Partial Multi-Label Learning with Label Distribution

Ning Xu, Yun-Peng Liu, Xin Geng*
MOE Key Laboratory of Computer Network and Information Integration, China
School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
{xning, yunpengliu, xgeng}@seu.edu.cn

Abstract

Partial multi-label learning (PML) aims to learn from training examples each associated with a set of candidate labels, among which only a subset are valid for the training example. The common strategy to induce predictive model is trying to disambiguate the candidate label set, such as identifying the ground-truth label via utilizing the confidence of each candidate label or estimating the noisy labels in the candidate label sets. Nonetheless, these strategies ignore considering the essential *label distribution* corresponding to each instance since the label distribution is not explicitly available in the training set. In this paper, a new partial multi-label learning strategy named PML-LD is proposed to learn from partial multi-label examples via *label enhancement*. Specifically, label distributions are recovered by leveraging the topological information of the feature space and the correlations among the labels. After that, a multi-class predictive model is learned by fitting a regularized multi-output regressor with the recovered label distributions. Experimental results on synthetic as well as real-world datasets clearly validate the effectiveness of PML-LD for solving PML problems.

Introduction

Partial multi-label learning deals with the problem where each training example is associated with a set of candidate labels, among which only a subset correspond to the ground-truth labels. In recent years, the need to learn from data with partial multi-labels naturally arises in many real-world applications (Zhou 2018; Xie and Huang 2018). For instance, in online object annotation (Figure 1), only some of the candidate labels given by the annotators are valid due to the potential unreliable annotators. Partial multi-label learning aims to induce a multi-label classifier from PML training examples, which can assign a set of proper labels for the unseen instance.

Formally, let $\mathcal{X} = \mathbb{R}^q$ be the q-dimensional feature space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ be the output space with c possible class labels. Given the PML training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq m\}$, the task of PML is to induce a multi-label predictor $f : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ from $\mathcal{D}$. Here, $\boldsymbol{x}_i \in \mathcal{X}$ is a q-dimensional feature vector and $Y_i \subseteq \mathcal{Y}$ is the set of candidate labels associated with $\boldsymbol{x}_i$. Partial multi-label learning takes the key assumption that the ground-truth labels $\widetilde{Y}_i \subseteq \mathcal{Y}$ corresponding to $\boldsymbol{x}_i$ reside in its candidate label set Y_i, i.e. $\widetilde{Y}_i \subseteq Y_i$, and therefore cannot be directly accessed by the learning algorithm. Intuitively, the basic strategy for coping with the PML problem is disambiguation, i.e. identifying the ground-truth labels from the candidate label sets. One recent attempt is utilizing the confidence of each candidate label being the ground-truth one (Xie and Huang 2018). Nonetheless, the confidence scores would be error-prone especially with the high proportion of false positive labels since it ignores the irrelevance of the non-candidate labels. Low-rank assumption is adopted to identify the noisy labels for disambiguation (Yu et al. 2018; Sun et al. 2019). For credible label elicitation techniques, the ground-truth labels are identified from the candidate label set to make final prediction on unseen instance (Fang and Zhang 2019).

In order to handle the ambiguity in partial multi-label learning, we can explicitly assign a *description degree* to each label instead of disambiguation. The description degrees $d_{\boldsymbol{x}}^{y_j}$ of all the labels constitute a real-valued vector

Figure 1: An example of partial multi-label learning. In online object annotation, among the set of five candidate labels given by the annotators, only three of them are valid ones (in red) including *house*, *mountain* and *tree*.

*Corresponding author

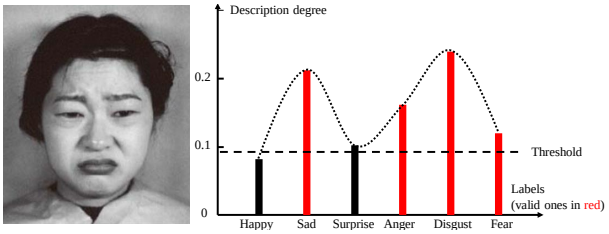


Figure 2: An example about the differentiation between candidate labels and non-candidate labels in PML

called *label distribution* (Geng 2016), which describes the instance more comprehensively than logical labels. Here $d_{\boldsymbol{x}}^{y_j} \in [0,1]$ and $\sum_y d_{\boldsymbol{x}}^{y} = 1$. Note that label distributions are more essential than logical labels in partial multi-label learning problems because the relevance or irrelevance of a label to an instance is relative in mainly three aspects:

- The differentiation between candidate labels and non-candidate labels is relative. In partial multi-label learning, the boundary between relevant and irrelevant labels is not clear, which may result in the partition to assign some irrelevant labels into the candidate label set. For example, in the facial expressions annotation (Figure 2), a facial expression often conveys a complex mixture of basic emotions (Zhou, Xue, and Geng 2015). The threshold chosen by an unreliable annotator leads to the candidate label set (e.g., *sad*, *surprise*, *anger*, *disgust* and *fear*) where *surprise* is not valid.

- The relevance among candidate labels is different rather than exactly equal. For example, a natural scene image may be annotated with the candidate labels sky, water, building and cloud simultaneously, but the relevance of each label to this image is different.

- The irrelevance of each non-candidate label may be very different. For example, for a car, the label airplane is more irrelevant than the label tank.

Although label distributions are not explicitly available in the partial multi-label training sets, they can be somehow recovered from the training set, a process named *label enhancement* (Xu, Tao, and Geng 2018). Accordingly, a novel partial multi-label learning algorithm named PML-LD, i.e., *Partial Multi-label Learning with Label Distribution*, is proposed in this paper. PML-LD recovers label distributions via leveraging the topological information of the feature space and the correlations among the labels. After that, a multi-class predictive model is learned by fitting a regularized multi-output regressor with the recovered label distributions.

The rest of this paper is organized as follows. Firstly, related works on partial multi-label learning are briefly reviewed. Secondly, technical details of the proposed approach are introduced. Thirdly, the results of the comparative experiments are reported. Finally, we conclude this paper.

Related Work

Partial multi-label learning is closely related to two popular learning frameworks, namely *multi-label learning* (Zhang and Zhou 2014; Gibaja and Ventura 2015; Zhou and Zhang 2017) and *partial label learning* (Cour, Sapp, and Taskar 2011; Liu and Dietterich 2012; Zhang, Yu, and Tang 2017).

In multi-label learning (MLL), each example is associated with multiple valid labels simultaneously. Based on the order of label correlations (Zhang and Zhou 2014) exploited for model training, multi-label learning approaches can be roughly grouped into three types. The simplest one is the first-order type which decomposes the problem into a series of binary classification problems, each for one label (Boutell et al. 2004; Zhang and Zhou 2007). The first-order approaches neglect the fact that the information of one label may be helpful for the learning of another label. The second-order approaches consider the correlations between pairs of class labels (Elisseeff and Weston 2002; Fürnkranz et al. 2008). But the second-order approaches such as CLR (Fürnkranz et al. 2008) and RankSVM (Elisseeff and Weston 2002) only focus on the difference between relevant label and irrelevant label. The high-order approaches consider the correlations among label subsets or all the class labels (Read et al. 2011; Tsoumakas, Katakis, and Vlahavas 2011). Both MLL and PML aim to induce the predictive model which can assign proper label set for unseen instance. Nonetheless, the task of PML is more challenging than MLL as the ground-truth is not directly accessible to PML learning algorithm.

In partial label learning (PLL), each example is associated with multiple candidate labels among which only one is valid. The task of partial label learning is to induce a multi-class predictive model which can assign one proper label for unseen instance, where existing PLL approaches work by disambiguating the candidate label set (Cour, Sapp, and Taskar 2011; Yu and Zhang 2017) or transforming partial label learning problem into canonical supervised learning problems (Zhang, Yu, and Tang 2017). Both PLL and PML learn from training examples with labeling noise where false positive labels reside in the candidate label set. Nonetheless, the task of PML is more challenging than PLL as a multi-label predictor rather than single-label predictor needs to be induced from PML training examples.

To solve the partial multi-label learning problem, the basic strategy for coping with the PML problem is disambiguation, i.e. identifying the ground-truth labels from the candidate label sets. One recent attempt is utilizing the confidence of each candidate label being the ground-truth one (Xie and Huang 2018). Nonetheless, the confidence scores would be error-prone especially with the high proportion of false positive labels since it ignores the irrelevance of the non-candidate labels. Low-rank assumption is adopted to identify the noisy labels for disambiguation (Yu et al. 2018; Sun et al. 2019). For credible label elicitation techniques, the ground-truth labels are identified from the candidate label set to make final prediction on unseen instance (Fang and Zhang 2019).

In the next section, a novel partial multi-label learning approach will be introduced. Different from existing partial

multi-label learning approaches, the label distributions are recovered and utilized to facilitate the learning procedure.

The Proposed Approach

Label Distribution Estimation

For each PML example $(\boldsymbol{x}, Y)$, let $\boldsymbol{l} = [l_{\boldsymbol{x}}^{y_1}, l_{\boldsymbol{x}}^{y_2}, ..., l_{\boldsymbol{x}}^{y_c}]^\top$ denote the c-dimensional logical vector w.r.t. the candidate label set: $l_{\boldsymbol{x}}^{y_i} = 1$ if $y_i \in Y$, otherwise $l_{\boldsymbol{x}}^{y_i} = 0$. Then, the logical label matrix $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, ..., \boldsymbol{l}_n]$ are constructed. Our aim is to recover the label distribution matrix $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_n]$ from the logical label matrix $\boldsymbol{L}$. To solve this problem, PML-LD assumes the parametric model

$$\boldsymbol{d}_i = \boldsymbol{W}^\top \varphi(\boldsymbol{x}_i) + \boldsymbol{s} = \hat{\boldsymbol{W}} \boldsymbol{\phi}_i, \tag{1}$$

where $\boldsymbol{W} = [\boldsymbol{w}^1, ..., \boldsymbol{w}^c]$ is a weight matrix and $\boldsymbol{s} \in \mathbb{R}^c$ is a bias vector. $\varphi(\boldsymbol{x})$ is a nonlinear transformation of $\boldsymbol{x}$ to a higher dimensional feature space. For convenient describing, $\hat{\boldsymbol{W}} = [\boldsymbol{W}^\top, \boldsymbol{s}]$ and $\boldsymbol{\phi}_i = [\varphi(\boldsymbol{x}_i); 1]$ are set. Accordingly, the goal of our method is to determine optimal model $\hat{\boldsymbol{W}}^*$ which minimizes

$$\hat{\boldsymbol{W}}^* = \arg\min_{\hat{\boldsymbol{W}}} \ L(\hat{\boldsymbol{W}}) + \lambda_1 Z(\hat{\boldsymbol{W}}) + \lambda_2 \Omega(\hat{\boldsymbol{W}}), \tag{2}$$

where L is a loss function, Ω is the functions to leverage the topological information of the feature space, and Z is the function to leverage the correlation among the labels. Note that label enhancement is essentially a pre-processing applied to the training set, which is different from standard supervised learning. Therefore, our optimization does not need to consider the overfitting problem.

Since the information in the label distributions is inherited from the initial logical labels, $L(\hat{\boldsymbol{W}})$ in Eq. (2) is defined as the least squares (LS) loss function

$$L(\hat{\boldsymbol{W}}) = \| \hat{\boldsymbol{W}} \boldsymbol{\Phi} - \boldsymbol{L} \|_F^2, \tag{3}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_n]$.

The local label correlations (Tsoumakas et al. 2009) are considered to provide helpful extra information to recover the label distributions from multi-labels. Specifically, the more correlative two labels are, the closer the corresponding description degrees should be. In other words, $\boldsymbol{d}^i$ should more be more similar to $\boldsymbol{d}^j$ if the i-th and j-th labels are more correlated. Here $\boldsymbol{d}^i$ is the vector constituted by all the description degrees of the i-th label, i.e., $\boldsymbol{d}^i = [d_{\boldsymbol{x}_1}^{y_i}, d_{\boldsymbol{x}_2}^{y_i}, ..., d_{\boldsymbol{x}_n}^{y_i}]$. Assuming that the training data can be separated into m groups $\{G^{(1)}, G^{(2)}, ..., G^{(m)}\}$, instances in the same group share the same subset of label correlations. Then the local label correlations are measured by the label correlation matrix $\boldsymbol{R}^{(k)}$ whose elements are $r_{ij}^{(k)}$. Therefore, $Z(\hat{\boldsymbol{W}})$ in Eq. (2) is defined as:

$$\begin{aligned} Z(\hat{\boldsymbol{W}}) &= \sum_k \sum_{i,j} r_{ij}^{(k)} \| \boldsymbol{d}^{i(k)} - \boldsymbol{d}^{j(k)} \|^2 \\ &= \sum_k \mathrm{tr}(\boldsymbol{\Phi}^{(k)\top} \hat{\boldsymbol{W}}^\top \boldsymbol{C}^{(k)} \hat{\boldsymbol{W}} \boldsymbol{\Phi}^{(k)}), \end{aligned} \tag{4}$$

where $\boldsymbol{d}^{(k)}$ is the label distributions corresponding to all the instance in $G^{(k)}$, $\boldsymbol{\Phi}^{(k)}$ is the feature matrix representing the higher dimensional features to the instance in $G^{(k)}$, $\boldsymbol{C}^{(k)} = \hat{\boldsymbol{R}}^{(k)} - \boldsymbol{R}^{(k)}$ is the Laplacian matrix, and $\hat{\boldsymbol{R}}^{(k)}$ is the diagonal matrix whose elements are $\hat{r}_{ii}^{(k)} = \sum_{j=1}^{n} r_{ij}^{(k)}$.

According to the smoothness assumption (Zhu, Lafferty, and Rosenfeld 2005), the points close to each other are more likely to share a label. Intuitively, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ have a high degree of similarity, as denoted by a_{ij}, then $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ should be near to one another. Therefore, the hidden label distributions can be mined from the training examples by leveraging the topological information of the feature space (Ning, An, and Xin 2018), which leads to the following function $\Omega(\hat{\boldsymbol{W}})$ in Eq. (2):

$$\begin{aligned} \Omega(\hat{\boldsymbol{W}}) &= \sum_{i,j} a_{ij} \| \boldsymbol{d}_i - \boldsymbol{d}_j \|^2 \\ &= \mathrm{tr}(\hat{\boldsymbol{W}} \boldsymbol{\Phi} \boldsymbol{G} \boldsymbol{\Phi}^\top \hat{\boldsymbol{W}}^\top), \end{aligned} \tag{5}$$

where each element a_{ij} in the local similarity matrix $\boldsymbol{A}$ can be calculated by $a_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2}\right)$ if $\boldsymbol{x}_i$ is among K-nearest neighbors of $\boldsymbol{x}_j$, otherwise $a_{ij} = 0$. Here K is set to be $c+1$. $\boldsymbol{G} = \hat{\boldsymbol{A}} - \boldsymbol{A}$ is the graph Laplacian and $\hat{\boldsymbol{A}}$ is the diagonal matrix whose elements are $\hat{a}_{ii} = \sum_{j=1}^{n} a_{ij}$.

Formulating the label enhancement problem into an optimization framework over Eq. (3), Eq. (5) and Eq. (4), the following optimization problem is obtained:

$$\begin{aligned} \min_{\hat{\boldsymbol{W}}} \quad & \| \hat{\boldsymbol{W}} \boldsymbol{\Phi} - \boldsymbol{L} \|_F^2 + \lambda_1 \sum_{k=1}^{m} \mathrm{tr}(\boldsymbol{\Phi}^{(k)\top} \hat{\boldsymbol{W}}^\top \boldsymbol{C}^{(k)} \hat{\boldsymbol{W}} \boldsymbol{\Phi}^{(k)}) \\ & + \lambda_2 \mathrm{tr}(\hat{\boldsymbol{W}} \boldsymbol{\Phi} \boldsymbol{G} \boldsymbol{\Phi}^\top \hat{\boldsymbol{W}}^\top). \end{aligned} \tag{6}$$

In this paper, instead of specifying any label correlation matrix, each Laplacian matrix $\boldsymbol{C}^{(k)}$ is learned directly. Note that optimization w.r.t. $\boldsymbol{C}^{(k)}$ may lead to the trivial solution $\boldsymbol{C}^{(k)} = \boldsymbol{0}$. To avoid the problems, $\boldsymbol{C}^{(k)}$ is decomposed as $\boldsymbol{E}^{(k)} \boldsymbol{E}^{(k)\top}$ and the constrain $\mathrm{diag}(\boldsymbol{E}^{(k)} \boldsymbol{E}^{(k)\top}) = \boldsymbol{1}$ is added. Then the following formulation is obtained:

$$\begin{aligned} \min_{\hat{\boldsymbol{W}}, \boldsymbol{E}} \quad & \| \hat{\boldsymbol{W}} \boldsymbol{\Phi} - \boldsymbol{L} \|_F^2 + \lambda_2 \mathrm{tr}(\hat{\boldsymbol{W}} \boldsymbol{\Phi} \boldsymbol{G} \boldsymbol{\Phi}^\top \hat{\boldsymbol{W}}^\top) \\ & + \lambda_1 \sum_{k=1}^{m} \mathrm{tr}(\boldsymbol{\Phi}^{(k)\top} \hat{\boldsymbol{W}}^\top \boldsymbol{E}^{(k)} \boldsymbol{E}^{(k)\top} \hat{\boldsymbol{W}} \boldsymbol{\Phi}^{(k)}) \\ \text{s.t.} \quad & \mathrm{diag}(\boldsymbol{E}^{(k)} \boldsymbol{E}^{(k)\top}) = \boldsymbol{1}, k = 1, 2, ..., m. \end{aligned} \tag{7}$$

If the best parameter $\hat{\boldsymbol{W}}^*$ is determined, the label distribution $\boldsymbol{d}_i$ can be generated through Eq. (1). Finally, $\boldsymbol{d}_i$ is normalized via the softmax normalization.

The Alternating Solution We solve the optimization problem in Eq. (7) in an alternating way, i.e., optimizing one of the two variables with the other fixed. When $\hat{\boldsymbol{W}}$ is fixed to solve $\boldsymbol{E}$, Eq. (7) can be reduced to m optimization problems, where the i-th one is:

$$\begin{aligned} \min_{\boldsymbol{E}^{(k)}} \quad & \mathrm{tr}(\boldsymbol{\Phi}^{(k)\top} \hat{\boldsymbol{W}}^\top \boldsymbol{E}^{(k)} \boldsymbol{E}^{(k)\top} \hat{\boldsymbol{W}} \boldsymbol{\Phi}^{(k)}) \\ \text{s.t.} \quad & \mathrm{diag}(\boldsymbol{E}^{(k)} \boldsymbol{E}^{(k)\top}) = \boldsymbol{1}. \end{aligned} \tag{8}$$

The optimization of Eq. (8) uses projected gradient descent. The gradient of the objective w.r.t. $\boldsymbol{E}_i$ is

$$\nabla_{\boldsymbol{E}^{(k)}} = 2\hat{\boldsymbol{W}}\boldsymbol{\Phi}^{(k)}\boldsymbol{\Phi}^{(k)^\top}\hat{\boldsymbol{W}}^\top\boldsymbol{E}^{(k)}. \tag{9}$$

To satisfy the constraint $\mathrm{diag}(\boldsymbol{E}^{(k)}\boldsymbol{E}^{(k)\top}) = \mathbf{1}$, each row of $\boldsymbol{E}^{(k)}$ is projected onto the unit norm ball after each update

$$\boldsymbol{e}_i^{(k)} \leftarrow \frac{\boldsymbol{e}_i^{(k)}}{\|\boldsymbol{e}_i^{(k)}\|}, \tag{10}$$

where $\boldsymbol{e}_i^{(k)}$ is the i-th row of $\boldsymbol{E}^{(k)}$.

When $\boldsymbol{E}$ is fixed to solve $\hat{\boldsymbol{W}}$, the task becomes:

$$\begin{aligned}\min_{\hat{\boldsymbol{W}}} \quad & \|\hat{\boldsymbol{W}}\boldsymbol{\Phi} - \boldsymbol{L}\|_F^2 + \lambda_2 \mathrm{tr}(\hat{\boldsymbol{W}}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{\Phi}^\top\hat{\boldsymbol{W}}^\top) \\ & + \lambda_1 \sum_{k=1}^{m} \mathrm{tr}(\boldsymbol{\Phi}^{(k)\top}\hat{\boldsymbol{W}}^\top\boldsymbol{E}^{(k)}\boldsymbol{E}^{(k)\top}\hat{\boldsymbol{W}}\boldsymbol{\Phi}^{(k)}).\end{aligned} \tag{11}$$

The optimization of Eq. (11) uses an effective quasi-Newton method BFGS (Nocedal and Wright 2006). As to the optimization of the target function $T(\hat{\boldsymbol{W}})$, the computation of BFGS is mainly related to the first-order gradient, which can be obtained through

$$\begin{aligned}\nabla_{\hat{\boldsymbol{W}}} =& 2\hat{\boldsymbol{W}}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top - 2\boldsymbol{L}\boldsymbol{\Phi}^\top + \lambda_2\hat{\boldsymbol{W}}\boldsymbol{\Phi}\boldsymbol{G}^\top\boldsymbol{\Phi}^\top + \lambda_2\hat{\boldsymbol{W}}\boldsymbol{\Phi}\boldsymbol{G}\boldsymbol{\Phi}^\top \\ & + 2\lambda_1 \sum_{k=1}^{m} (\boldsymbol{E}^{(k)}\boldsymbol{E}^{(k)^\top}\hat{\boldsymbol{W}}\boldsymbol{\Phi}^{(k)}\boldsymbol{\Phi}^{(k)^\top}).\end{aligned} \tag{12}$$

Predictive Model Induction

Following the first stage of label distribution recovery, the original PML training set $\mathcal{D}$ has been transformed into its essential counterpart: $\mathcal{E} = \{(\boldsymbol{x}_i, \boldsymbol{d}_i) | 1 \le i \le n\}$. In the second stage, PML-LD aims to induce the predictive model $f : \mathcal{X} \mapsto \mathcal{Y}$ based on $\mathcal{E}$. Considering that $\boldsymbol{d}_i$ for each training example in $\mathcal{E}$ are actually real-valued, it is natural to induce the predictive model by employing multi-output regression techniques. Similar to the MSVR (Chung et al. 2015; Sánchez-Fernández et al. 2004), we generalize a regressor to solve the multi-dimensional case. Then, PML-LD induces the regression model by minimizing the following loss function:

$$\Omega(\boldsymbol{\Theta}, \boldsymbol{b}) = \frac{1}{2}\sum_{j=1}^{c}\|\boldsymbol{\theta}^j\|^2 + \beta_1 \sum_{i=1}^{n}\Omega_{1i} + \beta_2 \sum_{i=1}^{n}\Omega_{2i}, \tag{13}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^c]$, $\boldsymbol{b} = [b^1, ..., b^c]$, Ω_1 and Ω_2 are the regression loss and the sign loss, respectively.

As shown in Eq.(13), the first term of $\Omega(\boldsymbol{\Theta}, \boldsymbol{b})$ controls the complexity of the induced model. In addition, the second term of $\Omega(\boldsymbol{\Theta}, \boldsymbol{b})$ is defined based on the ϵ-insensitive loss function:

$$\Omega_{1i} = \begin{cases} 0, & r_i < \epsilon \\ (r_i - \epsilon)^2, & r_i \ge \epsilon \end{cases} \tag{14}$$

For each example $(\boldsymbol{x}_i, \boldsymbol{d}_i)$ in $\mathcal{E}$, the corresponding input to the ϵ-insensitive loss function Ω_{1i} is set as: $r_i = \|\boldsymbol{e}_i\| = \sqrt{\boldsymbol{e}_i^\top \boldsymbol{e}_i}$ with $\boldsymbol{e}_i = \boldsymbol{d}_i - \varphi(\boldsymbol{x}_i)^\top\boldsymbol{\Theta} - \boldsymbol{b}$. In this way, the outputs of all linear predictors are considered simultaneously to yield a unique input to Ω_{1i} such that the dependencies among all the class labels can be exploited by the ϵ-insensitive term.

The third term of $\Omega(\boldsymbol{\Theta}, \boldsymbol{b})$ considers the partial multi-label loss for each example which is set as:

$$\Omega_{2i} = -\left(\frac{1}{|Y_i|}\cdot\mathbf{1}_{Y_i}^\top - \frac{1}{|\hat{Y}_i|}\cdot\mathbf{1}_{\hat{Y}_i}^\top\right)\left(\boldsymbol{\Theta}^\top\varphi(\boldsymbol{x}_i) + \boldsymbol{b}\right) \tag{15}$$

Here, for candidate label set Y_i and its complementary set $\hat{Y}_i$ in $\mathcal{Y}$, $\mathbf{1}_{Y_i}$ corresponds to a c-dimensional vector whose k-th element equals to 1 if $y_k \in Y_i$ and 0 otherwise. Similarly, $\mathbf{1}_{\hat{Y}_i}$ corresponds to a c-dimensional vector whose k-th element equals to 1 if $y_k \in \hat{Y}_i$ and 0 otherwise. In other words, the third term enforces the property that the average output from candidate labels should be larger than the average output from non-candidate ones (Cour, Sapp, and Taskar 2011; Zhang 2014).

To minimize $L(\boldsymbol{\Theta}, \boldsymbol{b})$, PML-LD employs the gradient-based iterative method named Iterative Re-Weighted Least Square (IRWLS) (Sánchez-Fernández et al. 2004). According to the representor's theorem (Smola 1999), under fairly general conditions, a learning problem can be expressed as a linear combination of the training examples in the feature space, i.e. $\boldsymbol{\theta}^j = \sum_i \boldsymbol{\eta}^j\varphi(\boldsymbol{x}_i)$. If we replace this expression into Eq. (7) and Eq. (13), it will generate the inner product$< \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j) >$, and then the kernel trick can be applied.

Virtual Label Bipartition PML-LD proceeds to predict the set of proper labels for $\boldsymbol{x}$ via virtual label Bipartition. According to (Li, Zhang, and Geng 2015), an extra virtual label y_0 is added into the original label set, i.e., the extended original label set $\mathcal{Y}' = \mathcal{Y} \cup \{y_0\} = \{y_0, y_1, ..., y_c\}$. In this paper, the origin value $l_{\boldsymbol{x}}^{y_0}$ is set to 0.5. Once the recovered label distribution and the predictive model have been learned on the extended original label set, the extended label distribution $\boldsymbol{d}^*$ corresponding to the test instance $\boldsymbol{x}^*$ can be predicted. Then, the predicted label set for $\boldsymbol{x}^*$ is determined as:

$$f(\boldsymbol{x}) = \{y_j \mid d_{\boldsymbol{x}}^{y_j} > d_{\boldsymbol{x}}^{y_0}, 1 \le j \le c\} \tag{16}$$

Experiments

Datasets

To thoroughly evaluate the performance of comparing approaches, a number of synthetic as well as real-world PML datasets have been employed for experimental studies. Table 1 summarizes characteristics of the experimental datasets used in this paper.

Specifically, a synthetic PML dataset is generated from one multi-label dataset by adding random labeling noise. For each multi-label example, some of its irrelevant labels are randomly chosen to form the candidate label set along with its relevant labels. As shown in Table 1, five benchmark multi-label datasets (Zhang and Zhou 2014) are used to generate synthetic PML datasets, including `image`,

Table 1: Characteristics of the PML experimental datasets. For each PML dataset, the average number of candidate labels (avg. #CLs) and the average number of ground-truth labels (avg. #GLs) are also recorded.

Dataset	#Examples	#Features	#Labels	avg. #CLs	avg. #GLs
emotions	593	72	6	3, 4, 5	1.86
image	2,000	294	5	2, 3, 4	1.23
scene	2,407	294	6	3, 4, 5	1.07
yeast	2,417	103	14	9, 10, 11, 12, 13	4.23
eurlex_sm	12,679	100	15	5, 6, 7, 8, 9, 10, 11, 12, 13, 14	1.53
music_emotion	6,833	98	11	5.29	2.42
music_style	6,839	98	10	6.04	1.44
mirflickr	10,433	100	7	3.35	1.77

Table 2: Experimental results of each comparing approach in terms of *hamming loss*, where the best performance (the smaller the better) is shown in bold face.

Dataset	avg.#CLS	PML-LD	PML-LC	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML
emotions	3	**0.180±0.014**	0.260±0.012	0.241±0.013	0.205±0.009	0.226±0.009	0.295±0.023	0.252±0.020
	4	**0.169±0.014**	0.258±0.015	0.250±0.017	0.206±0.012	0.232±0.012	0.306±0.021	0.257±0.012
	5	**0.218±0.023**	0.313±0.024	0.271±0.013	0.344±0.034	0.264±0.020	0.355±0.013	0.277±0.020
image	2	**0.151±0.010**	0.209±0.011	0.190±0.006	0.175±0.012	0.179±0.017	0.380±0.011	0.200±0.007
	3	**0.165±0.012**	0.219±0.008	0.200±0.013	0.195±0.013	0.192±0.009	0.416±0.010	0.206±0.010
	4	**0.186±0.009**	0.279±0.013	0.231±0.012	0.329±0.010	0.236±0.019	0.435±0.005	0.231±0.014
scene	3	**0.083±0.006**	0.155±0.008	0.139±0.006	0.117±0.002	0.114±0.005	0.351±0.005	0.114±0.006
	4	**0.098±0.005**	0.185±0.023	0.160±0.003	0.147±0.007	0.135±0.006	0.367±0.004	0.143±0.004
	5	**0.119±0.012**	0.237±0.010	0.193±0.011	0.383±0.014	0.198±0.025	0.380±0.006	0.176±0.015
yeast	9	**0.139±0.001**	0.229±0.007	0.215±0.005	0.207±0.010	0.237±0.013	0.438±0.006	0.265±0.004
	10	**0.139±0.002**	0.236±0.007	0.218±0.006	0.202±0.004	0.224±0.009	0.439±0.008	0.268±0.002
	11	**0.143±0.001**	0.237±0.008	0.224±0.005	0.210±0.008	0.222±0.009	0.448±0.004	0.270±0.004
	12	**0.143±0.001**	0.247±0.006	0.230±0.004	0.336±0.008	0.230±0.004	0.453±0.005	0.271±0.005
	13	**0.145±0.001**	0.270±0.008	0.268±0.004	0.697±0.004	0.232±0.006	0.465±0.006	0.282±0.005
eurlex_sm	5	**0.067±0.001**	0.112±0.001	0.082±0.002	0.067±0.001	0.075±0.002	0.083±0.002	0.087±0.002
	6	0.070±0.001	0.121±0.001	0.084±0.001	**0.067±0.000**	0.078±0.002	0.085±0.001	0.089±0.001
	7	0.071±0.001	0.130±0.001	0.083±0.001	**0.068±0.000**	0.080±0.001	0.084±0.001	0.089±0.001
	8	0.074±0.001	0.129±0.000	0.085±0.001	**0.068±0.001**	0.084±0.002	0.086±0.001	0.089±0.001
	9	0.076±0.001	0.126±0.002	0.087±0.002	**0.068±0.001**	0.088±0.002	0.087±0.002	0.092±0.001
	10	0.076±0.001	0.123±0.002	0.088±0.002	**0.071±0.002**	0.093±0.002	0.087±0.002	0.091±0.002
	11	0.080±0.001	0.124±0.001	0.088±0.001	**0.073±0.001**	0.098±0.002	0.091±0.002	0.091±0.001
	12	**0.082±0.001**	0.122±0.002	0.090±0.001	0.101±0.001	0.104±0.002	0.093±0.001	0.095±0.001
	13	**0.088±0.002**	0.122±0.001	0.097±0.001	0.855±0.002	0.126±0.003	0.098±0.002	0.099±0.003
	14	**0.091±0.002**	0.107±0.005	0.103±0.001	0.897±0.000	0.150±0.004	0.106±0.003	0.110±0.002
music_emotion	5.29	**0.123±0.002**	0.241±0.004	0.244±0.002	0.211±0.004	0.217±0.003	0.389±0.003	0.217±0.003
music_style	6.04	**0.109±0.002**	0.152±0.036	0.125±0.002	0.121±0.001	0.155±0.004	0.432±0.003	0.116±0.003
mirflickr	3.35	**0.062±0.045**	0.236±0.057	0.214±0.048	0.186±0.036	0.180±0.036	0.329±0.076	0.194±0.033

`emotions`, `scene`, `yeast`, and `eurlex_sm`. For each multi-label dataset, different settings are considered by varying the average number of candidate labels (avg. #CLs). Accordingly, a total of twenty-four synthetic PML datasets have been generated. Furthermore, three real-world PML datasets including `music_emotion`, `music_style` and `mirflickr` (Huiskes and Lew 2008) are also employed in this paper. For the real-world PML dataset, candidate labels are collected from web users which are further examined by human labellers to specify the ground-truth labels.

Methodology

The performance of PML-LD is compared against six state-of-the-art partial multi-label learning approaches, each configured with parameters suggested in respective literature:

- PML-LC (Xie and Huang 2018) which optimize labeling confidence and predictive model alternatively with label correlations [suggested configuration: $C_1 = 1$, C_2 with $\{1, 2, ..., 10\}$, C_3 with $\{1, 10, 100\}$].
- PML-FP (Xie and Huang 2018) which optimize labeling confidence and predictive model alternatively with feature prototypes [suggested configuration: $C_1 = 1$, C_2 with $\{1, 2, ..., 10\}$, C_3 with $\{1, 10, 100\}$].
- FPML (Yu et al. 2018) which adopts noisy labels estimation to learn from partial multi-label examples via low-rank approximation [suggested configuration: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 10$].
- PARTICLE-VLS (Fang and Zhang 2019) which adopts credible label elicitation technique to learn from partial multi-label examples and virtual label splitting for predictive model induction [suggested configuration: $k = 10$,

Table 3: Experimental results of each comparing approach in terms of *average precision*, where the best performance (the larger the better) is shown in bold face.

Data Set	avg.#CLS	PML-LD	PML-LC	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML
emotions	3	**0.804±0.021**	0.752±0.029	0.781±0.021	0.800±0.020	0.800±0.027	0.757±0.021	0.763±0.019
	4	0.789±0.026	0.753±0.027	0.758±0.039	**0.803±0.017**	0.792±0.022	0.751±0.022	0.754±0.028
	5	**0.741±0.028**	0.664±0.021	0.708±0.025	0.717±0.026	0.724±0.041	0.714±0.022	0.708±0.019
image	2	**0.809±0.020**	0.736±0.022	0.769±0.013	0.790±0.024	0.789±0.024	0.776±0.016	0.767±0.016
	3	**0.787±0.021**	0.698±0.016	0.751±0.018	0.779±0.017	0.781±0.014	0.745±0.013	0.745±0.017
	4	**0.762±0.017**	0.592±0.011	0.701±0.014	0.721±0.015	0.723±0.018	0.718±0.015	0.704±0.015
scene	3	**0.863±0.013**	0.718±0.008	0.762±0.015	0.830±0.009	0.826±0.013	0.801±0.015	0.814±0.013
	4	**0.839±0.010**	0.658±0.047	0.715±0.010	0.792±0.013	0.792±0.010	0.754±0.015	0.757±0.012
	5	**0.797±0.022**	0.546±0.031	0.644±0.024	0.703±0.012	0.712±0.019	0.699±0.024	0.686±0.030
yeast	9	**0.746±0.007**	0.713±0.013	0.738±0.011	0.744±0.007	0.722±0.007	0.558±0.008	0.734±0.005
	10	**0.744±0.007**	0.708±0.012	0.730±0.008	0.743±0.007	0.720±0.009	0.548±0.012	0.726±0.008
	11	**0.738±0.006**	0.699±0.014	0.723±0.009	0.738±0.006	0.712±0.008	0.527±0.008	0.712±0.007
	12	**0.728±0.005**	0.686±0.005	0.709±0.001	0.726±0.004	0.699±0.007	0.494±0.003	0.695±0.006
	13	**0.712±0.004**	0.654±0.009	0.651±0.004	0.704±0.003	0.688±0.001	0.475±0.005	0.650±0.005
eurlex_sm	5	**0.793±0.007**	0.486±0.006	0.707±0.009	0.789±0.005	0.779±0.004	0.713±0.008	0.676±0.006
	6	**0.778±0.005**	0.445±0.004	0.695±0.004	0.777±0.005	0.762±0.007	0.700±0.005	0.663±0.005
	7	0.769±0.003	0.417±0.009	0.690±0.007	**0.771±0.001**	0.759±0.006	0.701±0.006	0.658±0.010
	8	**0.754±0.011**	0.415±0.006	0.675±0.009	0.753±0.006	0.742±0.006	0.690±0.007	0.664±0.008
	9	0.734±0.006	0.429±0.014	0.661±0.004	**0.739±0.006**	0.729±0.009	0.681±0.005	0.643±0.004
	10	0.731±0.004	0.446±0.008	0.658±0.006	**0.736±0.005**	0.728±0.005	0.675±0.006	0.649±0.008
	11	0.709±0.005	0.444±0.008	0.653±0.007	**0.724±0.004**	0.710±0.005	0.649±0.009	0.644±0.005
	12	0.692±0.009	0.457±0.007	0.637±0.006	**0.704±0.002**	0.699±0.005	0.642±0.007	0.621±0.003
	13	0.662±0.010	0.475±0.008	0.607±0.004	**0.672±0.006**	0.665±0.005	0.604±0.008	0.597±0.014
	14	**0.619±0.006**	0.542±0.025	0.563±0.006	0.610±0.007	0.606±0.010	0.565±0.015	0.535±0.009
music_emotion	5.29	**0.630±0.010**	0.574±0.010	0.566±0.009	0.607±0.010	0.611±0.011	0.621±0.006	0.605±0.007
music_style	6.04	**0.737±0.003**	0.612±0.096	0.701±0.005	0.713±0.004	0.710±0.007	0.554±0.004	0.727±0.005
mirflickr	3.35	**0.835±0.090**	0.715±0.040	0.744±0.058	0.671±0.027	0.827±0.101	0.615±0.078	0.783±0.068

Table 4: Win/tie/loss counts of pairwise t-test (at 0.05 aignificance level) on PML-LD against each comparing approach.

	PML-LD **against**					
	PML-LC	PML-FP	PARTICLE-VLS	PARTICLE-MAP	PML-LRS	FPML
Ranking loss	25/2/0	22/5/0	21/6/0	19/8/0	23/4/0	23/4/0
Hamming loss	26/1/0	27/0/0	20/1/6	27/0/0	27/0/0	27/0/0
One-error	25/2/0	23/4/0	1/16/10	5/20/2	25/2/0	21/6/0
Coverage	23/4/0	20/7/0	7/19/1	17/10/0	21/6/0	20/7/0
Average precision	25/2/0	22/5/0	8/17/2	14/13/0	24/3/0	24/3/0
In Total	**124/11/0**	**114/21/0**	**57/59/19**	**82/51/2**	**120/15/0**	**115/20/0**

$\alpha = 0.95$, $thr = 0.9$].

- PARTICLE-MAP (Fang and Zhang 2019) which adopts credible label elicitation technique to learn from partial multi-label examples and maximum a posteriori (MAP) reasoning for predictive model induction [suggested configuration: $k = 10$, $\alpha = 0.95$, $thr = 0.9$].
- PML-LRS (Sun et al. 2019) which adopts low-rank and sparse decomposition scheme to learn from partial multi-label examples [suggested configuration: $\eta = 1$, $\gamma = 0.1$, $\beta = 1$].

For PML-LD, the parameter λ_1, λ_2, m, β_1 and β_2 are fix to 0.01, 0.01, 20, 1, 10 respectively. The kernel function in PML-LD is Gaussian kernel.

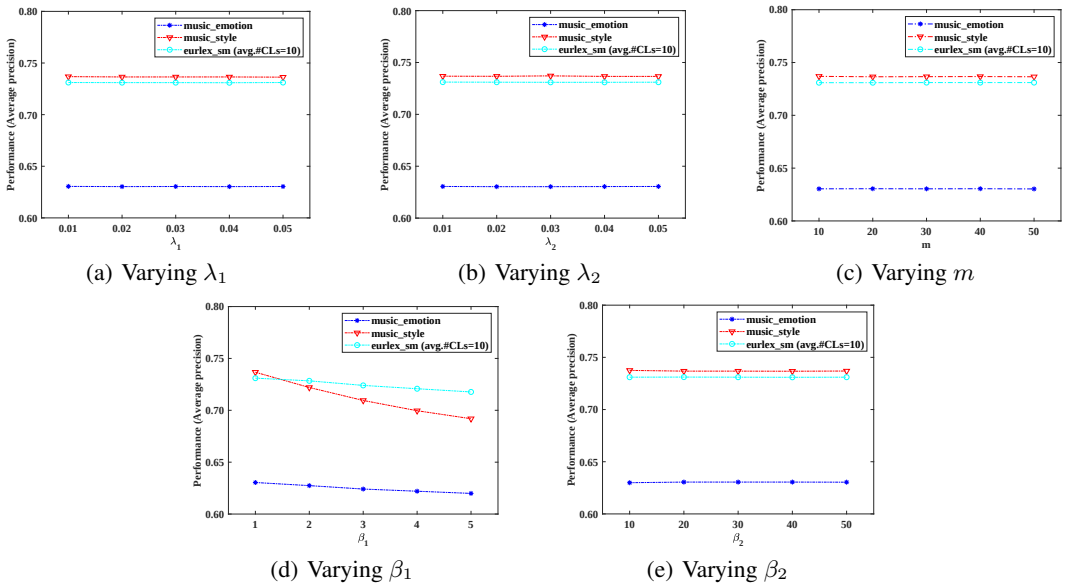
Five popular multi-label metrics *ranking loss*, *hamming loss*, *one-error*, *coverage*, and *average precision* are employed for performance evaluation, whose detailed definitions can be found in (Zhang and Zhou 2014; Gibaja and Ventura 2015). On each dataset, five-fold cross-validation is performed where the mean metric value as well as standard deviation are recorded for each comparing approach.

Experimental Results

Tables 2 and 3 report the detailed experimental results. Due to page limitation, we only show representative results on *hamming loss* and *average precision*. Those results on other evaluation measures are similar. For each dataset and evaluation metric, pairwise t-test based on five-fold cross-validation (at 0.05 significance level) is conducted to show whether the performance of PML-LD is significantly different to the comparing approach. Accordingly, Table 4 summarizes the resulting win/tie/loss counts over 27 datasets and 5 evaluation metrics.

Based on the experimental results of comparative studies, it is impressive to observe that:

- Across all the statistical tests, PML-LD achieves supe-



(a) Varying λ_1 (b) Varying λ_2 (c) Varying m

(d) Varying β_1 (e) Varying β_2

Figure 3: Parameter sensitivity analysis for PML-LD on `music_emotion`, `music_style` and `eurlex_sm`. (a) Classification accuracy of PML-LD changes as λ_1 increases from 0.01 to 0.05 with step-size 0.01 ($\lambda_2 = 0.01, m = 20, \beta_1 = 1, \beta_2 = 10$); (b) Classification accuracy of PML-LD changes as λ_2 increases from 0.01 to 0.05 with step-size 0.01 ($\lambda_1 = 0.01, m = 20, \beta_1 = 1, \beta_2 = 10$); (c) Classification accuracy of PML-LD changes as m increases from 10 to 50 with step-size 10 ($\lambda_1 = 0.01, \lambda_2 = 0.01, \beta_1 = 1, \beta_2 = 10$); (d) Classification accuracy of PML-LD changes as β_1 increases from 1 to 5 with step-size 1 ($\lambda_1 = 0.01, \lambda_2 = 0.01, m = 20, \beta_2 = 10$); (e) Classification accuracy of PML-LD changes as β_2 increases from 10 to 50 with step-size 10 ($\lambda_1 = 0.01, \lambda_2 = 0.01, m = 20, \beta_1 = 1$).

rior or at least comparable performance against PML-LC, PML-FP, PML-LRS and FPML. Especially, PML-LD achieves superior performance against PML-LC, PML-FP, PML-LRS and FPML in 91.9% cases (124 out of 135), 84.4% cases (114 out of 135), 88.9% cases (120 out of 135) and 85.2% cases (115 out of 135) respectively.

- PML-LD achieves comparable performance against PARTICLE-VLS and PARTICLE-MAP in 85.9% cases (116 out of 135) and 98.5% cases (133 out of 135) respectively. In addition, PML-LD achieves superior performance against PARTICLE-VLS and PARTICLE-MAP in 42.2% cases (57 out of 135) and 63.0% cases (82 out of 135) respectively.
- On the real-world PML datasets `music_emotion`, `music_style` and `mirflickr`, PML-LD achieves optimal performance in almost all cases. It is because that PML-LD can better recover the hidden label distributions in the real-world PML datasets.

Sensitivity Analysis

In this subsection, performance sensitivity of the proposed PML-LD approach w.r.t. its parameters λ_1, λ_2, m, β_1 and β_2 will be further analyzed.

Figure 3 illustrates how PML-LD performs under different parameter configurations. For clarity of illustration, three datasets `music_emotion`, `music_style` and `eurlex_sm` are chosen here for sensitivity analysis while similar observations also hold on other datasets.

As shown in Figure 3, it is obvious that the performance of PML-LD is relatively stable across a broad range of each parameter. This property is quite desirable as one can make use of PML-LD to achieve robust classification performance without the need of parameter fine-tuning. Therefore, the parameter configuration for PML-LD in Subsection 4.2 naturally follows from these observations.

Conclusion

In this paper, the problem of PML is studied where a novel approach PML-LD is proposed. Different from existing strategies, PML-LD considers the label distributions in the training datasets. Since the label distributions are not explicitly available in the training sets, PML-LD recovers the label distributions via leveraging the topological information of the feature space and the correlations among the labels, and then induces the predictive model based on multi-output regression analysis. Effectiveness of the proposed approach is validated via comprehensive experiments on both synthetic datasets and real-world PML datasets.

Acknowledgments

This research was supported by the National Key Research & Development Plan of China (No. 2017YFB1002801), the National Science Foundation of China (61622203), the Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and

the Collaborative Innovation Center of Wireless Communications Technology.

References

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

Chung, W.; Kim, J.; Lee, H.; and Kim, E. 2015. General dimensional multiple-output support vector regressions and their multiple kernel learning. *IEEE Transactions on Cybernetics* 45(11):2572–2584.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *Journal of Machine Learning Research* 12(May):1501–1536.

Elisseeff, A., and Weston, J. 2002. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14 (NIPS 2002)*, 681–687.

Fang, J.-P., and Zhang, M.-L. 2019. Partial multi-label learning via credible label elicitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3518–3525.

Fürnkranz, J.; Hüllermeier, E.; Mencía, E. L.; and Brinker, K. 2008. Multilabel classification via calibrated label ranking. *Machine Learning* 73(2):133–153.

Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1734–1748.

Gibaja, E., and Ventura, S. 2015. A tutorial on multilabel learning. *ACM Computing Surveys* 47(3):Article 52.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 39–43.

Li, Y.-K.; Zhang, M.-L.; and Geng, X. 2015. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *Proceedings of the 15th IEEE International Conference on Data Mining*, 251–260.

Liu, L., and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. In Bartlett, P.; Pereira, F. C. N.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 25*. Cambridge, MA: MIT Press. 557–565.

Ning, X.; An, T.; and Xin, G. 2018. Label enhancement for label distribution learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2926–2932.

Nocedal, J., and Wright, S. J. 2006. *Numerical optimization*. New York: Springer.

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

Sánchez-Fernández, M.; de Prado-Cumplido, M.; Arenas-García, J.; and Pérez-Cruz, F. 2004. SVM multiregression for nonliear channel estimation in multiple-input multiple-output systems. *IEEE Transactions on Signal Processing* 52(8):2298–2307.

Smola, A. J. 1999. *Learning with kernels*. Ph.D. Thesis, GMD, Birlinghoven, German.

Sun, L.; Feng, S.; Wang, T.; Lang, C.; and Jin, Y. 2019. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5016–5023.

Tsoumakas, G.; Dimou, A.; Spyromitros, E.; and Mezaris, V. 2009. Correlation-based pruning of stacked binary relevance models for multi-label learning. In *Proceedings of the 1st International Workshop on Learning from Multi-Label Data*, 101–116.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2011. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering* 23(7):1079–1089.

Xie, M.-K., and Huang, S.-J. 2018. Partial multi-label learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 4302–4309.

Xu, N.; Tao, A.; and Geng, X. 2018. Label enhancement for label distribution learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2926–2932.

Yu, F., and Zhang, M.-L. 2017. Maximum margin partial label learning. *Machine Learning* 106(4):573–593.

Yu, G.; Chen, X.; Domeniconi, C.; Wang, J.; Li, Z.; Zhang, Z.; and Wu, X. 2018. Feature-induced partial multi-label learning. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1398–1403.

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, 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.

Zhang, M.-L.; Yu, F.; and Tang, C.-Z. 2017. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2155–2167.

Zhang, M.-L. 2014. Disambiguation-free partial label learning. In *Proceedings of the 14th SIAM International Conference on Data Mining*, 37–45.

Zhou, Z.-H., and Zhang, M.-L. 2017. Multi-label learning. In Sammut, C., and Webb, G. I., eds., *Encyclopedia of Machine Learning and Data Mining, 2nd Edition*. Berlin: Springer.

Zhou, Y.; Xue, H.; and Geng, X. 2015. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM International Conference on Multimedia*, 1247–1250.

Zhou, Z.-H. 2018. A brief introduction to weakly supervised learning. *National Science Review* 5(1):44–53.

Zhu, X.; Lafferty, J.; and Rosenfeld, R. 2005. *Semi-supervised learning with graphs*. Carnegie Mellon University, language technologies institute, school of computer science.