Leveraging Latent Label Distributions for Partial Label Learning

Lei Feng and **Bo An**
School of Computer Science and Engineering, Nanyang Technological University, Singapore
feng0093@e.ntu.edu.sg, boan@ntu.edu.sg

Abstract

In partial label learning, each training example is assigned a set of candidate labels, only one of which is the ground-truth label. Existing partial label learning frameworks either assume each candidate label of equal confidence or consider the ground-truth label as a latent variable hidden in the indiscriminate candidate label set, while the different labeling confidence levels of the candidate labels are regrettably ignored. In this paper, we formalize the different labeling confidence levels as the latent label distributions, and propose a novel unified framework to estimate the latent label distributions while training the model simultaneously. Specifically, we present a biconvex formulation with constrained local consistency and adopt an alternating method to solve this optimization problem. The process of alternating optimization exactly facilitates the mutual adaption of the model training and the constrained label propagation. Extensive experimental results on controlled UCI datasets as well as real-world datasets clearly show the effectiveness of the proposed approach.

1 Introduction

Partial label (PL) learning is a specific type of weakly supervised learning [Cour *et al.*, 2011], in which each instance is associated with a set of candidate labels. However, only one of the candidate labels is the ground-truth label, which is concealed in the training process. This learning problem is also termed as *ambiguous label learning* [Hüllermeier and Beringer, 2006; Zeng *et al.*, 2013; Chen *et al.*, 2014; Chen *et al.*, 2017] or *superset label learning* [Liu and Dietterich, 2012; Liu and Dietterich, 2014; Hüllermeier and Cheng, 2015; Gong *et al.*, 2017]. Since precisely labeled data are too expensive to be collected in reality, partial label learning has various application domains, such as ecoinformatics [Liu and Dietterich, 2012], image annotation [Cour *et al.*, 2009; Zeng *et al.*, 2013] and web mining [Luo and Orabona, 2010], etc.

Formally speaking, suppose we have m training examples $\mathcal{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m]^\top \in \mathbb{R}^{m\times n}$ with n dimensions, and their candidate label sets are denoted by $\{S_1, S_2, \cdots, S_m\}$, respectively. The ground-truth labels of these m examples are $\{y_1, y_2, \cdots, y_m\}$ with $y_i \in S_i$ ($i \in [m]$), while they are not directly accessible in the model training. Given the label space denoted by $\mathcal{Y} = \{1, 2, \cdots, l\}$, the task of partial label learning is to learn a function: $f : \mathcal{X} \to \mathcal{Y}$ from the imprecisely labeled training set $\mathcal{D} = \{(\mathbf{x}_i, S_i)|i \in [m]\}$ to accurately predict the label of the test example.

To learn with such PL examples, the key is how to properly deal with the candidate labels. To this end, there are mainly two learning frameworks, including the average-based framework and the identification-based framework. For the average-based framework, each candidate label is treated equally in the model training [Cour *et al.*, 2011; Zhang *et al.*, 2016]. For the identification-based framework, the ground-truth label is considered as a latent variable hidden in the indiscriminate candidate label set [Jin and Ghahramani, 2003; Liu and Dietterich, 2012; Chen *et al.*, 2014; Nguyen and Caruana, 2008; Yu and Zhang, 2016]. They all make predictions by aggregating the modeling outputs of the candidate labels without discrimination, while the confidence of each candidate label being the ground-truth label is regrettably ignored. As a consequence, these approaches may be suboptimal, since each candidate label normally makes different contributions to the model training.

To address this problem, we formalize the different labeling confidence levels of the candidate labels as the latent label distributions, and propose the LALO (partial label learning with LAtent Label distributiOns) approach. LALO first introduces a novel unified framework that estimates the latent label distributions while training the model simultaneously, and then presents a biconvex formulation with constrained local consistency, finally adopts an alternating method to solve this optimization problem. On the one hand, the inductive model is discriminatively trained by minimizing the least squares loss of fitting the latent label distributions. On the other hand, the latent label distributions are regularized by the modeling outputs via a constrained label propagation procedure specifically for the PL properties. Through the mutual promotion of the model training and the label propagation, the ground-truth label can be identified by optimally estimating the label distributions. The effectiveness of LALO is validated by experiments on 4 controlled UCI datasets and 5 real-world datasets.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 introduces the LALO approach. Section 4 presents the technical details of the alternating optimization method. Section 5 reports the experimental results of comparative studies. In the end, Section 6 concludes this paper and discusses future research issues.

2 Related Work

Due to the difficulty in dealing with ambiguous labeling information of PL examples, there are only two common partial label learning frameworks, including the average-based framework and the identification-based framework.

The average-based framework normally treats each candidate label equally in the model training, and averages the modeling outputs of all the candidate labels for predictions. Following this framework, some instance-based approaches [Hüllermeier and Beringer, 2006; Zhang and Yu, 2015] predict a test instance by averaging the candidate labeling information of its neighbors. In addition, some parametric approaches assume a parametric model $F(\mathbf{x}_i, y; \theta)$ [Cour *et al.*, 2011; Zhang *et al.*, 2016] that discriminates the average modeling output of the candidate labels from that of the non-candidate labels, i.e., $\max(\sum_{i=1}^{m}(\frac{1}{|S_i|}\sum_{y\in S_i}F(\mathbf{x}_i, y; \theta) - \frac{1}{|\hat{S}_i|}\sum_{y\in\hat{S}_i}F(\mathbf{x}_i, y; \theta)))$ where S_i and $\hat{S}_i$ denote the candidate and non-candidate label set respectively. Although this framework is intuitive, the obvious drawback is that the ground-truth label may be overwhelmed by other candidate (false positive) labels without discrimination.

Instead of maximizing the average modeling output of all the candidate labels, the identification-based framework aims at directly maximizing the modeling output of exactly one candidate label, which is distinguished as the ground-truth label. Existing approaches following this framework consider the ground-truth label as a latent variable determined by $y_i = \arg\max_{y\in S_i} F(\mathbf{x}_i, y; \theta)$. Generally, the objective function is optimized according to the maximum likelihood criterion: $\max(\sum_{i=1}^{m}\log(\sum_{y\in S_i}\frac{1}{|S_i|}F(\mathbf{x}_i, y, \theta)))$ [Jin and Ghahramani, 2003; Liu and Dietterich, 2012] or the maximum margin criterion: $\max(\sum_{i=1}^{m}(\max_{y\in S_i}\frac{1}{|S_i|}F(\mathbf{x}_i, y; \theta) - \max_{y\in\hat{S}_i}\frac{1}{|\hat{S}_i|}F(\mathbf{x}_i, y; \theta)))$ [Nguyen and Caruana, 2008; Yu and Zhang, 2016]. Because of indiscriminately targeting the ground-truth label within the candidate label set, the identification-based framework is sensitive to the false positive labels that co-occur with the ground-truth label.

In a nutshell, the above learning frameworks train the model with the modeling outputs of the candidate labels indiscriminate (i.e., the same weight $\frac{1}{|S_i|}$), while the different labeling confidence levels of the candidate labels are regrettably ignored. To address this problem, a novel unified partial label learning framework will be introduced in the next section. Following this framework, a biconvex formulation is presented to estimate the latent label distributions while training the model simultaneously.

3 The LALO Approach

For each training example, we receive a feature vector $\mathbf{x}_i \in \mathbb{R}^n$ and its corresponding label vector $\mathbf{y}_i \in \{0,1\}^l$ with l labels. Suppose m denotes the number of training examples, $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $\mathbf{Y} \in \{0,1\}^{m\times l}$ are the instance matrix and label matrix, respectively. In this setting, $y_{ij} = 1$ means the i-th training sample is assigned the j-th label.

Existing partial label learning frameworks indiscriminately train the model with noise-corrupted label matrix $\mathbf{Y} \in \{0,1\}^{m\times l}$, in which the labeling confidence of each candidate label is not discriminated. However, each candidate normally makes different contributions to the model training. To capture the labeling confidence (relative importance) of each candidate label, we propose to train the model with the latent label distributions. Specifically, for a training example $\mathbf{x}_i \in \mathbb{R}^n$, its latent label distribution is denoted by $\mathbf{p}_i \in [0,1]^l$. By arranging the label distributions of m training examples, we form the label distribution matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m]^\top \in [0,1]^{m\times l}$. By substituting $\mathbf{Y} \in \{0,1\}^{m\times l}$ with $\mathbf{P} \in [0,1]^{m\times l}$, we thus propose a novel unified framework that estimates the latent label distributions while training the model simultaneously:

$$\min \sum_{i=1}^{m} L(\mathbf{x}_i, \mathbf{p}_i, \mathbf{f}) + \lambda\Omega(\mathbf{f}) + \mu\Psi(\mathbf{P}) \tag{1}$$

where L denotes the prescribed loss function, Ω controls the complexity of the model $\mathbf{f}$, Ψ aims to guarantee an accurate estimation of the label distribution matrix $\mathbf{P}$, and λ, μ are parameters trading off these three terms.

Unlike the average-based framework and the identification-based framework, our proposed framework naturally treats the modeling outputs of the candidate labels in a discriminative manner due to the label distribution matrix $\mathbf{P}$, which can indicate the different contributions of the candidate labels. To optimally estimate $\mathbf{P}$, we assume it should have the following property: *local consistency*, i.e., nearby (similar) instances are supposed to have similar label distributions. Specifically, if the i-th instance $\mathbf{x}_i$ is similar to the j-th instance $\mathbf{x}_j$, their corresponding label distributions $\mathbf{p}_i$ and $\mathbf{p}_j$ should also be similar. In order to characterize the similarity between instances, we construct the similarity matrix $\mathbf{S} = [s_{ij}]_{m\times m}$ by the symmetry-favored k-NN graph [Liu and Chang, 2009]. Specifically, $s_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma^2)$ if $j \in \mathcal{N}_i$, otherwise $s_{ij} = 0$. The set $\mathcal{N}_i$ saves the indices of the k-nearest neighbors of $\mathbf{x}_i$, and the parameter σ is defined by $\sigma = \sum_{i=1}^{m}\|\mathbf{x}_i - \mathbf{x}_{i_k}\|_2/m$ where $\mathbf{x}_{i_k}$ denotes the k-th nearest neighbor of $\mathbf{x}_i$. To ensure that $\mathbf{S}$ is symmetric, we finally set $\mathbf{S} = \mathbf{S} + \mathbf{S}^\top$. In this way, we define $\Psi(\mathbf{P})$ as follows:

$$\begin{aligned} \Psi(\mathbf{P}) &= \sum_{i,j}^{m} s_{ij}\left\|\frac{\mathbf{p}_i}{\sqrt{d_{ii}}} - \frac{\mathbf{p}_j}{\sqrt{d_{jj}}}\right\|_2^2 \\ \text{s.t.} \quad & \sum\nolimits_j p_{ij} = 1, \quad \forall\, i \in [m] \\ & \mathbf{0}_{m\times l} \leq \mathbf{P} \leq \mathbf{Y} \end{aligned} \tag{2}$$

where $d_{ii} = \sum_j s_{ij}$ is the degree of the vertex $\mathbf{x}_i$ in the graph corresponding to the similarity matrix, and $\mathbf{0}_{m\times l} \in \{0\}^{m\times l}$. The first constraint formalizes the labeling confidence levels of all the labels as label distributions. The second constraint guarantees that the ground-truth label is strictly in the candidate label set, and the labeling confidence of each non-candidate label must be 0. Following the above settings, we propose to train the model by minimizing the least squares loss of fitting the label distributions:

$$L(\mathbf{x}_i, \mathbf{p}_i, \mathbf{f}) = \left\|\mathbf{x}_i\mathbf{W} + \mathbf{b}^\top - \mathbf{p}_i\right\|_2^2 \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{n\times l}$ and $\mathbf{b} \in \mathbb{R}^l$ are the model parameters. For the regularization term to control the model complexity, we adopt the widely-used squared Frobenius norm of $\mathbf{W}$:

$$\Omega(\mathbf{f}) = \|\mathbf{W}\|_F^2 \tag{4}$$

Finally, to further facilitate a kernel extension for the general nonlinear case, we present the formulation as a constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b},\mathbf{P}} \sum_i \|\mathbf{e}_i\|_2^2 + \lambda\|\mathbf{W}\|_F^2 + \mu\sum_{i,j}^m s_{ij}\left\|\frac{\mathbf{p}_i}{\sqrt{d_{ii}}} - \frac{\mathbf{p}_j}{\sqrt{d_{jj}}}\right\|_2^2 \\ \text{s.t.} \quad \mathbf{p}_i = \mathbf{z}_i\mathbf{W} + \mathbf{b}^\top + \mathbf{e}_i, \quad \forall\, i \in [m] \\ \sum\nolimits_j p_{ij} = 1, \quad \forall\, i\in[m] \\ \mathbf{0}_{m\times l} \le \mathbf{P} \le \mathbf{Y} \end{aligned} \tag{5}$$

where $\mathbf{z}_i = \phi(\mathbf{x}_i)$ and $\phi(\cdot): \mathbb{R}^n \to \mathbb{R}^h$ is a feature mapping that maps the feature space to some higher (maybe infinite) dimensional Hilbert space with h dimensions.

4 Alternating Optimization

Obviously, the optimization problem (5) is a biconvex problem [Gorski *et al.*, 2007], and we solve this problem in an alternating way. Specifically, we first optimize the objective function with respective to $\mathbf{W}$ and $\mathbf{b}$ when $\mathbf{P}$ is fixed, and then optimize the objective function with respective to $\mathbf{P}$ when $\mathbf{W}$ and $\mathbf{b}$ are both fixed. This procedure is repeated until convergence or the maximum number of iterations is reached.

Updating W and b

When $\mathbf{P} \in \mathbb{R}^{m\times l}$ is fixed, the optimization problem (5) with respective to $\mathbf{W}$ and $\mathbf{b}$ can be stated as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{b}} \operatorname{tr}(\mathbf{\Xi}^\top\mathbf{\Xi}) + \lambda\operatorname{tr}(\mathbf{W}^\top\mathbf{W}) \\ \text{s.t.} \quad \mathbf{P} = \mathbf{Z}\mathbf{W} + \mathbf{1}_m\mathbf{b}^\top + \mathbf{\Xi} \end{aligned} \tag{6}$$

where $\mathbf{\Xi} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_m]^\top \in \mathbb{R}^{m\times l}$, $\operatorname{tr}(\cdot)$ is the trace norm operator with the property $\operatorname{tr}(\mathbf{W}^\top\mathbf{W}) = \|\mathbf{W}\|_F^2$, and $\mathbf{1}_m = [1, 1, \cdots, 1]^\top \in \mathbb{R}^m$. Then, the Lagrangian of this problem can be expressed as:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \mathbf{b}, \mathbf{\Xi}, \mathbf{A}) = \operatorname{tr}(\mathbf{\Xi}^\top\mathbf{\Xi}) + \lambda\operatorname{tr}(\mathbf{W}^\top\mathbf{W}) - \\ \operatorname{tr}(\mathbf{A}^\top(\mathbf{Z}\mathbf{W} + \mathbf{1}_m\mathbf{b}^\top + \mathbf{\Xi} - \mathbf{P})) \end{aligned} \tag{7}$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m]^\top \in \mathbb{R}^{m\times l}$ is the matrix that stores the Lagrange multipliers. In this way, the following equations will be induced according to the KKT conditions:

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathbf{\Xi}} = 0 \Rightarrow \mathbf{A} = 2\mathbf{\Xi}, \frac{\partial\mathcal{L}}{\partial\mathbf{A}} = 0 \Rightarrow \mathbf{Z}\mathbf{W} + \mathbf{1}_m\mathbf{b}^\top + \mathbf{\Xi} = \mathbf{P} \\ \frac{\partial\mathcal{L}}{\partial\mathbf{W}} = 0 \Rightarrow \mathbf{W} = \frac{1}{2\lambda}\mathbf{Z}^\top\mathbf{A}, \frac{\partial\mathcal{L}}{\partial\mathbf{b}} = 0 \Rightarrow \mathbf{A}^\top\mathbf{1}_m = \mathbf{0}_l \end{aligned} \tag{8}$$

Above linear equations can be solved by following steps:

$$\begin{aligned} \mathbf{Z}\mathbf{W} + \mathbf{1}_m\mathbf{b}^\top + \mathbf{\Xi} = \mathbf{P} \\ \frac{1}{2\lambda}\mathbf{Z}\mathbf{Z}^\top\mathbf{A} + \mathbf{1}_m\mathbf{b}^\top + \frac{1}{2}\mathbf{A} = \mathbf{P} \end{aligned} \tag{9}$$

Here, we define the positive definite matrix $\mathbf{H}$ by $\mathbf{H} = \frac{1}{2\lambda}\mathbf{K} + \frac{1}{2}\mathbf{I}_{m\times m}$ and $\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{m\times m}$ is given by its elements $k_{ij} = \phi(\mathbf{x}_i)\phi(\mathbf{x}_j)^\top = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathcal{K}(\cdot,\cdot)$ is the kernel function. For LALO, Gaussian kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/(2\sigma'^2))$ is employed with σ' set to the average distance of all pairs of training examples. The matrix $\mathbf{I}_{m\times m}$ is an identity matrix with m rows and m columns. Then we can obtain:

$$\begin{aligned} \mathbf{H}\mathbf{A} + \mathbf{1}_m\mathbf{b}^\top &= \mathbf{P} \\ \mathbf{A} + \mathbf{H}^{-1}\mathbf{1}_m\mathbf{b}^\top &= \mathbf{H}^{-1}\mathbf{P} \\ \mathbf{1}_m^\top\mathbf{H}^{-1}\mathbf{1}_m\mathbf{b}^\top &= \mathbf{1}_m^\top\mathbf{H}^{-1}\mathbf{P} \\ \mathbf{b}^\top &= \frac{\mathbf{1}_m^\top\mathbf{H}^{-1}\mathbf{P}}{\mathbf{1}_m^\top\mathbf{H}^{-1}\mathbf{1}_m} \end{aligned} \tag{10}$$

For computational convenience, we define $\mathbf{s} = \mathbf{1}_m^\top\mathbf{H}^{-1} \in \mathbb{R}^{1\times m}$, and the results are reported as follows:

$$\begin{aligned} \mathbf{b}^\top &= \frac{\mathbf{s}\mathbf{P}}{\mathbf{s}\mathbf{1}_m} \\ \mathbf{A} &= \mathbf{H}^{-1}(\mathbf{P} - \mathbf{1}_m\mathbf{b}^\top) \end{aligned} \tag{11}$$

Updating P

When $\mathbf{W}$ and $\mathbf{b}$ are fixed, the modeling output matrix $\mathbf{Q} \in \mathbb{R}^{m\times l}$ is denoted by $\mathbf{Q} = \mathbf{Z}\mathbf{W} + \mathbf{1}_m\mathbf{b}^\top = \frac{1}{2\lambda}\mathbf{K}\mathbf{A} + \mathbf{1}_m\mathbf{b}^\top$, then $\mathbf{\Xi} = \mathbf{P} - \mathbf{Q}$. By eliminating $\mathbf{\Xi}$, we can obtain:

$$\begin{aligned} \min_{\mathbf{P}} \|\mathbf{P} - \mathbf{Q}\|_F^2 + \mu\sum_{i,j}^m s_{ij}\left\|\frac{\mathbf{p}_i}{\sqrt{d_{ii}}} - \frac{\mathbf{p}_j}{\sqrt{d_{jj}}}\right\|_2^2 \\ \text{s.t.} \quad \sum\nolimits_j p_{ij} = 1, \quad \forall\, i \in [m] \\ \mathbf{0}_{m\times l} \le \mathbf{P} \le \mathbf{Y} \end{aligned} \tag{12}$$

Here, this optimization problem is actually a constrained label propagation problem [Zhou *et al.*, 2004], where μ specifies the relative amount of labeling information from the neighbor points and the modeling outputs. The first constraint guarantees that a label distribution is consistently assigned to each instance in the process of label propagation. The second constraint guarantees that labels are only propagated among candidate labels. While in semi-supervised settings [Zhu and Goldberg, 2009], labels are normally propagated from labeled examples to unlabeled examples. In addition, traditional label propagation problems normally treat the observed label matrix $\mathbf{Y}$ as the initial label matrix. In contrast, since the observed label matrix $\mathbf{Y}$ is a noise-corrupted version in partial label learning, we take the modeling output matrix $\mathbf{Q}$ as the initial label matrix for each optimization iteration, thereby adjusting the confidence level of each candidate label iteratively. The optimization problem (12) can be reformulated as a standard Quadratic Programming (QP) problem, which can be solved by any off-the-shelf QP tools. The detailed information is given in Appendix A.

Algorithm 1 The LALO Algorithm

Inputs:
$\mathcal{D}$: the PL training set $\mathcal{D} = \{(\mathbf{X}, \mathbf{Y})\}$
k: the number of nearest neighbors used for the similarity matrix
λ, μ: the parameters trading off each term in the loss function
Output:
y: the predicted label for the test example $\mathbf{x}$
Process:
construct the similarity matrix by the symmetry-favored k-NN graph;
calculate the kernel matrix $\mathbf{K} = [\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)]_{m \times m}$;
initialize $\mathbf{P}$ according to (13);
repeat
  update $\mathbf{b}$ and $\mathbf{A}$ according to (11);
  update $\mathbf{Q} = \frac{1}{2\lambda}\mathbf{K}\mathbf{A} + \mathbf{1}_m \mathbf{b}^\top$;
  calculate $\widetilde{\mathbf{p}}$ by solving (16) with a general QP procedure;
  update $\mathbf{P}$ by reshaping $\widetilde{\mathbf{p}} \in \mathbb{R}^{ml}$ into $\mathbf{P} \in \mathbb{R}^{m \times l}$;
until convergence or the maximum number of iterations.
return the predicted label y according to (14).

At the beginning of the alternating optimization, we initialize the label distribution matrix $\mathbf{P} = [p_{ij}]_{m \times l}$ as follows:

$$p_{ij} = \begin{cases} \frac{1}{\sum_j y_{ij}}, & \text{if } y_{ij} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

After $\widetilde{\mathbf{p}}$ is figured out, we can easily obtain the label distribution matrix $\mathbf{P}$ by reshaping $\widetilde{\mathbf{p}} \in \mathbb{R}^{ml}$ into $\mathbf{P} \in \mathbb{R}^{m \times l}$. After the completion of the optimization process, the predicted label y of the test example $\mathbf{x}$ by LALO is given as follows:

$$y = \arg\max_{k \in [l]} \sum_{i=1}^{m} a_{ik}\mathcal{K}(\mathbf{x}, \mathbf{x}_i) + b_k \tag{14}$$

The pseudo code of LALO is presented in Algorithm 1. Since the proposed formulation (5) is biconvex, it can be solved by the alternating optimization method with guaranteed convergence [Gorski *et al.*, 2007], and we set the maximum number of iterations as 50.

5 Experiments

5.1 Experimental Setup

In this section, we conduct extensive experiments on artificial (i.e., controlled UCI datasets) and real-world datasets to evaluate the performance of LALO. The main characteristics of these datasets are reported in Table 1.

Following the widely-used controlling protocol [Chen *et al.*, 2014; Cour *et al.*, 2011; Liu and Dietterich, 2012; Yu and Zhang, 2016; Zhang and Yu, 2015; Zhang *et al.*, 2016; Zhang *et al.*, 2017], each UCI dataset is controlled by three parameters p, r and ϵ to generate artificial PL datasets. Here, p controls the proportion of training examples that are partially labeled, r controls the number of false positive labels within the candidate label set, and ϵ controls the co-occurring probability of a specific false positive label and the ground-truth label.

In addition, we have also collected 5 real-world PL datasets[1], including Soccer Player [Zeng *et al.*, 2013], Lost [Cour *et al.*, 2011], Yahoo! News [Guillaumin *et al.*, 2010], FG-NET [Panis and Lanitis, 2014], and MSCRCv2 [Liu and Dietterich, 2012]. These real-world datasets come from several application domains. For *automatic face naming* (Lost, Soccer Player and Yahoo! News), each face cropped from an image or a video frame is considered as an instance and the names extracted from the corresponding image captions or video subtitles work as candidate labels. For *objective classification* (MSRCv2), image segments are considered as instances and objects appearing in the same image work as candidate labels. For *facial age estimation* (FG-NET), each human face is represented as an instance, and the age annotations obtained by crowdsouring are candidate labels. Besides, the average number of the candidate labels (Avg. CLs) for each real-world dataset is also recorded in Table 1.

The performance of LALO is compared with five state-of-the-art partial label learning algorithms, each configured with recommended parameters according to the respective literature:

- PL-KNN [Hüllermeier and Beringer, 2006]: an k-nearest neighbor approach following the average-based framework. (Recommended configuration: $k = 10$).
- CLPL [Cour *et al.*, 2011]: a parametric approach following the average-based framework. (Recommended configuration: SVM with squared hinge loss).
- IPAL [Zhang and Yu, 2015]: an instance-based approach following the average-based framework. (Recommended configuration: $\alpha = 0.95, k = 10, T = 100$)
- PL-SVM [Nguyen and Caruana, 2008]: a maximum margin approach following the identification-based framework. (Recommended configuration: regularization parameter pool with $\{10^{-3}, \cdots, 10^3\}$).
- LSB-CMM [Liu and Dietterich, 2012]: a maximum likelihood approach following the identification-based framework. (Recommended configuration: l mixture components).

The parameters employed by LALO are set as $k = 10, \lambda = 0.05, \mu = 0.005$. The sensitivity analysis of LALO's parameter configuration is conducted in Subsection 5.3. On each artificial and real-world dataset, ten runs of 50%/50% random train/test splits are performed, and the averaged accuracies (with standard deviations) are recorded for all algorithms. In addition, we use the t-test at 0.05 significance level for two independent samples to investigate whether LALO is significantly superior/inferior to the comparing algorithms for all experiments.

5.2 Experimental Results

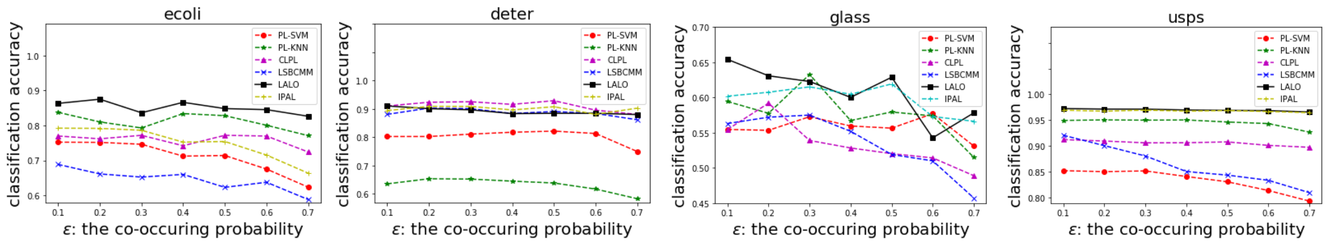
Controlled UCI Datasets

Figure 1 reports the classification accuracy of each algorithm as ϵ ranges from 0.1 to 0.7 with step size 0.1 when p and r are

[1]These data sets are publicly avaible at: http://cse.seu.edu.cn/PersonalPage/zhangml/Resources.htm#partial_data

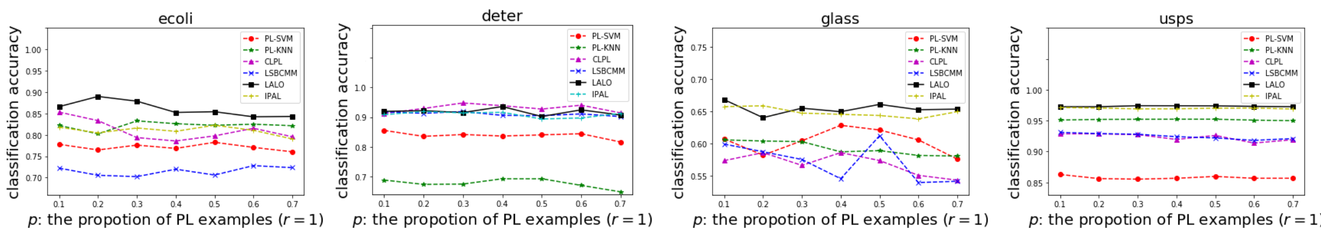
	controlled UCI datasets				real-world datasets				
Dataset	glass	usps	letter	deter	Lost	FG-NET	MSRCv2	Soccer Player	Yahoo! News
Examples	214	9298	20000	358	1122	1002	1758	17472	22991
Features	10	256	16	23	108	262	48	279	163
Classes	5	10	26	6	16	78	32	171	219
Avg. CLs	-	-	-	-	2.23	7.48	3.16	2.09	1.91

Table 1: Characteristics of the experimental datasets.



(a) ecoli (b) deter (c) glass (d) usps

Figure 1: Classification performance on controlled UCI datasets with ϵ ranging from 0.1 to 0.7 ($p = 1, r = 1$).



(a) ecoli (b) deter (c) glass (d) usps

Figure 2: Classification performance on controlled UCI datasets with p ranging from 0.1 to 0.7 ($r = 1$).

both fixed at 1. For each ground-truth label $y \in \mathcal{Y}$, one extra label $y' \neq y$ is selected as the coupled label that co-occurs with y in the candidate label set with probability ϵ, and any other label is chosen to be the false positive label with the probability $1 - \epsilon$. Figure 2 reports the classification accuracy of each algorithm as p ranges from 0.1 to 0.7 with step size 0.1 when r is set to 1. In this setting, r labels are randomly selected as the false positive labels in the candidate label set for the PL examples. In addition, we also do experiments on controlled UCI datasets as p ranges from 0.1 to 0.7 with r set to 2 and 3. Due to the limited space, these results are not reported here[2], while they are quite similar to that in Figure 2 ($r = 1$).

As shown in Figure 1 to 2, LALO outperforms the comparing algorithms in most cases. Besides, the detailed win/tie/loss counts between LALO and other comparing algorithms are recorded in Table 2. Out of the 112 results (4 UCI datasets×28 configurations), we can find that LALO can achieve superior or at least comparable performance against all comparing algorithms in most cases, and lose to them in only a few cases.

	PL-KNN	CLPL	IPAL	PL-SVM	LSB-CMM
vary ϵ ($p, r = 1$)	25/2/1	21/2/5	16/12/0	26/2/0	23/5/0
vary p ($r = 1$)	26/2/0	24/1/3	13/15/0	27/1/0	23/5/0
vary p ($r = 2$)	26/2/0	21/4/3	11/16/1	28/0/0	21/5/2
vary p ($r = 3$)	24/4/0	21/2/5	23/12/3	27/1/0	21/7/0
In Total	**101/10/1**	**87/9/16**	**63/65/4**	**108/4/0**	**88/22/2**

Table 2: Win/tie/loss counts on the controlled UCI datasets between LALO and the comparing algorithms.

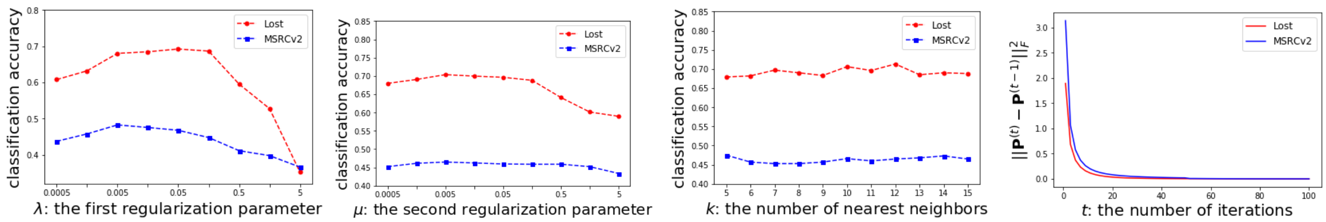
Real-World Datasets

The predictive accuracy of each algorithm on real-world datasets is recorded in Table 3. Note that the average number of candidate labels (Avg. CLs) of the dataset FG-NET is quite large, which causes an extremely low classification accuracy of each algorithm. For better evaluation of this facial age estimation task, two extra experiments are conducted on the dataset FG-NET where a test example is considered to be correctly classified if the difference between the predicted age and the ground-truth age is no more than 3 years

[2]Figures and code package for LALO are publicly available at: https://sites.google.com/site/ramber1995paper/publications

	LALO	PL-KNN	CLPL	IPAL	PL-SVM	LSB-CMM
Lost	0.693±0.024	0.332±0.030•	0.670±0.024	0.576±0.035•	0.639±0.056•	0.591±0.019•
MSRCv2	0.465±0.013	0.417±0.012•	0.375±0.020•	0.476±0.019	0.417±0.027•	0.431±0.008•
Soccer Player	0.523±0.005	0.494±0.004•	0.347±0.004•	0.525±0.006	0.430±0.004•	0.506±0.006•
Yahoo! News	0.613±0.004	0.403±0.004•	0.457±0.005•	0.565±0.004•	0.615±0.002	0.594±0.007•
FG-NET	0.073±0.006	0.037±0.008•	0.047±0.017•	0.054±0.006•	0.058±0.010•	0.056±0.008•
FG-NET(MAE3)	0.424±0.011	0.284±0.035•	0.240±0.045•	0.347±0.021•	0.343±0.022•	0.344±0.026•
FG-NET(MAE5)	0.569±0.020	0.438±0.033•	0.343±0.055•	0.512±0.020•	0.473±0.016•	0.478±0.025•

Table 3: Classification accuracy of each algorithm on the real-world datasets. Furthermore, •/○ indicates whether LALO is statistically superior/inferior to the comparing algorithm.



(a) Vary λ ($\mu = 0.005, k = 10$) (b) Vary μ ($\lambda = 0.05, k = 10$) (c) Vary k ($\lambda = 0.05, \mu = 0.005$) (d) Convergence curves

Figure 3: Parameter sensitivity analysis of LALO on the real-world datasets Lost and MSRCv2.

(MAE3) or 5 years (MAE5). As shown in Table 3, it is obvious that LALO significantly outperforms all the counterpart algorithms on these real-world datasets except for CLPL on Lost, PL-SVM on Yahoo! News, and IPAL on MSRCv2 and Soccer Player, and LALO is never significantly outperformed by any comparing algorithm.

5.3 Sensitivity Analysis

We also study the sensitivity of LALO with respect to its three parameters λ, μ and k. Figure 3 shows the performance of LALO under different parameter configurations. From Figure 3, we can easily find that the parameter configuration specified for LALO in Subsection 5.1 ($\lambda = 0.05$, $\mu = 0.005$, $k = 10$) naturally follows the sensitivity curves. In addition, Figure 3 also reports the difference of the label distribution matrix $\mathbf{P}$ between two successive iterations. We can easily observe that $\left\|\mathbf{P}^{(t)} - \mathbf{P}^{(t-1)}\right\|_F^2$ gradually decreases to 0 as t increases. Therefore, the convergence of LALO is demonstrated.

6 Conclusion

In this paper, we propose a novel unified partial label learning framework and present a biconvex formulation to leverage the latent label distributions for the model training. Extensive experimental results validate the effectiveness of the proposed approach named LALO. Since LALO serves as a bridge between the model training and label propagation, this work can be naturally extended to inductive semi-supervised learning based on label propagation. Besides, it is also interesting to exploit the consistency of the feature space and the label space in other manners.

Acknowledgements

The authors want to thank Prof. Min-Ling Zhang for help on the controlled UCI datasets and the comparing algorithms. This work was supported by MOE, NRF, and NTU.

A Quadratic Programming Formulation

To solve the problem (12), we let $\widetilde{\mathbf{p}} = \text{vec}(\mathbf{P}) \in [0,1]^{ml}$ where $\text{vec}(\cdot)$ is the vectorization operator. Likewise, $\widetilde{\mathbf{q}} = \text{vec}(\mathbf{Q}) \in \mathbb{R}^{ml}$ and $\widetilde{\mathbf{y}} = \text{vec}(\mathbf{Y}) \in \{0,1\}^{ml}$. To deal with the equality constraint using $\widetilde{\mathbf{p}}$, we pick up the indices of $\widetilde{\mathbf{p}}$ by defining a set $\mathcal{C} = \{\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_{m-1}\}$ as follows:

$$j \in \mathcal{C}_i \quad \text{if } j\%m = i, \forall\, j \in [ml] \tag{15}$$

Using these notations, the problem (12) can be written as:

$$\begin{aligned} \min_{\widetilde{\mathbf{p}}} \; & \frac{1}{2}\widetilde{\mathbf{p}}^\top \widetilde{\mathbf{H}}\widetilde{\mathbf{p}} + \widetilde{\mathbf{f}}^\top \widetilde{\mathbf{p}} \\ \text{s.t.} \; & \sum\nolimits_{j \in \mathcal{C}_i} \widetilde{p}_j = 1, \quad \forall\, \mathcal{C}_i \subseteq \mathcal{C} \\ & \mathbf{0}_{ml} \leq \widetilde{\mathbf{p}} \leq \widetilde{\mathbf{y}} \end{aligned} \tag{16}$$

where $\widetilde{\mathbf{f}} = -2\widetilde{\mathbf{q}}$, and $\widetilde{\mathbf{H}} \in \mathbb{R}^{ml \times ml}$ is defined as follows:

$$\widetilde{\mathbf{H}} = \begin{bmatrix} \widetilde{\mathbf{T}} & \mathbf{0}_{m\times m} & \cdots & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} & \widetilde{\mathbf{T}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} & \cdots & \mathbf{0}_{m\times m} & \widetilde{\mathbf{T}} \end{bmatrix} \tag{17}$$

Here, $\widetilde{\mathbf{T}}$ is a square matrix defined by $\widetilde{\mathbf{T}} = 2((\mu+1)\mathbf{I}_{m\times m} - \mu\mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}) \in \mathbb{R}^{m\times m}$ where $\mathbf{D}$ is a diagonal matrix with its diagonal element defined by $d_{ii} = \sum_j s_{ij}$. In this way, the optimization problem (16) can be efficiently solved by any off-the-shelf QP toolbox.

References

[Chen *et al.*, 2014] 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):2076–2088, 2014.

[Chen *et al.*, 2017] Ching-Hui Chen, Vishal M Patel, and Rama Chellappa. Learning from ambiguously labeled face images. *arXiv preprint arXiv:1702.04455*, 2017.

[Cour *et al.*, 2009] Timothee Cour, Benjamin Sapp, Chris Jordan, and Ben Taskar. Learning from ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 919–926. Miami, FL, 2009.

[Cour *et al.*, 2011] Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12(5):1501–1536, 2011.

[Gong *et al.*, 2017] Chen Gong, Tongliang Liu, Yuanyan Tang, Jian Yang, Jie Yang, and Dacheng Tao. A regularization approach for instance-based superset label learning. *IEEE Transactions on Cybernetics*, 2017.

[Gorski *et al.*, 2007] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: A survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

[Guillaumin *et al.*, 2010] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multiple instance metric learning from automatically labeled bags of faces. *Lecture Notes in Computer Science*, 63(11):634–647, 2010.

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

[Hüllermeier and Cheng, 2015] Eyke Hüllermeier and Weiwei Cheng. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer, 2015.

[Jin and Ghahramani, 2003] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems*, pages 921–928, 2003.

[Liu and Chang, 2009] Wei Liu and Shih-Fu Chang. Robust multi-class transductive learning with graphs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 381–388, 2009.

[Liu and Dietterich, 2012] Li-Ping Liu and Thomas G Dietterich. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems*, pages 548–556, 2012.

[Liu and Dietterich, 2014] Li-Ping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, pages 1629–1637, 2014.

[Luo and Orabona, 2010] Jie Luo and Francesco Orabona. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems*, pages 1504–1512, 2010.

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

[Panis and Lanitis, 2014] Gabriel Panis and Andreas Lanitis. An overview of research activities in facial age estimation using the fg-net aging database. In *European Conference on Computer Vision*, pages 737–750, 2014.

[Yu and Zhang, 2016] Fei Yu and Min-Ling Zhang. Maximum margin partial label learning. In *Proceedings of Asian Conference on Machine Learning*, pages 96–111, 2016.

[Zeng *et al.*, 2013] Zi-Nan Zeng, Shi-Jie Xiao, Kui Jia, Tsung-Han Chan, Sheng-Hua Gao, Dong Xu, and Yi Ma. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 708–715, 2013.

[Zhang and Yu, 2015] 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, 2015.

[Zhang *et al.*, 2016] Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. Partial label learning via feature-aware disambiguation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1335–1344, 2016.

[Zhang *et al.*, 2017] Min-Ling Zhang, Fei Yu, and Cai-Zhi Tang. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2155–2167, 2017.

[Zhou *et al.*, 2004] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2004.

[Zhu and Goldberg, 2009] Xiao-Jin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.