2011), semi-supervised multi-label learning method SSML (Zhao and Guo 2015), a state-of-the-art supervised multi-label learning method ML-KNN (Zhang and Zhou 2007) and three naive methods that directly decompose the task into multiple binary classification problems via treating labels independently. Particularly, BSVM trains multiple supervised SVMs each for one label, which is the baseline method. WellSVM (Li et al. 2013) and S4VM (Li and Zhou 2015) are two promising binary semi-supervised SVMs. We further compare with a variant of our proposal that does not using the unlabeled instances. We call it as SSWL-wo. LIBSVM (Chang and Lin 2011) package is employed for the implementation for the BSVM method and RBF kernel with the recommended parameter is employed. For our SSWL method and the SSWL-wo method, 5 nearest neighbor graph is used for the instance matrix in all the experiments.

We measure the classification results in terms of three multi-label evaluation criteria that are both instance-wise and label-wise effective (Wu and Zhou 2017), i.e., Micro-F1, Macro-F1 and Hamming Loss (H.L.). Hamming Loss evaluates the fraction of misclassified instance-label pairs; Macro-F1 and Micro-F1 which take both precision and recall into account. The larger the value of Micro-F1 and Macro-F1, the better the performance. For hamming loss, the smaller the value, the better the performance. More details about the evaluation metric please refer to (Zhang and Zhou 2014).

For each dataset, we consider the incomplete label ratio (I. L. Ratio) by randomly dropping $\{0\%, 20\%, 40\%, 60\%\}$ of the observed labels on the labeled training data. We compared all methods using the same data setting for each data set. For all the methods, we conducted parameter selection for each evaluation metric by performing 5-fold cross-validation on the training set. For our approach, we selected the trade-off parameters α, β and ζ from $\{10^{-2}, \ldots, 10^{2}\}$. To reduce statistical variability, results are averaged over 10 independent repetitions.

Text Categorization Task

The text classification task is collected from SIAM Text Mining Competition (TMC). Each document is an aviation safety report documenting one or more problems that occurred on certain flights. The goal is to label the documents with respect to what types of problems they describe. Each document may belong to more than one class. *TMC* dataset (Srivastava and Zane-Ulman 2005) is a large text dataset with 28,596 instances and 22 labels in total. We used its short version, each instance contains 500 features. We randomly selected 1500 instances for training (500 labeled and 1000 unlabeled) and used the rest for testing.

Results are shown in Table 2. It can be seen that SSWL obtains quite promising performance. It achieves the best performance on 9 of 12 subtasks, while the other comparison methods have achieved the best performance on up to one subtask. SSWL-wo also obtains good performance but is not that good as SSWL. This suggests that the use of unlabeled data can help to further improve performance. Our proposal works better than state-of-the-art weak-label learning and semi-supervised multi-label learning algorithms. This shows that in semi-supervised weak-label learning, taking both semi-supervised and weak-label data into account is beneficial. The approaches with direct decomposition achieve sub-optimal performance.

Gene Function Analysis Task

The second task is to predict the gene function classes of the Yeast Saccharomyces cerevisiae, which is one of the best studied organisms. The Yeast data set (Elisseeff and Weston 2001)is a gene function classification dataset with 2417 examples and 14 class labels. Each gene is expressed with 103 microarray expression features. The average number of labels for each instance is 4.24±1.57. We randomly selected 1500 instances for training (500 labeled and 1000 unlabeled) and used the rest for testing.

Results are shown in Table 3. It can be seen that SSWL performs significantly better than the other approaches on three evaluation metrics. Moreover, SSWL also consistently perform robustly as the ratio of missing labels changes. This result further verifies that it is important to take both semi-supervised and weak-label data into account to handling semi-supervised weak-label learning.

Table 2: Experimental results (mean±std) on *TMC*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.640 ± .001**	.639 ± .001	.612 ± .003	.615 ± .001	.638 ± .001	.501 ± .001	.578 ± .002	.487 ± .002
	20%	**.602 ± .003**	.578 ± .001	.556 ± .002	.596 ± .002	.580 ± .002	.213 ± .002	.506 ± .001	.292 ± .001
	40%	**.582 ± .001**	.455 ± .004	.356 ± .002	.461 ± .003	.423 ± .001	.032 ± .001	.365 ± .003	.023 ± .002
	60%	**.570 ± .002**	.505 ± .001	.113 ± .002	.563 ± .002	.160 ± .022	.012 ± .001	.215 ± .002	.007 ± .003
Macro-F1(↑)	0%	.618 ± .002	**.620 ± .003**	.586 ± .002	.588 ± .002	.613 ± .002	.467 ± .001	.545 ± .002	.464 ± .002
	20%	**.582 ± .001**	.568 ± .001	.519 ± .001	.567 ± .001	.543 ± .002	.175 ± .001	.457 ± .002	.244 ± .001
	40%	**.566 ± .003**	.409 ± .005	.295 ± .003	.413 ± .002	.368 ± .002	.024 ± .001	.309 ± .002	.017 ± .001
	60%	**.553 ± .002**	.494 ± .001	.089 ± .005	.537 ± .001	.125 ± .029	.008 ± .001	.279 ± .002	.004 ± .001
H.L.(↓)	0%	**.065 ± .002**	.069 ± .001	.067 ± .002	.076 ± .001	.067 ± .001	.082 ± .002	.085 ± .002	.080 ± .001
	20%	.075 ± .002	.075 ± .001	.072 ± .001	.078 ± .002	**.071 ± .002**	.092 ± .002	.087 ± .005	.088 ± .001
	40%	**.079 ± .001**	.082 ± .002	.083 ± .002	.086 ± .002	.080 ± .001	.099 ± .001	.092 ± .002	.100 ± .001
	60%	.087 ± .003	.089 ± .002	.097 ± .003	**.084 ± .002**	.095 ± .003	.101 ± .002	.111 ± .003	.101 ± .002

Table 3: Experimental results (mean±std) on *yeast*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.647 ± .001**	.619 ± .002	.612 ± .002	.623 ± .003	.584 ± .001	.625 ± .002	.592 ± .001	.623 ± .002
	20%	**.638 ± .001**	.626 ± .002	.534 ± .003	.618 ± .005	.510 ± .002	.506 ± .001	.511 ± .001	.509 ± .001
	40%	**.604 ± .002**	.554 ± .003	.394 ± .003	.379 ± .004	.152 ± .001	.103 ± .002	.432 ± .001	.188 ± .003
	60%	**.616 ± .002**	.568 ± .002	.241 ± .002	.209 ± .002	.046 ± .031	.002 ± .002	.320 ± .002	.019 ± .007
Macro-F1(↑)	0%	**.635 ± .001**	.594 ± .001	.582 ± .003	.600 ± .001	.557 ± .002	.602 ± .001	.578 ± .002	.592 ± .002
	20%	**.618 ± .001**	.613 ± .001	.494 ± .002	.593 ± .004	.476 ± .001	.470 ± .001	.476 ± .002	.478 ± .002
	40%	**.574 ± .001**	.538 ± .005	.359 ± .004	.340 ± .002	.126 ± .001	.083 ± .001	.397 ± .001	.145 ± .002
	60%	**.595 ± .002**	.554 ± .001	.194 ± .003	.177 ± .002	.039 ± .025	.001 ± .001	.280 ± .005	.016 ± .003
H.L.(↓)	0%	**.207 ± .001**	.209 ± .001	.211 ± .001	.213 ± .002	.215 ± .002	.208 ± .001	.214 ± .001	.209 ± .001
	20%	**.210 ± .001**	.216 ± .001	.221 ± .002	.211 ± .002	.224 ± .002	.225 ± .001	.253 ± .002	.224 ± .001
	40%	**.225 ± .001**	.252 ± .004	.246 ± .003	.251 ± .001	.286 ± .002	.294 ± .001	.257 ± .001	.279 ± .002
	60%	**.231 ± .003**	.268 ± .003	.278 ± .002	.275 ± .003	.299 ± .010	.305 ± .001	.286 ± .002	.302 ± .002

Scene Classification Task

The third task is a scene classification problem. In natural scene classification, each scene image may belong to several classes simultaneously. Through analyzing images with known label sets, a multi-label learning method will automatically predict the sets of labels for unseen images. The above process of semantic scene classification can be applied to many areas. The Scene Image data set (Zhang and Zhou 2007) contains 2,000 natural scene images and 5 labels. Each image has on average 1.24 ± 0.44 labels and is represented as 294-dimensional vector. We randomly selected 1500 instances for training (500 labeled and 1000 unlabeled) and the rest ones are used for testing.

Results in Table 4 show that SSWL also achieve highly competitively performance with compared methods. It also clearly shows the advantage of our proposed approach which is able to exploit unlabeled data, as well as weak-label learning to enhance the performance.

Image Annotation Task

The last task is an image annotation problem. A key issue in image annotation is the correlations among the labels. The labels do not exist in isolation. We use the Microsoft Research image annotation data set (msrc) to verify that our method can help predict. *msrc* is a labeled image dataset with 591 images in 23 object classes. Each image has on average 2.51 ± 1.22 labels and is represented as a vector with 960 GIST features (Oliva and Torralba 2001). We randomly selected 80% of the data for training (30% labeled and 50% unlabeled) and used the rest 20% for testing.

Results are shown in Table 5. Overall, these results demonstrate the benefit of handling incomplete labels in the learning process. It also clearly shows the advantage of our proposed approach which is able to exploit unlabeled data, as well as weak-label learning to enhance the performance. Our algorithm usually converges quickly. Again, the results demonstrate the effectiveness of our proposal algorithm.

Table 4: Experimental results (mean±std) on *SceneImage*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	.589 ± .003	.583 ± .002	**.613** ± **.002**	.528 ± .002	.499 ± .002	.456 ± .001	.489 ± .001	.538 ± .002
	20%	**.572** ± **.002**	.545 ± .001	.558 ± .003	.419 ± .001	.394 ± .002	.389 ± .002	.458 ± .003	.392 ± .001
	40%	**.540** ± **.002**	.534 ± .002	.395 ± .001	.220 ± .001	.174 ± .001	.094 ± .001	.289 ± .001	.205 ± .001
	60%	**.521** ± **.003**	.517 ± .002	.251 ± .002	.000 ± .001	.019 ± .010	.000 ± .000	.300 ± .001	.010 ± .002
Macro-F1(↑)	0%	**.576** ± **.002**	.553 ± .001	.567 ± .001	.454 ± .003	.407 ± .001	.362 ± .001	.466 ± .001	.437 ± .001
	20%	**.550** ± **.003**	.505 ± .001	.494 ± .002	.320 ± .001	.295 ± .002	.295 ± .002	.417 ± .001	.284 ± .001
	40%	**.523** ± **.002**	.499 ± .001	.306 ± .001	.140 ± .001	.113 ± .001	.056 ± .001	.207 ± .002	.134 ± .002
	60%	**.510** ± **.004**	.495 ± .001	.165 ± .002	.000 ± .001	.011 ± .005	.000 ± .000	.230 ± .001	.005 ± .003
H.L.(↓)	0%	.184 ± .001	.186 ± .001	**.167** ± **.002**	.192 ± .001	.193 ± .001	.192 ± .002	.245 ± .002	.167 ± .001
	20%	.199 ± .001	.201 ± .001	**.182** ± **.002**	.199 ± .001	.203 ± .001	.204 ± .002	.236 ± .002	.194 ± .001
	40%	**.206** ± **.001**	.214 ± .001	**.206** ± **.001**	.219 ± .001	.227 ± .001	.240 ± .002	.226 ± .001	.221 ± .001
	60%	**.208** ± **.002**	.222 ± .001	.220 ± .002	.249 ± .001	.247 ± .003	.250 ± .002	.229 ± .002	.248 ± .002

Table 5: Experimental results (mean±std) on *msrc*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	.590 ± .003	.550 ± .003	.505 ± .003	.533 ± .003	**.604** ± **.002**	.355 ± .001	.351 ± .001	.487 ± .001
	20%	**.571** ± **.001**	.513 ± .003	.444 ± .001	.472 ± .002	.569 ± .003	.307 ± .001	.331 ± .002	.292 ± .002
	40%	.507 ± .001	.449 ± .005	.216 ± .002	.500 ± .001	**.533** ± **.002**	.122 ± .002	.316 ± .001	.047 ± .003
	60%	**.465** ± **.001**	.367 ± .002	.118 ± .003	.244 ± .001	.462 ± .001	.034 ± .001	.294 ± .001	.007 ± .001
Macro-F1(↑)	0%	.562 ± .003	.505 ± .004	.412 ± .002	.465 ± .004	**.612** ± **.001**	.291 ± .002	.378 ± .001	.464 ± .001
	20%	**.522** ± **.001**	.419 ± .004	.302 ± .002	.390 ± .002	.518 ± .002	.216 ± .001	.369 ± .002	.244 ± .002
	40%	**.491** ± **.002**	.396 ± .004	.151 ± .003	.424 ± .002	.472 ± .001	.072 ± .002	.317 ± .002	.032 ± .002
	60%	**.430** ± **.001**	.298 ± .001	.082 ± .002	.166 ± .001	.391 ± .002	.037 ± .001	.282 ± .001	.004 ± .001
H.L.(↓)	0%	**.083** ± **.001**	.085 ± .001	**.083** ± **.002**	.093 ± .002	.108 ± .002	.094 ± .001	.154 ± .001	.092 ± .001
	20%	**.085** ± **.001**	.091 ± .001	.086 ± .001	.101 ± .001	.105 ± .002	.095 ± .001	.163 ± .002	.088 ± .001
	40%	.090 ± .001	.098 ± .002	.098 ± .003	**.083** ± **.002**	.100 ± .001	.101 ± .001	.182 ± .002	.100 ± .002
	60%	**.092** ± **.001**	.099 ± .001	.103 ± .002	.100 ± .002	.094 ± .001	.106 ± .001	.201 ± .001	.101 ± .001

Conclusion

In this paper, we consider semi-supervised weak-label learning problem where the relevant label sets of instances are not only partially known, but also probably completely unknown. This is a new kind of multi-label learning problem that to the best of our knowledge, has not been thoroughly studied before. To address this problem, we propose the SSWL method. Both instance similarity and label similarity are considered for the complement of missing labels. Moreover, ensemble of multiple models is employed which is more robust than a single model when the label information is insufficient. We formulate the objective as a bi-convex optimization and present an efficient block coordinate descend solution. Experiments on a number of real tasks validate the effectiveness of SSWL in handling the semi-supervised weak-label learning problem.

There are many interesting future works. For example, our current proposal adopts a learning method for the label similarity where some prior knowledge may be not leveraged. More flexible methods that is able to incorporate the domain knowledge are worth trying in the future. Moreover, the study for transductive weak-label problem is an interesting issue in the future.

References

Bucak, S. S.; Jin, R.; and Jain, A. K. 2011. Multi-label learning with incomplete class assignments. In *Proceedings of 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR11), Colorado Springs, CO*, 2801–2808.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27.

Chapelle, O.; Schölkopf, B.; and Zien, A., eds. 2006. *Semi-Supervised Learning*. MIT Press.

Chen, G.; Song, Y.; Wang, F.; and Zhang, C. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In *Proceedings of the SIAM International Conference on Data Mining (SDM08), Atlanta, GA*, 410–419.

Elisseeff, A., and Weston, J. 2001. A kernel method for multi-labelled classification. *Advances in Neural Information Processing Systems 14 (NIPS01), Vancouver, Canada* 681–687.

Gorski, J.; Pfeuffer, F.; and Klamroth, K. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research* 66(3):373–407.

Guo, Y., and Schuurmans, D. 2012. Semi-supervised multi-label classification. *Machine Learning and Knowledge Discovery in Databases* 355–370.

Horn, R. A., and Johnson, C. R. 1991. Topics in matrix analysis. *Cambridge University Presss, Cambridge* 37:39.

Kong, X.; Ng, M. K.; and Zhou, Z.-H. 2013. Transductive multilabel learning via label set propagation. *IEEE Transactions on Knowledge and Data Engineering* 25(3):704–719.

Li, Y.-F., and Zhou, Z.-H. 2015. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(1):175–188.

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

Lin, Z.; Ding, G.; Hu, M.; Wang, J.; and Ye, X. 2013. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR13), Portland, OR*, 1618–1625.

Liu, Y.; Jin, R.; and Yang, L. 2006. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI06), Boston, MA*, 421–426.

Møller, M. F. 1993. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* 6(4):525–533.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Sindhwani, V.; Niyogi, P.; and Belkin, M. 2005. A co-regularization approach to semi-supervised learning with multiple views. In *ICML05 Workshop on Learning with Multiple Views, Bonn, Germany*, 74–79.

Srivastava, A. N., and Zane-Ulman, B. 2005. Discovering recurring anomalies in text reports regarding complex space systems. In *IEEE Aerospace Conference, Big Sky, MT*, 3853–3862.

Sun, Y.-Y., and Zhang, Yin an Zhou, Z.-H. 2010. Multi-label learning with weak label. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI10), Atlanta, GA*, 593–598.

Tseng, P. 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* 109(3):475–494.

Wang, B.; Tu, Z.; and Tsotsos, J. K. 2013. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV13), Sydney, Australia*, 425–432.

Weber, R.; Schek, H.-J.; and Blott, S. 1998. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB98), New York City, NY*, 194–205.

Wu, X.-Z., and Zhou, Z.-H. 2017. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning (ICML17), Sydney, Australia*, 3780–3788.

Wu, L.; Jin, R.; and Jain, A. K. 2013. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(3):716–727.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup matrix completion with side information: Application to multi-label learning. *Advances in Neural Information Processing Systems 26 (NIPS13), Lake Tahoe, NE* 2301–2309.

Yang, S.-J.; Jiang, Y.; and Zhou, Z.-H. 2013. Multi-instance multi-label learning with weak label. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI13), Beijing, China*, 1862–1868.

Zhan, W., and Zhang, M.-L. 2017. Inductive semi-supervised multi-label learning with co-training. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD17), Halifax, Canada*, 1305–1314.

Zhang, M.-L., and Zhou, Z.-H. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.

Zhang, Y., and Zhou, Z.-H. 2010. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data* 4(3):14.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhao, F., and Guo, Y. 2015. Semi-supervised multi-label learning with incomplete labels. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI15), Buenos Aires, Argentina*, 4062–4068.

Zhou, Z.-H. 2012. *Ensemble methods: foundations and algorithms*. CRC press.

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review*.

Zhu, X., and Goldberg, A. B. 2009. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers.

Zhu, S.; Ji, X.; Xu, W.; and Gong, Y. 2005. Multi-labelled classification using maximum entropy method. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR05), New York City, NY*, 274–281.

Zhu, G.; Yan, S.; and Ma, Y. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the 18th ACM International Conference on Multimedia (ACMMM10), Firenze, Italy*, 461–470.