JMLR: Workshop and Conference Proceedings 30:1–16, 2015 ACML 2015

Maximum Margin Partial Label Learning

Fei Yu YUF@SEU.EDU.CN

Min-Ling Zhang ZHANGML@SEU.EDU.CN

School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China

Abstract

Partial label learning deals with the problem that each training example is associated with a set of *candidate* labels, and only one among the set is the ground-truth label. The basic strategy to learn from partial label examples is disambiguation, i.e. by trying to recover the ground-truth labeling information from the candidate label set. As one of the major machine learning techniques, maximum margin criterion has been employed to solve the partial label learning problem. Therein, disambiguation is performed by optimizing the margin between the maximum modeling output from candidate labels and that from non-candidate labels. However, in this formulation the margin between the ground-truth label and other candidate labels is not differentiated. In this paper, a new maximum margin formulation for partial label learning is proposed which aims to directly maximize the margin between the ground-truth label and all other labels. Specifically, an alternating optimization procedure is utilized to coordinate *ground-truth label identification* and *margin maximization*. Extensive experiments show that the derived partial label learning approach achieves competitive performance against other state-of-the-art comparing approaches.

1. Introduction

Partial label learning deals with the problem that each training example is associated with a set of candidate labels, among which only one label is valid (Cour et al., 2011; Zhang, 2014). In recent years, many real-world learning tasks were solved under the framework of partial label learning such as web mining (Jie and Orabona, 2010), multimedia content analysis (Cour et al., 2011; Zeng et al., 2013), ecoinformatics (Liu and Dietterich, 2012), etc.

Formally speaking, let $\mathcal{X} = \mathbb{R}^d$ be the d-dimensional instance space and $\mathcal{Y} = \{1, 2, \ldots, q\}$ be the label space with q possible class labels. Furthermore, let $\mathcal{D} = \{(\mathbf{x}_i, S_i) \mid 1 \leq i \leq m\}$ denote the partial label training set where $\mathbf{x}_i \in \mathcal{X}$ is a d-dimensional feature vector $(x_{i1}, x_{i2}, \ldots, x_{id})^\top$ and $S_i \subseteq \mathcal{Y}$ is the associated candidate label set. The task of partial label learning is to induce a multi-class classifier $f : \mathcal{X} \to \mathcal{Y}$ based on $\mathcal{D}$. In partial label learning, the ground-truth label y_i of $\mathbf{x}_i$ is hidden in its candidate label set S_i and not accessible during training phase.[1]

The basic strategy to learn from partial label examples is disambiguation, which aims to identify the ground-truth label from candidate label set. Maximum margin criterion is one of the major

1. Partial label learning is also termed as *ambiguous label learning* (Hüllermeier and Beringer, 2006; Chen et al., 2014), *soft label learning* (Côme et al., 2008), or *superset label learning* (Liu and Dietterich, 2014) in some literatures. Furthermore, there are some studies which admit noisy candidate label set without containing the ground-truth label (Cid-Sueiro, 2012).

techniques in designing machine learning approaches. Earlier attempts towards maximum margin partial label learning perform disambiguation by optimizing the margin between the maximum modeling output from candidate labels and that from non-candidate labels (Nguyen and Caruana, 2008). In other words, given the parametric model $\mathbf{\Theta}$ and $\mathbf{x}_i$'s modeling output $F(\mathbf{x}_i, y; \mathbf{\Theta})$ on each label $y \in \mathcal{Y}$, existing formulation aims to maximize the following predictive difference over each instance: $\max_{y_j \in S_i} F(\mathbf{x}_i, y_j; \mathbf{\Theta}) - \max_{y_k \notin S_i} F(\mathbf{x}_i, y_k; \mathbf{\Theta})$. Nonetheless, the above margin does not consider the predictive difference between the ground-truth label (i.e. y_i) and other candidate labels (i.e. $S_i \setminus \{y_i\}$), which may lead to suboptimal performance for the resulting maximum margin partial label learning approach.

Note that the goal of partial label learning is to induce a multi-class classifier $f : \mathcal{X} \to \mathcal{Y}$, the canonical *multi-class margin*, i.e. $F(\mathbf{x}_i, y_i; \mathbf{\Theta}) - \max_{\tilde{y}_i \neq y_i} F(\mathbf{x}_i, \tilde{y}_i; \mathbf{\Theta})$, would be a natural choice for learning from partial label examples. Here, the predictive difference between the ground-truth label and all other labels can be fully taken into account. In light of this, a new learning approach named M3PL, i.e. *MaxiMum Margin Partial Label learning*, is proposed in this paper. The major difficulty in making use of multi-class margin for partial label training examples lies in that the ground-truth labeling information is not accessible to the learning algorithm. To tackle this issue, an alternating optimization procedure is employed by M3PL to iteratively identify the ground-truth label and maximize the multi-class margin. Comparative studies against other well-established partial label learning approaches clearly validate the effectiveness of the proposed formulation.

The rest of this paper is organized as follows. Section 2 briefly discusses related work. Section 3 presents details of the proposed M3PL approach. Section 4 reports the result of comparative experiments. Finally, Section 5 summarizes the paper and indicates several future research issues.

2. Related Work

Partial label learning can be regarded as one of the *weakly-supervised* learning frameworks, which lies between the two ends of the supervision spectrum, i.e. supervised learning with explicit supervision and unsupervised learning with blind supervision. Learning with weak supervision has found wide applications in solving various learning tasks as obtaining explicit and sufficient supervision information in real-world scenarios is generally hard (Pfahringer, 2012). In particular, partial label learning is related to several popular weakly-supervised learning frameworks such as *semi-supervised learning*, *multi-instance learning* and *multi-label learning*.

Semi-supervised learning (Chapelle et al., 2006; Zhu and Goldberg, 2009) learns from training examples which are either explicitly labeled with a single label or unlabeled without any labeling information, while for partial label learning the training examples are partially labeled with a set of candidate labels. Multi-instance learning (Dietterich et al., 1997; Amores, 2013) learns from training examples with labels assigned to a bag of instances, while for partial label learning the candidate labels are assigned to single instances. Multi-label learning (Tsoumakas et al., 2010; Zhang and Zhou, 2014) learns from training examples each associated with multiple valid labels, while for partial label learning only one of the candidate labels associated with the instance is valid.

In recent years, several approaches have been proposed to solving partial label learning problem by utilizing major machine learning techniques, such as maximum likelihood (ML) estimation (Jin and Ghahramani, 2003; Liu and Dietterich, 2012), convex optimization (Cour et al., 2011), k-nearest neighbors (Hüllermeier and Beringer, 2006; Zhang and Yu, 2015), sparse coding (Chen et al., 2014), error-correcting output codes (ECOC) (Zhang, 2014), etc. Specifically, maximum margin

criterion has also been applied to design partial label learning approaches (Nguyen and Caruana, 2008). Given the (linear) classification model $\boldsymbol{\Theta} = \{(\mathbf{w}_p, b_p) \mid 1 \leq p \leq q\}$, existing partial label maximum margin formulation aims to solve the following optimization problem (OP):

OP 1: Existing Maximum Margin Formulation

$$\begin{aligned} \min_{\boldsymbol{\Theta}, \boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{p=1}^{q} ||\mathbf{w}_p||^2 + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.}: \quad & \max_{y_j \in S_i} (\mathbf{w}_{y_j}^\top \cdot \mathbf{x}_i + b_{y_j}) - \max_{y_k \notin S_i} (\mathbf{w}_{y_k}^\top \cdot \mathbf{x}_i + b_{y_k}) \geq 1 - \xi_i \\ & \xi_i \geq 0 \qquad\qquad \forall i \in \{1, 2, \ldots, m\} \end{aligned}$$

Here, $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_m\}$ correspond to the set of slack variables and C is the regularization parameter. As shown in **OP 1**, existing formulation focuses on differentiating the maximum output from candidate labels against that from non-candidate labels. One potential issue lies in that the predictive difference between the ground-truth label and other candidate labels are not considered in this formulation, which may lead to suboptimal performance for the resulting partial label learning approach.

In the next section, a new maximum margin formulation towards partial label learning is proposed, which aims to maximize the margin between the ground-truth label and all other labels in the label space.

3. The M3PL Approach

3.1. Proposed Formulation

Following the same notations in Section 1, the training set $\mathcal{D}$ consists of m partial label examples $(\boldsymbol{x}_i, S_i)$ $(1 \leq i \leq m)$ with $\mathbf{x}_i \in \mathcal{X}$ and $S_i \subseteq \mathcal{Y}$. In addition, let $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ be the (unknown) ground-truth label assignments for the training examples. Under the partial label learning assumption, the ground-truth label of each instance $\mathbf{x}_i$ should reside in its candidate label set S_i. Therefore, the feasible solution space for ground-truth label assignments corresponds to $\mathcal{S} = S_1 \times S_2 \times \cdots \times S_m$.

As usual, M3PL assumes a maximum margin learning system with q linear classifiers $\boldsymbol{\Theta} = \{(\mathbf{w}_p, b_p) \mid 1 \leq p \leq q\}$, one for each class label. Once the ground-truth label assignments $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ are fixed, M3PL aims to maximize the canonical multi-class margin over each instance $\mathbf{x}_i$, i.e.: $(\mathbf{w}_{y_i}^\top \cdot \mathbf{x}_i + b_{y_i}) - \max_{\tilde{y}_i \neq y_i} (\mathbf{w}_{\tilde{y}_i}^\top \cdot \mathbf{x}_i + b_{\tilde{y}_i})$. By introducing slack variables $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_m\}$ to accommodate margin relaxations, the maximum margin formulation considered by M3PL corresponds to the following optimization problem:

OP 2: Proposed Maximum Margin Formulation

$$\begin{aligned} \min_{\mathbf{y},\boldsymbol{\Theta},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{p=1}^{q}||\mathbf{w}_p||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.}: \quad & (\mathbf{w}_{y_i}^\top \cdot \mathbf{x}_i + b_{y_i}) - \max_{\tilde{y}_i \neq y_i}(\mathbf{w}_{\tilde{y}_i}^\top \cdot \mathbf{x}_i + b_{\tilde{y}_i}) \geq 1 - \xi_i \\ & \xi_i \geq 0 \qquad\qquad \forall i \in \{1,2,\ldots,m\} \\ & \mathbf{y} \in \mathcal{S} \\ & \sum_{i=1}^{m}\mathbb{I}(y_i = p) = n_p \qquad \forall p \in \{1,2,\ldots,q\} \end{aligned}$$

As shown in **OP 2**, the first two constraints impose the maximum margin criterion over each training example, and the third constraint imposes that the ground-truth label assignment $\mathbf{y}$ should be confined within the feasible solution space $\mathcal{S}$. The fourth constraint, i.e. $\sum_{i=1}^{m}\mathbb{I}(y_i = p) = n_p$, serves as an additional restriction on $\mathbf{y}$ in terms of its compatibility with the prior class distribution.[2] Here, n_p represents the prior number of examples for the p-th class label in $\mathcal{Y}$.

Intuitively, by sharing equal labeling confidence $\frac{1}{|S_i|}$ among each candidate label in S_i, the prior number can be roughly estimated as:

$$\hat{n}_p = \sum_{i=1}^{m}\mathbb{I}(p \in S_i) \cdot \frac{1}{|S_i|} \tag{1}$$

Obviously, $\sum_{p=1}^{q}\hat{n}_p = m$ holds. Furthermore, let $\lfloor \hat{n}_p \rfloor$ denote the integer part of $\hat{n}_p$ and $r = m - \sum_{p=1}^{q}\lfloor \hat{n}_p \rfloor$ denote the residual number after the rounding operation. Then, the integer value n_p used in the fourth constraint is determined as:

$$n_p = \begin{cases} \lfloor \hat{n}_p \rfloor + 1 & if\ p\ is\ among\ the\ r\ class\ labels\ with\ least\ \hat{n}_p\ values \\ \lfloor \hat{n}_p \rfloor & otherwise \end{cases} \tag{2}$$

Accordingly, $\sum_{p=1}^{q} n_p = m$ still holds.

Note that **OP 2** involves the optimization of mixed-type variables (i.e. integer variables $\mathbf{y}$ and real-valued variables $\boldsymbol{\Theta}$), which is difficult to be optimized simultaneously. In the following subsection, an alternating optimization procedure is employed to update $\mathbf{y}$ and $\boldsymbol{\Theta}$ in an iterative manner.

3.2. Alternating Optimization

3.2.1. FIX $\mathbf{y}$, UPDATE $\boldsymbol{\Theta}$

By fixing the ground-truth label assignments $\mathbf{y} = (y_1, y_2, \ldots, y_m)$, **OP 2** turns to be the following optimization problem:

2. $\mathbb{I}(a)$ is an indicator function which returns 1 if the argument a is true, and 0 otherwise.

OP 3: Classification Model Optimization

$$\begin{aligned} \min_{\boldsymbol{\Theta},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{p=1}^{q}||\mathbf{w}_p||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.}: \quad & (\mathbf{w}_{y_i}^\top \cdot \mathbf{x}_i + b_{y_i}) - \max_{\tilde{y}_i \neq y_i}(\mathbf{w}_{\tilde{y}_i}^\top \cdot \mathbf{x}_i + b_{\tilde{y}_i}) \geq 1 - \xi_i \\ & \xi_i \geq 0 \qquad\qquad \forall i \in \{1, 2, \ldots, m\} \end{aligned}$$

As shown in **OP 3**, the resulting optimization problem is identical to the well-studied single-label multi-class maximum margin formulation (Crammer and Singer, 2001; Hsu and Lin, 2002). Therefore, any off-the-shelf implementations on multi-class SVM can be used here to fulfill the optimization task (Fan et al., 2008).

3.2.2. FIX $\boldsymbol{\Theta}$, UPDATE $\mathbf{y}$

By fixing the classification model $\boldsymbol{\Theta} = \{(\mathbf{w}_p, b_p) \mid 1 \leq p \leq q\}$, **OP 2** turns to be the following optimization problem:

OP 4: Ground-truth Label Assignment Optimization (Version 1)

$$\begin{aligned} \min_{\mathbf{y},\boldsymbol{\xi}} \quad & \sum_{i=1}^{m}\xi_i \\ \text{s.t.}: \quad & \xi_i \geq 1 - \eta_i^{y_i} \\ & \xi_i \geq 0 \qquad \forall i \in \{1, 2, \ldots, m\} \\ & \mathbf{y} \in \mathcal{S} \\ & \sum_{i=1}^{m}\mathbb{I}(y_i = p) = n_p \qquad \forall p \in \{1, 2, \ldots, q\} \end{aligned}$$

Here, $\eta_i^{y_i}$ represents the multi-class margin on $\mathbf{x}_i$ by taking y_i as its ground-truth label, i.e.:

$$\eta_i^{y_i} = (\mathbf{w}_{y_i}^\top \cdot \mathbf{x}_i + b_{y_i}) - \max_{\tilde{y}_i \neq y_i}(\mathbf{w}_{\tilde{y}_i}^\top \cdot \mathbf{x}_i + b_{\tilde{y}_i}) \tag{3}$$

By replacing $\xi_i = \max(0, 1 - \eta_i^{y_i})$ according to the first two constraints, **OP 4** can be transformed into the following equivalent form:

OP 5: Ground-truth Label Assignment Optimization (Version 2)

$$\begin{aligned} \min_{\mathbf{y}} \quad & \sum_{i=1}^{m}\max(0, 1 - \eta_i^{y_i}) \\ \text{s.t.}: \quad & \mathbf{y} \in \mathcal{S} \\ & \sum_{i=1}^{m}\mathbb{I}(y_i = p) = n_p \qquad \forall p \in \{1, 2, \ldots, q\} \end{aligned}$$

Let $\mathbf{Z} = [z_{pi}]_{q\times m}$ denote the labeling matrix for training examples with binary values, where $z_{pi} = 1$ indicates that the p-th class label in $\mathcal{Y}$ is the ground-truth label for $\mathbf{x}_i$. Accordingly, set the coefficient matrix $\mathbf{C} = [c_{pi}]_{q\times m}$ as follows:

$$\forall 1 \leq p \leq q,\ 1 \leq i \leq m: \quad c_{pi} = \begin{cases} \max(0, 1 - \eta_i^p) & if \ \ p \in S_i \\ M & otherwise \end{cases} \tag{4}$$

Here, M is a user-specified constant with large value so as to refrain from assigning ground-truth label outside the candidate label set.[3] Based on the above definitions, **OP 5** can be rewritten in the following form:

OP 6: Ground-truth Label Assignment Optimization (Version 3)

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \sum_{p=1}^{q} \sum_{i=1}^{m} c_{pi} \cdot z_{pi} \\ \text{s.t.}: \quad & \sum_{p=1}^{q} z_{pi} = 1 \qquad \forall i \in \{1, 2, \ldots, m\} \\ & \sum_{i=1}^{m} z_{pi} = n_p \qquad \forall p \in \{1, 2, \ldots, q\} \\ & z_{pi} \in \{0, 1\} \end{aligned}$$

Here, the first constraint $\sum_{p=1}^{q} z_{pi} = 1$ ensures that the ground-truth label for each training example is unique. In addition, the second constraint $\sum_{i=1}^{m} z_{pi} = n_p$ enforces the constraint w.r.t. prior class distribution.

Note that **OP 6** corresponds to a binary integer programming (BIP) problem, which is generally NP-hard to solve. Fortunately, **OP 6** falls into a special case of BIP where the constraint matrix is totally unimodular and the right-hand sides of the constraints are integers (Papadimitriou and Steiglitz, 1998). In this case, the original BIP problem can be equivalently solved in its linear programming (LP) relaxation form by replacing the integer constraint $z_{pi} \in \{0, 1\}$ with the weaker interval constraint $z_{pi} \in [0, 1]$:

OP 7: Ground-truth Label Assignment Optimization (Version 4)

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \sum_{p=1}^{q} \sum_{i=1}^{m} c_{pi} \cdot z_{pi} \\ \text{s.t.}: \quad & \sum_{p=1}^{q} z_{pi} = 1 \qquad \forall i \in \{1, 2, \ldots, m\} \\ & \sum_{i=1}^{m} z_{pi} = n_p \qquad \forall p \in \{1, 2, \ldots, q\} \\ & 0 \leq z_{pi} \leq 1 \end{aligned}$$

Thereafter, solution to the relaxation problem **OP 7** can be efficiently found by employing standard LP solvers such as simplex algorithm or interior point algorithm (Boyd and Vandenberghe, 2004).

3. In this paper, M is set to be 10^5.

Algorithm 1 The M3PL Approach

Inputs:

$\mathcal{D}$: the partial label training set, $\{(\mathbf{x}_i, S_i) \mid 1 \leq i \leq m\}$ ($\mathbf{x}_i \in \mathcal{X}, S_i \subseteq \mathcal{Y}$)

$C_{\max}$: the maximum value for regularization parameter

$\mathbf{x}^*$: the unseen instance.

Outputs:

y^*: the predicted class label for $\mathbf{x}^*$

Process:

Initialize the regularization parameter: $C = 10^{-5} \cdot C_{\max}$;
Initialize the coefficient matrix $\mathbf{C}$ according to Eq.(5);
Solve the LP problem **OP 7**, and then initialize the ground-truth label assignment $\mathbf{y}$ with $y_i = \arg\max_{1 \leq p \leq q} z_{pi}$ $(1 \leq i \leq m)$;
while $C < C_{\max}$ **do**
  $C = \min\{(1 + \Delta)C,\ C_{\max}\}$;
  Initialize the objective function value in **OP 2**: $Obj = +\infty$;
  repeat
    $Obj_{\text{old}} = Obj$;
    Solve the multi-class maximum margin problem **OP 3**, and then update the classification model $\boldsymbol{\Theta}$;
    Set the coefficient matrix $\mathbf{C}$ according to Eq.(4);
    Solve the LP problem **OP 7**, and then update the ground-truth label assignment $\mathbf{y}$ with $y_i = \arg\max_{1 \leq p \leq q} z_{pi}$ $(1 \leq i \leq m)$;
    Calculate the new objective function value in **OP 2**: $Obj = \frac{1}{2}\sum_{p=1}^{q} ||\mathbf{w}_p||^2 + C\sum_{i=1}^{m} \max(0, 1 - \eta_i^{y_i})$;
  until $Obj_{\text{old}} - Obj < \delta$
end while
return $y^* = \arg\max_{p \in \mathcal{Y}}\ \mathbf{w}_p^\top \cdot \mathbf{x}^* + b_p$;

3.3. Iterative Implementation

To initialize the alternating optimization procedure, M3PL sets the initial coefficient matrix $\mathbf{C}$ by resorting to candidate label sets:

$$\forall 1 \leq p \leq q,\ 1 \leq i \leq m: \quad c_{pi} = \begin{cases} \frac{1}{|S_i|} & if \ \ p \in S_i \\ M & otherwise \end{cases} \tag{5}$$

By solving **OP 7** based on initial coefficients, the ground-truth label assignment $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ would be $y_i = \arg\max_{1 \leq p \leq q} z_{pi}$. Then, the classification model $\boldsymbol{\Theta}$ is updated by solving **OP 3** and the alternating optimization procedure iterates. The iteration procedure terminates once the objective function value in **OP 2** decreases less than δ after one round of alternating update.

Instead of specifying some fixed value for the regularization parameter C, M3PL chooses to gradually increase the value of C within an outer annealing loop. Similar strategy has been used in solving other weakly-supervised learning problems (Joachims, 1999; Chapelle et al., 2008) to reduce the risk of the learning algorithm being getting stuck with local minimum solution.

Table 1: Characteristics of the experimental data sets.

Controlled UCI Data Sets				Configurations
Data set	**#Examples**	**#Features**	**#Class Labels**	
glass	214	10	5	(I) $r = 1, p \in \{0.1, 0.2, \ldots 0.7\}$
vehicle	846	18	4	(II) $r = 2, p \in \{0.1, 0.2, \ldots 0.7\}$
segment	2310	18	7	(III) $r = 3, p \in \{0.1, 0.2, \ldots 0.7\}$
satimage	6435	36	7	(IV) $p = 1, r = 1, \epsilon \in \{0.1, 0.2, \ldots 0.7\}$

Real-World Data Sets					
Data Set	**# Examples**	**# Features**	**# Class Labels**	**avg. #CLs**	**Domain**
Lost	1122	108	16	2.23	automatic face naming (Cour et al., 2011)
BirdSong	4998	38	13	2.18	bird song classification (Briggs et al., 2012)
MSRCv2	1758	48	23	3.16	object classification (Liu and Dietterich, 2012)
Yahoo! News	22991	163	219	1.91	automatic face naming (Guillaumin et al., 2010)
Soccer Player	17472	279	171	2.09	automatic face naming (Zeng et al., 2013)

The pseudo-code of M3PL is summarized in **Algorithm 1**.[4] Given the partial label training examples, M3PL firstly initializes the regularization parameter C and the ground-truth label assignment (Steps 1-3). After that, the classification model and ground-truth label assignment is alternatively optimized until convergence (Steps 7-13). An outer loop is used to gradually increase the value of C by a factor of $1 + \Delta$ (Step 5). Finally, the unseen instance is classified based on the learned classification model (Step 15). [5]

4. Experiment

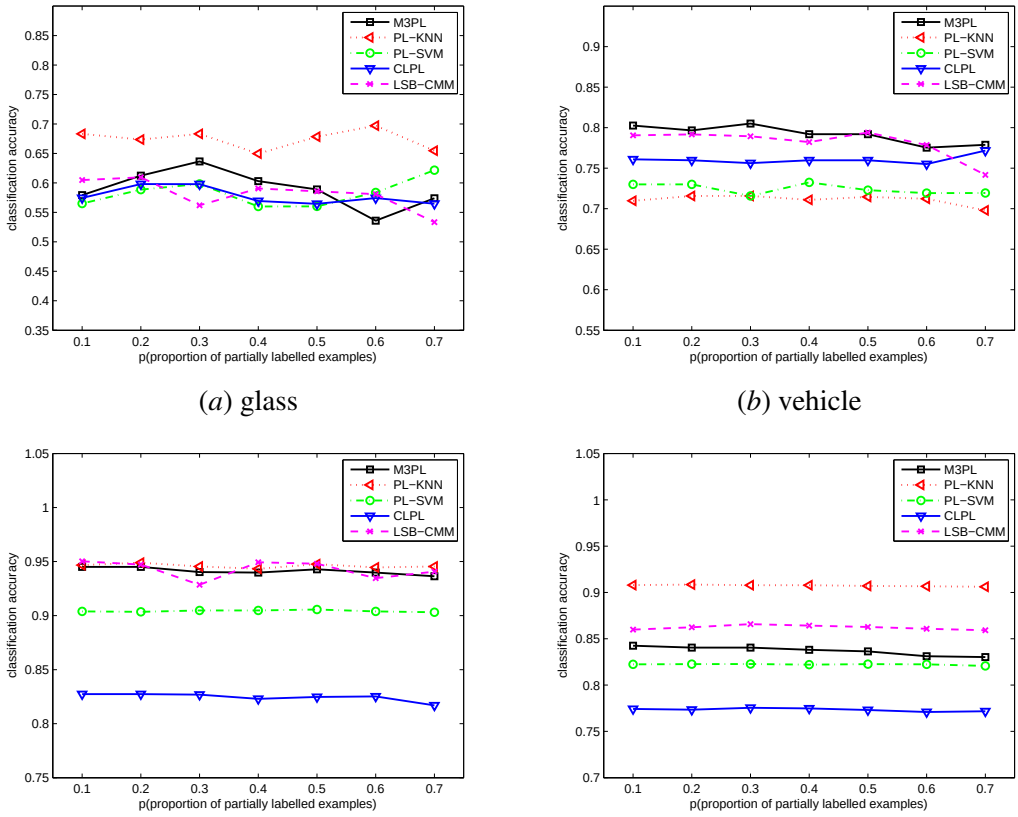
4.1. Experimental Settings

In this section, two series of experiments are conducted to evaluate the performance of the proposed approach, with one series on controlled UCI data sets (Bache and Lichman, 2013) and the other one on real-world partial label data sets. Table 1 summarizes characteristics of the employed data sets.

Following the controlling protocol over multi-class UCI data set (Cour et al., 2011; Chen et al., 2014; Liu and Dietterich, 2012; Zhang, 2014), an artificial partial label data set can be generated under different configurations of three controlling parameters p, r and ϵ. Here, p controls the proportion of examples with partial labeling (i.e. $|S_i| > 1$), r controls the number of candidate labels other than the ground-truth label (i.e. $|S_i| = r + 1$), and ϵ controls the co-occurring probability between one extra candidate label and the ground-truth label. As shown in Table 1, a total of 28 (4x7) configurations are considered for each of the four UCI data sets.

The real-world partial label data sets are collected from several task domains, such as *automatic face naming* including `Lost` (Cour et al., 2011), `Yahoo! News` (Guillaumin et al., 2010), `Soccer Player` (Zeng et al., 2013), *bird song classification* including `BirdSong` (Briggs et al., 2012), and *object classification* including `MSRCv2` (Liu and Dietterich, 2012). For the automatic face naming task, faces cropped from an image or video frame are represented as instances while

4. Code package available at http://cse.seu.edu.cn/PersonalPage/zhangml/files/M3PL.zip.
5. In this paper, **OP 3** (Step 9) and **OP 7** (Step 11) are solved by adopting the LibLinear toolbox (Fan et al., 2008) and CVX toolbox (Grant and Boyd, 2014) respectively. Furthermore, Δ is set to be 0.5 following (Chapelle et al., 2008) and δ is set to be 10^{-4}.



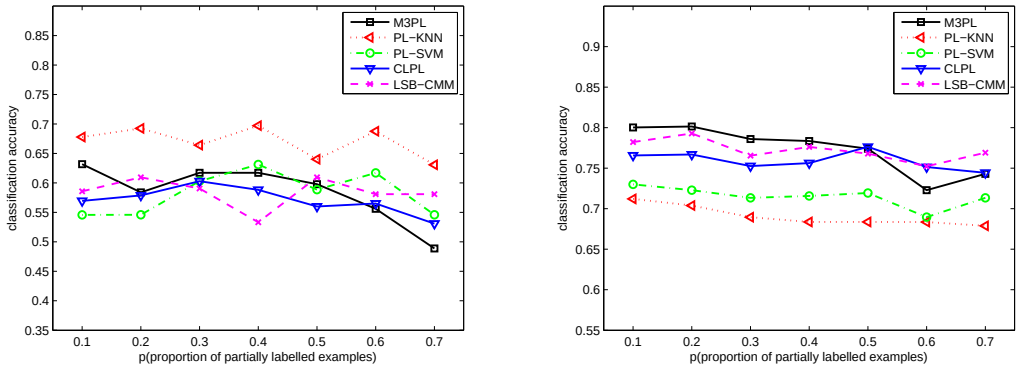
(*a*) glass (*b*) vehicle

(*c*) segment (*d*) satimage

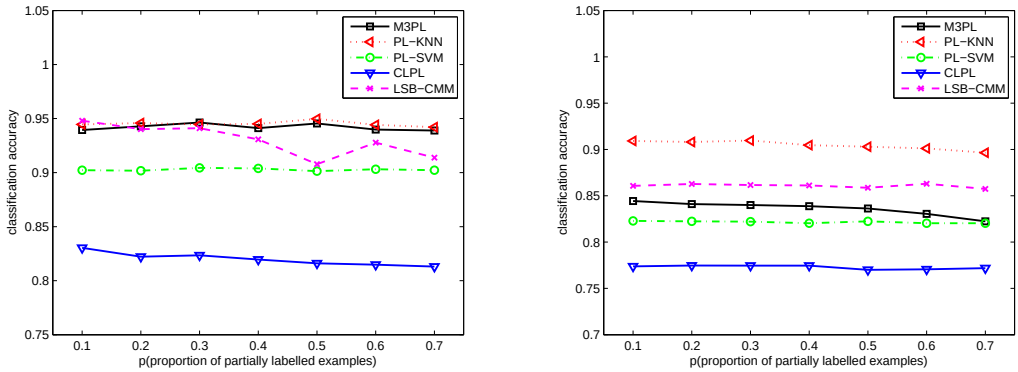
Figure 1: Classification accuracy of each comparing algorithm changes as p (proportion of partially labeled example) increases from 0.1 to 0.7 ($r = 1$).

names extracted from the associated captions or subtitles are regarded as candidate labels. For the bird song classification task, singing syllables of the birds are represented as instances while bird species jointly singing during a 10-seconds period are regarded as candidate labels. For the object classification task, image segmentations are represented as instances while objects appearing within the same image are regarded as candidate labels.

Four well-established partial label learning approaches are employed for comparative studies, each implemented with parameter setup suggested in respective literatures: 1) An existing maximum margin partial label learning approach PL-SVM (Nguyen and Caruana, 2008) [suggested setup: regularization parameter pool with $\{10^{-3}, \ldots, 10^{3}\}$]; 2) The k-nearest neighbor partial label learning approach PL-KNN (Hüllermeier and Beringer, 2006) [suggested setup: k=10]; 3) The convex optimization partial label learning approach CLPL (Cour et al., 2011) [suggested setup: SVM with squared hinge loss]; 4) The maximum likelihood partial label learning algorithm LSB-CMM (Liu and Dietterich, 2012) [suggested setup: q mixture components]. Accordingly, for M3PL the parameter $C_{\max}$ is chosen among $\{10^{-2}, \ldots, 10^{2}\}$ via cross-validation.



(*a*) glass (*b*) vehicle



(*c*) segment (*d*) satimage

Figure 2: Classification accuracy of each comparing algorithm changes as p (proportion of partially labeled example) increases from 0.1 to 0.7 ($r = 2$).

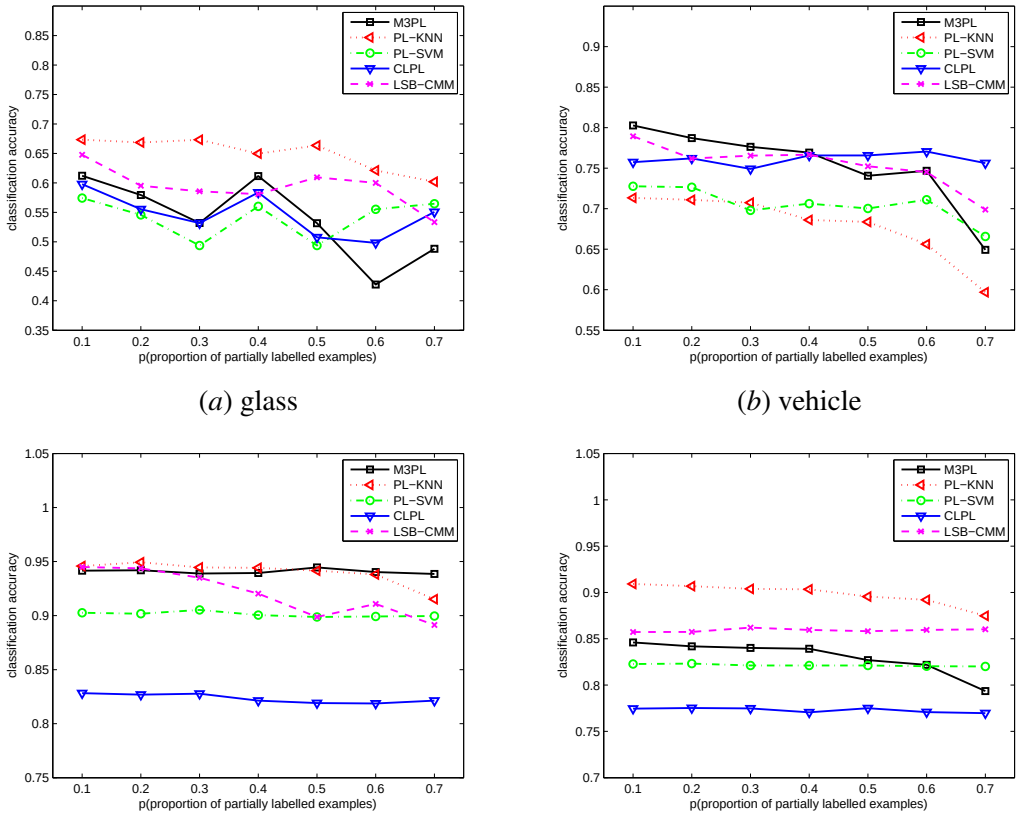
In this paper, ten-fold cross-validation is performed on each data set where the mean predictive accuracies as well as standard deviations are recorded for all comparing approaches.

4.2. Experimental Result

4.2.1. CONTROLLED UCI DATA SETS

Figures 1 to 3 illustrate the classification accuracy of each comparing algorithm as p increases from 0.1 to 0.7 with step-size 0.1 ($r = 1, 2, 3$). For any partially labeled example, its candidate label set consists of the ground-truth label along with r additional labels randomly chosen from $\mathcal{Y}$. Figure 4 illustrates the classification accuracy of each comparing algorithm as ϵ increases from 0.1 to 0.7 with step-size 0.1 ($p = 1, r = 1$). For any label $y \in \mathcal{Y}$, one extra label $y' \in \mathcal{Y}$ is designated as the coupling label which co-occurs with y in the candidate label set with probability ϵ. Otherwise, any other class label would be randomly chosen to co-occur with y.

As shown in Figures 1 to 4, M3PL achieves competitive performance against the comparing algorithms in most cases. Based on pairwise t-test at 0.05 significance level, Table 2 summarizes



(*a*) glass (*b*) vehicle

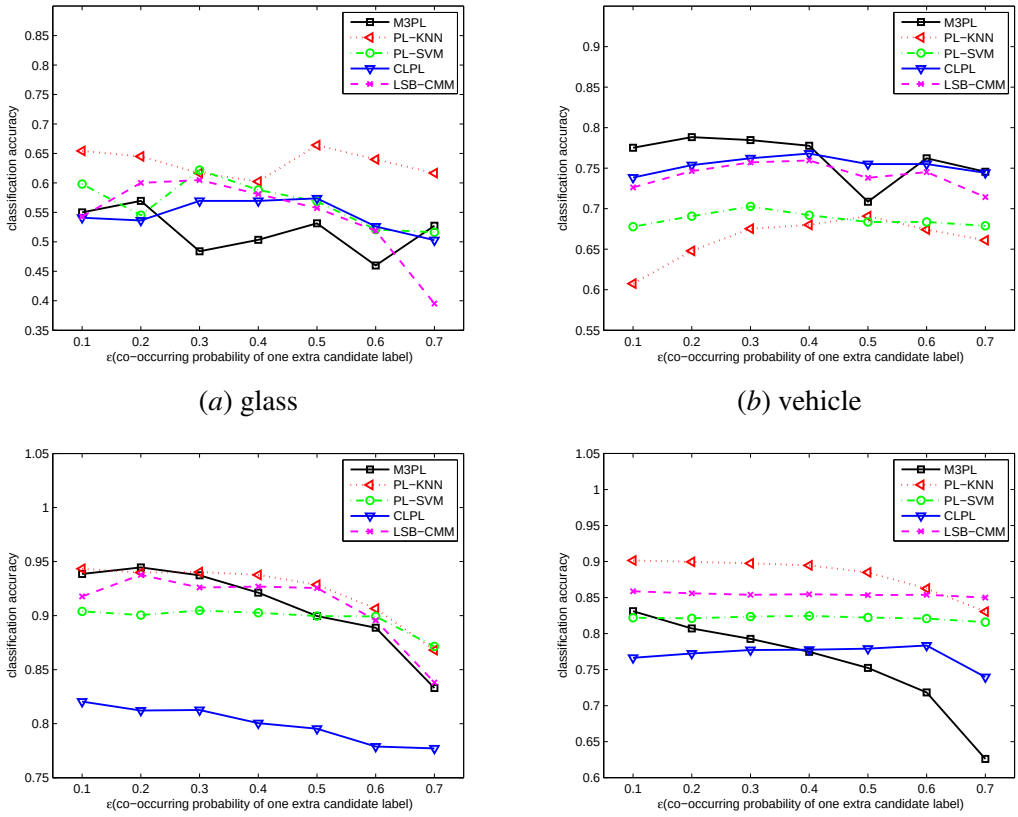
(*c*) segment (*d*) satimage

Figure 3: Classification accuracy of each comparing algorithm changes as p (proportion of partially labeled example) increases from 0.1 to 0.7 ($r = 3$).

the win/tie/loss counts between M3PL and the comparing algorithms. Out of the 112 statistical comparisons (28 configurations x 4 data sets), the following observations can be made:

- Compared to the existing maximum margin counterpart PL-SVM (Nguyen and Caruana, 2008), M3PL achieves superior performance in 51.8% cases and only loses in 8.0% cases. These results clearly indicate the advantage of the proposed formulation against existing partial label maximum margin formulation;

- Compared to PL-KNN (Hüllermeier and Beringer, 2006), CLPL (Cour et al., 2011) and LSB-CMM (Liu and Dietterich, 2012), M3PL achieves superior or at least comparable performance in 60.7%, 96.4% and 75.0% cases respectively. These results validate the ability of M3PL in achieving state-of-the-art generalization performance for partial label learning problem.



(*a*) glass (*b*) vehicle

(*c*) segment (*d*) satimage

Figure 4: Classification accuracy of each comparing algorithm changes as ϵ (co-occurring probability of the coupling label) increases from 0.1 to 0.7 ($p = 1$, $r = 1$).

Table 2: Win/tie/loss counts (pairwise t-test at 0.05 significance level) on the classification performance of M3PL against each comparing algorithm.

	M3PL **against**			
	PL-SVM	PL-KNN	CLPL	LSB-CMM
[Figure 1]	18/10/0	7/13/8	14/14/0	0/21/7
[Figure 2]	15/13/0	6/12/10	14/14/0	2/20/6
[Figure 3]	15/12/1	7/10/11	15/13/0	3/19/6
[Figure 4]	10/10/8	6/7/15	9/15/4	1/18/9
In Total	**58/45/9**	**26/42/44**	**52/56/4**	**6/78/28**

Table 3: Classification accuracy (mean±std) of each comparing algorithm on the real-world partial label data sets. In addition, •/◦ indicates whether M3PL is statistically superior/inferior to the comparing algorithm on each data set (pairwise t-test at 0.05 significate level).

	MSRCv2	Lost	BirdSong	Yahoo! News	Soccer Player
M3PL	0.521±0.030	0.767±0.043	0.709±0.010	0.655±0.009	0.446±0.013
PL-SVM	0.482±0.043•	0.729±0.040•	0.663±0.032•	0.636±0.010•	0.443±0.014•
CLPL	0.413±0.039•	0.742±0.038	0.632±0.017•	0.462±0.009•	0.368±0.010•
PL-KNN	0.448±0.037•	0.424±0.041•	0.614±0.024•	0.457±0.010•	0.497±0.014◦
LSB-CMM	0.456±0.031•	0.707±0.055•	0.717±0.024	0.648±0.007	0.525±0.015◦

Table 4: Transductive accuracy (mean±std) of each comparing algorithm on the real-world partial label data sets. In addition, •/◦ indicates whether M3PL is statistically superior/inferior to the comparing algorithm on each data set (pairwise t-test at 0.05 significate level).

	MSRCv2	Lost	BirdSong	Yahoo! News	Soccer Player
M3PL	0.732±0.025	0.860±0.006	0.855±0.030	0.870±0.002	0.761±0.010
M3PL†	0.735±0.025	0.876±0.007	0.861±0.048	0.881±0.002	0.766±0.009
PL-SVM	0.653±0.024•	0.887±0.012◦	0.825±0.012•	0.871±0.002	0.688±0.014•
CLPL	0.656±0.010•	0.894±0.005◦	0.822±0.004•	0.834±0.002•	0.680±0.010•
PL-KNN	0.616±0.006•	0.615±0.036•	0.772±0.021•	0.692±0.010•	0.492±0.015•
LSB-CMM	0.524±0.007•	0.721±0.010•	0.716±0.014•	0.872±0.001	0.704±0.002•

4.2.2. REAL-WORLD DATA SETS

Table 3 reports the performance of each comparing algorithm on the real-world partial label data sets. Based on the results of ten-fold cross-validation, pairwise t-tests at 0.05 significance level between M3PL and the comparing algorithms are recorded as well.

As shown in Table 3, it is impressive that M3PL significantly outperforms its maximum margin counterpart PL-SVM on all real-world data sets. Furthermore, M3PL achieves superior performance against CLPL, PL-KNN and LSB-CMM on the `MSRCv2` data set, and achieves superior or at least comparable performance against them on the `Lost`, `BirdSong` and `Yahoo! News` data sets. On the `Soccer Player` data set, M3PL significantly outperforms CLPL and is inferior to PL-KNN and LSB-CMM.

In addition to inductive performance on unseen examples, transductive accuracies of each comparing algorithm on training examples are also reported in Table 4. Here, for each training example $(\mathbf{x}_i, S_i)$, the prediction on its ground-truth label is made by consulting the candidate label set, i.e.: $y_i = \arg\max_{y \in S_i} F(\mathbf{x}_i, y; \boldsymbol{\Theta})$. Generally, transductive performance reflects the disambiguation ability of the partial label learning approach in recovering ground-truth labeling information from candidate label set. Similar to Table 3, pairwise t-tests at 0.05 significance level between M3PL and the comparing algorithms are also recorded in Table 4. Furthermore, as the training procedure of M3PL terminates, the identified ground-truth label assignment $\mathbf{y}$ can be also used as the disambiguation predictions on the training examples. The resulting transductive performance is reported in Table 4 as well (denoted as M3PL†) for reference purpose.

As shown in Table 4, M3PL significantly outperforms all the comparing algorithms on the MSRCv2, BirdSong, and Soccer Player data sets. On the Lost data set, M3PL achieves superior performance against PL-KNN and LSB-CMM and is inferior to PL-SVM and CLPL. On the Yahoo! News data set, M3PL achieves superior or at least comparable performance against the other comparing algorithms. As expected, M3PL and M3PL† show similar transductive performance over each real-world data set.

5. Conclusion

In this paper, the partial label learning problem is tackled by adopting a new formulation of the maximum margin criterion. Specifically, the canonical multi-class margin is directly optimized by the proposed approach with an alternating optimization procedure. Experimental studies on controlled UCI data sets and real-world partial label data sets validate the effectiveness of the derived M3LP approach.

In the future, it is worth studying whether better performance could be gained by incorporating kernel trick into the proposed M3PL approach. Furthermore, it is also interesting to investigate other ways to solve the proposed maximum margin formulation **OP 2** other than utilizing alternating optimization.

References

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

Kevin Bache and Moshe Lichman. UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine, 2013. URL [http://archive.ics.uci.edu/ml].

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Forrest Briggs, Xiaoli Z. Fern, and Raviv Raich. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 534–542, Beijing, China, 2012.

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

Olivier Chapelle, Vikas Sindhwani, and Sathiya S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9:203–233, 2008.

Yi-Chen Chen, Vishal M. Patel, Rama Chellappa, and P. Jonathon Phillips. Ambiguously labeled learning using dictionaries. *IEEE Transactions on Information Forensics and Security*, 9(12): 2976–2088, 2014.

Jesús Cid-Sueiro. Proper losses for learning from partial labels. In Fernando Pereira, Chris J. C. Burges, Leon Bottou, and Kilian Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1574–1582. MIT Press, Cambridge, MA, 2012.

Etienne Côme, Latifa Oukhellou, Thierry Denœux, and Patrice Aknin. Mixture model estimation with soft labels. In Didier Dubois, Maria Asuncion Lubiano, Henri Prade, María Angeles Gil, Przemyslaw Grzegorzewski, and Olgierd Hryniewicz, editors, *Advances in Soft Computing 48*, pages 165–174. Springer, Berlin, 2008.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multiple instance metric learning from automatically labeled bags of faces. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Lecture Notes in Computer Science 6311*, pages 634–647. Springer, Berlin, 2010.

Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

Eyke Hüllermeier and Jürgen Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

Luo Jie and Francesco Orabona. Learning from candidate labeling sets. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1504–1512. MIT Press, Cambridge, MA, 2010.

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 921–928. MIT Press, Cambridge, MA, 2003.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, Bled, Slovenia, 1999.

Liping Liu and Thomas G. Dietterich. A conditional multinomial mixture model for superset label learning. In Fernando Pereira, Chris J. C. Burges, Leon Bottou, and Kilian Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 557–565. MIT Press, Cambridge, MA, 2012.

Liping Liu and Thomas G. Dietterich. Learnability of the superset label learning problem. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1629–1637, Beijing, China, 2014.

Nam Nguyen and Rich Caruana. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 551–559, Las Vegas, NV, 2008.

Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, Mineola, NY, 1998.

Bernhard Pfahringer. Leaning with weak supervision: Charting the territory. In *Keynote at the 1st International Workshop on Learning with Weak Supervision*, in conjunction with ACML 2012, Singapore, 2012.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–686. Springer, Berlin, 2010.

Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 708–715, Portland, OR, 2013.

Min-Ling Zhang. Disambiguation-free partial label learning. In *Proceedings of the 14th SIAM International Conference on Data Mining*, pages 37–45, Philadelphia, PA, 2014.

Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 4048–4054, Buenos Aires, Argentina, 2015.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge & Data Engineering*, 26(8):1819–1937, 2014.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.