PP-PLL: Probability Propagation for Partial Label Learning

Kaiwei Sun, Zijian Min, and Jin Wang (✉)

Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
{sunkw,wangjin}@cqupt.edu.cn
{s170201098}@stu.cqupt.edu.cn

Abstract. Partial label learning (PLL) is a weakly supervised learning framework which learns from the data where each example is associated with a set of candidate labels, among which only one is correct. Most existing approaches are based on the disambiguation strategy, which either identifies the valid label iteratively or treats each candidate label equally based on the averaging strategy. In both cases, the disambiguation strategy shares a common shortcoming that the ground-truth label may be overwhelmed by the false positive candidate labels, especially when the number of candidate labels becomes large. In this paper, a probability propagation method for partial label learning (PP-PLL) is proposed. Specifically, based on the manifold assumption, a biconvex regular function is proposed to model the linear mapping relationships between input features and output true labels. In PP-PLL, the topological relations among training samples are used as additional information to strengthen the mutual exclusiveness among candidate labels, which helps to prevent the ground-truth label from being overwhelmed by a large number of candidate labels. Experimental studies on both artificial and real-world data sets demonstrate that the proposed PP-PLL method can achieve superior or comparable performance against the state-of-the-art methods.

Keywords: Partial label learning · Disambiguation strategy · Manifold assumption · Biconvex regular function.

1 Introduction

In many real-world scenarios, data with explicit label information is hard to obtain. Thus, we have to face with the problem of learning from ambiguous data. Recently, partial label learning (PLL) provides an effective solution to cope with this problem and has been widely used in many real-world applications such as automatic image annotation [3], web mining [13], ecoinformatics [12], etc. Partial label learning is regarded as a weakly-supervised learning where each sample is associated with a set of candidate labels, among which only one is correct [2]. During the training process, the correct label of each training sample is concealed in its candidate label set and not directly accessible to the learning algorithm.

Since the exact labeling information is concealed in the candidate label set, the key to partial label learning is to disambiguate labels in candidate label set. To this end, many disambiguation methods have been proposed to extract the ground-truth label from the ambiguously labeled data. These methods can be categorized into two groups, i.e. identification based disambiguation strategies (IDS) and averaging based disambiguation strategies (ADS). The IDS methods regard the ground-truth label as a latent variable which is identified via iterative refining procedure [10, 12, 15, 17, 19]. The ADS methods treat each candidate label equally and make the final prediction by averaging the modeling outputs [2, 21]. Although IDS and ADS methods have yielded relatively good performance for partial label learning, they still suffer from some defects. Due to some misleading information in the candidate label set, both IDS and ADS methods have the risk that the ground-truth label may be overwhelmed by false positive labels, especially when the number of partially labeled training samples or the size of candidate label set becomes large[19].

To extract as much useful information about the ground-truth label as possible from the partially labeled data, many weakly-supervised learning algorithms assume that there exists a potential structure in the feature space of data, which helps to reveal the mapping from input features to ground-truth labels. Clustering based assumption and manifold based assumption are among the most common ones of them[24]. In the clustering based assumption, data samples are clustered into several clusters based on some similarity criterion such as Euclidean distance, and samples within the same cluster are assumed to belong to the same label. The manifold based assumption can be viewed as the extension of clustering based assumption. It assumes that the feature space of data follows a manifold structure, and the output of each sample is similar to its neighbors. Furthermore, manifold assumption based disambiguation strategies (MADS) have also been proposed to alleviate the negative impact of false positive labels [5, 14, 19, 21]. However, the existing MADS methods ignore the mapping relationships from input features to ground-truth label and excessively rely on the potential topological structure of feature space, which makes the prediction trend to be the frequent labels.

In this paper, a probability propagation method for partial label learning (PP-PLL) is proposed. In PP-PLL, based on the manifold assumption we further assume that neighboring samples have similar label distribution, and we utilize the maximum entropy model to form a biconvex objective function. The objective function is then optimized by the alternating method, which can be regarded as a process of probability propagation. Different from the strategies mentioned above, our proposed PP-PLL method utilizes the potential topological structure of feature space as additional information, which strengthens the exclusiveness among labels and mitigates the risk of the ground-truth label being overwhelmed by candidate labels. Furthermore, in the process of probability propagation the mapping from input features to the ground-truth labels is modeled, which makes it less dependent on the intrinsic topological, and more accurately distinguishes the ground-truth label from false positive labels in the

candidate label set. Compared with many state-of-the-art partial label learning methods, our proposed method can achieve better generalization performance and superior prediction performance.

The rest of this paper is organized as follows. In Section 2, we briefly introduce related works. The concrete formulation of our proposed PP-PLL method is presented in Section 3. In Section 4, the optimization of our model is presented. Section 5 provides experimental studies on various data sets, followed by the conclusions and future works in Section 6.

2 Related work

In partial label learning framework, the label information is no longer unique and explicit. Real semantic information is concealed in the candidate label set, making the learning from data extremely difficult. Existing methods for partial label learning can be roughly grouped into three categories: ADS (Averaging-based Disambiguation Strategies) ,IDS (Identification-based Disambiguation Strategies) and MADS (Manifold Assumption-based Disambiguation Strategies).

ADS methods identify the ground-truth label via giving the label in candidate label set the same weight for each sample, and then obtain prediction by averaging the outputs from all candidate labels or the candidate labels in its neighbors. Following such strategy, ADS methods can be further divided into discrimination-based learning and instance-based learning. For the discrimination-based learning, Cour et al.[2, 3] suppose that a parametric model $F(\mathbf{x}_i, y; \theta)$ discriminates the average modeling output of candidate labels from non-candidate labels as much as possible. For the instance-based learning, Hüllermeier and Beringer[9] suppose that the model predicts unseen instance by aggregating the weight of its neighbors' candidate labels. Although ADS methods are intuitive with strong explanatory, the critical defect is that the false positive labels in each set of candidate labels have greater advantages in weight assignment, especially when the size of each candidate label set becomes large.

Different from ADS, existing IDS approaches consider the ground-truth label as a latent variable, determined directly as $\hat{y}_i = \arg\max_{y \in \mathcal{S}_i} F(\boldsymbol{x}_i, y; \boldsymbol{\theta})$. Furthermore, the objective function is defined according to the maximum likelihood criterion [7, 10, 12, 23]: $\sum_{i=1}^{m} \log\left(\sum_{y \in S_i} F(\boldsymbol{x}_i, y; \boldsymbol{\theta})\right)$ which is generally refined iteratively via utilizing Expectation-Maximization (EM) procedure [4], or the maximum margin criterion [15, 18]: $\sum_{i=1}^{m}\left(\max_{y \in S_i} F(\boldsymbol{x}_i, y; \boldsymbol{\theta}) - \max_{y \notin S_i} F(\boldsymbol{x}_i, y; \boldsymbol{\theta})\right)$ which is optimized via the Pegasos method that alternately performs sub-gradient descent and projection operations to update the model iteratively. Experimental results demonstrate that IDS have achieved more desirable performance than ADS. Nonetheless, the information from the false positive labels in all sets of candidate labels would mislead the model into updating towards the wrong direction, especially when the number of partially labeled training samples become large.

The strategies mentioned above utilize the set of candidate labels to construct partial label learning algorithms. However, their performance improvements are

usually limited by false positive labels. To break through this limitation, manifold assumption-based disambiguation strategies (MADS) are proposed to extract as much useful labeling information as possible from the ambiguously labeled data through manifold assumption. To the best of our knowledge, the concept of neighbor samples in partial label learning was first proposed by Hüllermeier and Beringer [9]. However, it is unable to guarantee that the prediction of each sample is similar to its neighbors. This is why we generalize it into ADS. Following manifold assumption, existing MADS can be divided into nonparametric and parametric model. Regardless of the model proposed, a weighted graph of k-nearest neighbors should be constructed at first stage. At second stage, the prediction is obtained directly by label propagation [5, 19] for nonparametric, and by a feature-aware disambiguation for parametric model [21]. Different from IDS and ADS, MADS can extract additional information from the ambiguously labeled data, however, existing MADS excessively relies on the potential topological structure of feature space.

In the next section, a novel partial label learning approach named PP-PLL will be introduced. To address the problem mentioned above, PP-PLL utilizes the character of the optimizing a biconvex formulation presented in this paper to achieve probability propagation.

3 The PP-PLL Method

Let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional feature space, and $\mathcal{Y} = \{1, 2, \ldots, q\}$ be a label set with q class labels. Partial label learning is aimed at learning a classifier $f : \mathcal{X} \to \mathcal{Y}$ from training data $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) \,|1 \leq i \leq m\}$ to predict the ground-truth label of the unseen samples, where $\boldsymbol{x}_i \in \mathcal{X}$ is a d-dimensional feature vector $(x_{i1}, x_{i2}, \ldots, x_{id})^\top$, and $S_i \subseteq \mathcal{Y}$ is the candidate label set associated with $\boldsymbol{x}_i$. The ground-truth label y_i for $\boldsymbol{x}_i$ is concealed in S_i, i.e. $y_i \in S_i$, and is not directly accessible to the learning algorithm.

Let $\mathcal{F}$ denote the set of $m \times q$ matrices with nonnegative entries. A matrix $\boldsymbol{F} = \left[F_1^\top, \ldots, F_m^\top\right]^\top \in \mathcal{F}$ corresponds to ultimate label probability distribution of m partial label samples, and each sample $\boldsymbol{x}$ is labeled as $\hat{y}_i = \arg\max_{j \leq q} F_{ij}$. Therefore, one of the main goals is to obtain the ultimate label distribution matrix $\boldsymbol{F}$. To this end, some existing partial label learning approaches [7, 10, 12] regard the ground-truth label as a latent variable and estimate the ground-truth label by an iterative procedure. Although this kind of strategies have the capability of mapping from input features to ground-truth label, they are failed to correct the wrong updating direction caused by false positive labels during the iterative learning process.

Accordingly, we proposed PP-PLL under the assumption that the probability distribution of candidate labels for each sample is similar to its neighbors. At first stage, we construct a weighted graph $G = (V, E)$ over the ambiguously labeled data, where each sample is considered as a node of the graph. In order to characterize the manifold structure of feature space via conducting some affinity relationship, $E = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \,|\boldsymbol{x}_i \in k\mathrm{NN}(\boldsymbol{x}_j), i \neq j\}$ is denoted as

the set of directed edges from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$ in graph G if $\boldsymbol{x}_i$ belongs to the k-nearest neighbors of $\boldsymbol{x}_j$. Furthermore, $\mathbf{W} = [w_{ij}]_{m\times m}$ is denoted as the non-negative weight matrix where $w_{ij} = 0$ if $(\boldsymbol{x}_i, \boldsymbol{x}_j) \notin E$. Otherwise, the j-th weight column $\mathbf{w}_{\cdot j} = (w_{i_1 j}, w_{i_2 j}, \ldots, w_{i_k j})^\top$ is denoted as the k-nearest neighbors' optimal weight column corresponding to the j-th sample via optimizing the following linear least square problem:

$$\min_{\mathbf{w}_{\cdot j}} \quad \left\| \boldsymbol{x}_j - \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in E} w_{ij} \cdot \boldsymbol{x}_i \right\|_2^2 \tag{1}$$

$$\text{s.t.} \quad w_{ij} \geq 0 \quad (\forall\, (\boldsymbol{x}_i, \boldsymbol{x}_j) \in E, 0 \leq i, j \leq m)$$

The OP(1) can be re-written as

$$\min_{\mathbf{w}_{\cdot j}} \quad \left(\boldsymbol{x}_j - \boldsymbol{X}_j^\top \cdot \mathbf{w}_{\cdot j}\right)^\top \cdot \left(\boldsymbol{x}_j - \boldsymbol{X}_j^\top \cdot \mathbf{w}_{\cdot j}\right) \tag{2}$$

As shown in OP(2), the $k \times d$ matrix $\boldsymbol{X}_j = (\boldsymbol{x}_{\boldsymbol{i_1}}, \boldsymbol{x}_{\boldsymbol{i_2}}, \ldots, \boldsymbol{x}_{\boldsymbol{i_k}})^\top$ denotes the k-nearest neighbors of $\boldsymbol{x}_j$. We further convert OP(2) into a standard quadratic programming (QP) problem:

$$\min_{\mathbf{w}_{\cdot j}} \quad \frac{1}{2}\mathbf{w}_{\cdot j}^\top \left(2\boldsymbol{X}_j \boldsymbol{X}_j^\top\right) \mathbf{w}_{\cdot j} - 2\boldsymbol{x}_j^\top \boldsymbol{X}_j^\top \mathbf{w}_{\cdot j} \tag{3}$$

$$\text{s.t.} \quad w_{ij} \geq 0 \quad (\forall\, (\boldsymbol{x}_i, \boldsymbol{x}_j) \in E, 0 \leq i, j \leq m)$$

Therefore, the optimized weight of OP(3) can be obtained through any off-the-shelf QP method. Although the restriction $\sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in E} w_{ij} = 1$ is to avoid probability divergence during subsequent iterative probability propagation procedure, it would cause some linear combinations of k-nearest neighbors far away from the center sample. As a consequence, we would rather apply the normalization of each weight column than embed restriction $\sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in E} w_{ij} = 1$ for the j-th sample. In other words, for each weight column, we utilize the following normalized column vector to replace primary weight column vector:

$$\mathbf{h}_{\cdot j} = \mathbf{w}_{\cdot j} / \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in E} w_{ij} \quad (0 \leq j \leq m) \tag{4}$$

At second stage, we develop a novel regularization framework that incorporates probabilistic propagation with the maximum entropy model:

$$\mathcal{J}\left(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{F}\right) = \mathcal{L}\left(\mathcal{D}, \boldsymbol{F}, \boldsymbol{\theta}\right) + \lambda \Omega\left(\boldsymbol{\theta}\right) + \mu \mathcal{Q}\left(\boldsymbol{F}\right) \tag{5}$$

As shown in Eq.(5), the first term $\mathcal{L}$ in the object function $\mathcal{J}$ is denoted as fidelity term with a definition of the conditional probability matrix of the

ground-truth labels $\boldsymbol{\mathcal{C}} = [p(y_i = j|\boldsymbol{x}_i, \boldsymbol{\theta})]_{m\times q}$. The definition of $\boldsymbol{\mathcal{C}}$ is shown as:

$$P\left(y_i = j|\boldsymbol{x}_i, \boldsymbol{\theta}\right) = \begin{cases} \exp\left(\boldsymbol{\theta}_j^\top \boldsymbol{x}\right) / \sum_{j'\in S_i} \exp\left(\boldsymbol{\theta}_{j'}^\top \boldsymbol{x}\right), & \text{if } \ j \in S_i \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{d\times q}$ is a parameter matrix learned from the object function $\mathcal{J}$. This term suggests that the finally obtained label distribution matrix $\boldsymbol{F}$ is closed to the maximum entropy model which builds a linear discriminative mapping from input features to ground-truth labels smoothly. Meanwhile, we choose to apply the Kullback-Leibler divergence of $\boldsymbol{F}$ relative to $\boldsymbol{\mathcal{C}}$ rather than the quadratic form to preserve the convex properties of the object function $\mathcal{J}$ with respect to $\boldsymbol{\theta}$. Therefore $\mathcal{L}$ is formalized as:

$$\mathcal{L}\left(\mathcal{D}, \boldsymbol{F}, \boldsymbol{\theta}\right) = \sum_{i=1}^{m} \sum_{j\in S_i} \boldsymbol{F}_{ij} \log \frac{\boldsymbol{F}_{ij}}{\boldsymbol{\mathcal{C}}_{ij}} \tag{7}$$

The second term Ω in the object function $\mathcal{J}$ is aimed at avoiding parameter redundancy caused by conditional probability matrix, which is defined as an Frobenius norm:

$$\Omega\left(\boldsymbol{\theta}\right) = \frac{1}{2} \left\|\boldsymbol{\theta}\right\|_F^2 \tag{8}$$

The third term $\mathcal{Q}\left(\boldsymbol{F}\right)$ in the object function $\mathcal{J}$ is formalized a smoothness constraint which is to ensure the probability distribution candidate labels of each sample not to vary too much from its k-nearest neighbors to satisfy the realization of the manifold assumption. Based on the above description, $\mathcal{Q}$ can be defined as:

$$\mathcal{Q}\left(\boldsymbol{F}\right) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} \boldsymbol{F}_i - \frac{1}{\sqrt{D_{jj}}} \boldsymbol{F}_j \right\|_2^2 \tag{9}$$

where w_{ij} is the similarity weight between the i-th sample and the j-th sample in graph G, and D_{ll} is the l-th diagonal element in diagonal matrix $\mathbf{D} = \operatorname{diag}\left[\sum_{i=1}^{m} w_{i,1}, \sum_{i=1}^{m} w_{i,2}, \ldots, \sum_{i=1}^{m} w_{i,m}\right]$. As shown in Eq.(9), minimizing $\mathcal{Q}$ will force $\boldsymbol{F}_i$ $(i = 1, 2, \ldots, m)$ to get closer to $\boldsymbol{F}_j$ (if $\boldsymbol{x}_j \in k\mathrm{NN}\left(\boldsymbol{x}_i\right)$) when w_{ij} is larger.

Finally, a novel regularization framework that incorporates probabilistic label propagation with maximum likelihood criterion is presented as a constrained optimization problem:

$$\min_{\boldsymbol{\theta}, \boldsymbol{F}} \quad \sum_{i=1}^{m} \sum_{j\in S_i} \boldsymbol{F}_{ij} \log \frac{\boldsymbol{F}_{ij}}{\boldsymbol{\mathcal{C}}_{ij}} + \frac{\lambda}{2} \left\|\boldsymbol{\theta}\right\|_F^2 + \frac{\mu}{2} \sum_{i,j=1}^{n} w_{ij} \left\| \frac{\boldsymbol{F}_i}{\sqrt{D_{ii}}} - \frac{\boldsymbol{F}_j}{\sqrt{D_{jj}}} \right\|_2^2 \tag{10}$$

$$\text{s.t.} \quad \sum_{j=1}^{q} \mathbf{F}_{ij} = 1, \mathbf{F}_{ij} \geq 0, \quad \forall i = 1, 2, \ldots, m$$

4 Optimization

Apparently, OP(10) is convex with respect to $\boldsymbol{\theta}$ when $\boldsymbol{F}$ is fixed, and it is also convex with respect to $\boldsymbol{F}$ when $\boldsymbol{\theta}$ is fixed. Therefore, OP(10) is regarded as a biconvex problem which can be solved in an alternating way [6]. Specifically, we first optimize OP(10) regarding $\boldsymbol{F}$ when $\boldsymbol{\theta}$ is treated as a constant, and then optimize OP(10) regarding $\boldsymbol{\theta}$ when $\boldsymbol{F}$ is substituted by $\boldsymbol{F}^*$ which is the optimized value of $\boldsymbol{F}$ in previous step.

4.1 Updating $\boldsymbol{F}$

When $\boldsymbol{\theta}$ is assumed as a constant, the conditional probability matrix $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{m\times q}$ corresponding to $\boldsymbol{\theta}$ is also considered as a constant.Therefore the optimization of OP(10) can be simplified to

$$\min_{\boldsymbol{F}} \quad \sum_{i=1}^{m}\sum_{j\in S_i} \boldsymbol{F}_{ij}\log\frac{\boldsymbol{F}_{ij}}{\boldsymbol{\mathcal{C}}_{ij}} + \frac{\mu}{2}\sum_{i,j=1}^{n} w_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}\boldsymbol{F}_i - \frac{1}{\sqrt{D_{jj}}}\boldsymbol{F}_j\right\|_2^2 \tag{11}$$

$$\text{s.t.} \quad \sum_{j=1}^{q}\mathbf{F}_{ij} = 1, \mathbf{F}_{ij} \geq 0, \quad \forall i = 1, 2, \ldots, m$$

which is similar to a label propagation problem [22]. The first term of the OP(11) guarantees that the ultimate label distribution $\boldsymbol{F}$ should be close to constant matrix $\boldsymbol{\mathcal{C}}$, which is denoted as the mapping relationship from input features to the ground-truth label. The second term guarantees that the ultimate label distribution $\boldsymbol{F}$ of each sample should be close to its k-nearest neighbors, which satisfies manifold assumption. In this paper, we present another convex function with respect to $\boldsymbol{F}$:

$$\mathcal{OB} = \frac{1}{2}\left\|\boldsymbol{F}-\boldsymbol{\mathcal{C}}\right\|_F^2 + \frac{\mu}{2}\sum_{i,j=1}^{n} w_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}\boldsymbol{F}_i - \frac{1}{\sqrt{D_{jj}}}\boldsymbol{F}_j\right\|_2^2 \tag{12}$$

As shown in $\mathcal{OB}$, the first term in OP(11) is replaced by the quadratic form, which is convex regarding $\boldsymbol{F}$, and the optimal solution of $\mathcal{OB}$ can be obtained directly via derivation rather than traditional Lagrangian method which is time-consuming. Through label propagation method, the obtained optimal solution of $\mathcal{OB}$ is the approximation of the solution of OP(11). Differentiating the $\mathcal{OB}$ below with respect to $\boldsymbol{F}$, we have

$$\left.\frac{\partial\mathcal{OB}}{\partial\boldsymbol{F}}\right|_{\boldsymbol{F}=\tilde{\boldsymbol{F}^*}} = \tilde{\boldsymbol{F}^*} - \boldsymbol{\mathcal{C}} + \mu\left(\tilde{\boldsymbol{F}^*} - \boldsymbol{H}\tilde{\boldsymbol{F}^*}\right) = \mathbf{0} \tag{13}$$

where $\boldsymbol{H}$ is equal to the $m\times m$ matrix $(\mathbf{h}_{\cdot 1}, \mathbf{h}_{\cdot 2}, \cdots, \mathbf{h}_{\cdot m})$. Since $\boldsymbol{I} - \frac{\mu}{1+\mu}\boldsymbol{H}$ is invertible, we have

$$\tilde{\boldsymbol{F}^*} = \frac{1}{1+\mu}\left(\boldsymbol{I} - \frac{\mu}{1+\mu}\boldsymbol{H}\right)^{-1}\boldsymbol{\mathcal{C}} \tag{14}$$

In order to satisfy the constraints in OP(11), $\tilde{\boldsymbol{F}}^*$ is re-scaled into $\boldsymbol{F}^*$ via consulting the sample in the ambiguously labeled data, which is similar to the E-step in PL-EM [10]:

$$\forall 1 \leq i \leq m: \quad \boldsymbol{F}^*_{i,j} = \begin{cases} \tilde{\boldsymbol{F}}^*{}_{i,j} / \sum_{j' \in S_i} \tilde{\boldsymbol{F}}^*{}_{i,j'}, & \text{if } \ j \in S_i \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

4.2 Updating $\boldsymbol{\theta}$

When $\boldsymbol{F} \in \mathbb{R}^{m \times q}$ is replaced by $\boldsymbol{F}^*$, we have

$$\min_{\boldsymbol{\theta}} \quad \sum_{i=1}^{m} \sum_{j \in S_i} \boldsymbol{F}^*_{ij} \log \frac{\boldsymbol{F}^*_{ij}}{\boldsymbol{\mathcal{C}}_{ij}} + \frac{\lambda}{2} \left\| \boldsymbol{\theta} \right\|_F^2 \tag{16}$$

which is optimized via L-BFGS [11]. Apparently, the process of optimizing OP(16) is similar to M-step in PL-EM, which models the mapping relationship from input features to the ground-truth label.

At the beginning of optimization, it is necessary to initialize the conditional probability matrix $\boldsymbol{\mathcal{C}} = [p(y_i = j | \boldsymbol{x}_i, \boldsymbol{\theta})]_{m \times q}$ as follows:

$$p(y_i = j | \boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{cases} \frac{1}{|S_i|} & \text{if } j \in S_i \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Then we iteratively update the parameter $\boldsymbol{\theta}$ by combining label propagation with PL-EM algorithm, which is collectively called probability propagation procedure. During the testing phase, the conditional probability matrix $\boldsymbol{\mathcal{C}}'$ of each unseen sample $\boldsymbol{x}'$ is calculated as:

$$\boldsymbol{\mathcal{C}}' = \left[\exp\left(\boldsymbol{\theta}_k^\top \boldsymbol{x}' \right) / \sum_{j' \in \mathcal{Y}} \exp\left(\boldsymbol{\theta}_{j'}^\top \boldsymbol{x}' \right) \right]_{1 \times q} \tag{18}$$

And then, the ultimate label distribution $\boldsymbol{F}'$ of each unseen sample $\boldsymbol{x}'$ can be calculated according to Eq.(14) and Eq.(15). Finally, the predicted label y' of each unseen sample $\boldsymbol{x}'$ is given as follows:

$$y' = \arg\max_{k \in \mathcal{Y}} \left[\boldsymbol{F}'_{1,k} \right]_{1 \times q} \tag{19}$$

The complete procedure of PP-PLL is presented in Algorithm 1, where we creatively embed alternating optimization method into PL-EM algorithm to update parameter $\boldsymbol{\theta}$. At first, given a partial label training dataset, a weighted graph is constructed via asymmetric k-NN graph (Steps 1-9). And then, an probability propagation procedure based on EM procedure with alternating optimization is implemented to calculate the optimal parameters (Step 10-15). Finally, the predicted label of the unseen data is obtained according to the optimal parameters(Step 16).

Algorithm 1 PP-PLL

Input:
$\mathcal{D}$: the PL training set $\{(\boldsymbol{x}_i, S_i) \,|1 \leq i \leq m\}$
k: the number of nearest neighbors used for the similarity matrix
λ, μ: the parameters trading off each term in the object function
T: the number of iterations
$\boldsymbol{x}'$: the unseen data
Output:
y': the predicted label for $\boldsymbol{x}'$
Process:

Construct weight graph $G = (V, E)$ by the asymmetric k-NN graph with $V = \{\boldsymbol{x}_i | 1 \leq i \leq m\}$ and $E = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \,|\boldsymbol{x}_i \in k\mathrm{NN}(\boldsymbol{x}_j), i \neq j\}$;
Initialize weight matrix $\mathbf{W} = [w_{ij}]_{m \times m}$ with $w_{ij} = 0$;
for $j = 1$ **to** m **do**
  Determine the j-th weight column corresponding to the j-th sample $\hat{\mathbf{w}}_{\cdot j} = (\hat{w}_{i_1 j}, \hat{w}_{i_2 j}, \ldots, \hat{w}_{i_k j})^\top$ via solving OP(3);
  Normalize the $\hat{\mathbf{w}}_{\cdot j}$ to $\hat{\mathbf{h}}_{\cdot j} = \hat{\mathbf{w}}_{\cdot j} / \sum_{a=1}^{k} \hat{w}_{i_a j} = \left(\hat{h}_{i_1 j}, \hat{h}_{i_2 j}, \ldots, \hat{h}_{i_k j}\right)^\top$
  for $\boldsymbol{x}_{i_a} \in k\mathrm{NN}(\boldsymbol{x}_j)$ **do**
    Set $w_{i_a j} = \hat{h}_{i_a j}$;
  end for
end for
Initial $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{q}}$ according to Eq.(17);
for $t = 1$ **to** T **do**
  Update $\boldsymbol{F}$ according to Eq.(15);
  Caculate $\boldsymbol{\theta}$ by solving OP(16);
  Update $\boldsymbol{\mathcal{C}}$ by updated $\boldsymbol{\theta} \in \mathbb{R}^{d \times q}$;
end for
Return the predicted label y' according to Eq.(18) and Eq.(19).

5 Experiments

5.1 Experimental Setup

To verify the performance of the proposed PP-PLL method, we conduct extensive experiments on four controlled UCI datasets and five real-world datasets. Characteristics of the experimental datasets are summarized in Table 1.

Controlled UCI Datasets To generate artificial PL datasets, controlled UCI datasets are controlled by two parameters p and r, where p controls the proportion of partially labeled samples, and r controls the size of distracting labels set in the candidate label set.

Real-world Datasets In addition, we have also collected five real-world datasets which are partially labeled. The real-world datasets can be summarized into four task domains:

- **Bird Song Classification**: Spectrogram of the birds are considered as instances while candidate labels is composed of bird species jointly singing [1].
- **Automatic Face Naming**: Each face recognized from images or a videos are considered as instances and the names extracted from the corresponding image captions or video subtitles are regarded as candidate labels, such as Yahoo! News [8] and Lost [2];
- **Facial Age Estimation**: Human faces constitute the instance space and candidate labels is composed of the ages annotated by ten crowd-sourced labels and the ground-truth ages, such as FG-NET [16];
- **Objective Classification**: Image segmentations are considered as instances and the objects appearing within the same image are represented as the candidate labels, such as MSRCv2 [12].

The average number of the candidate labels (Avg. CLs) for each real-world dataset is also recorded in Table 1.

	Controlled UCI datasets				Real-world datasets				
Dataset	Glass	Ecoli	Segment	Letter	Lost	FG-NET	MSRCv2	BirdSong	Yahoo! News
Examples	214	336	2310	20000	1122	1002	1758	4,998	22991
Features	10	7	18	16	108	262	48	38	163
Classes	7	8	7	26	16	78	32	13	219
Avg. CLs	-	-	-	-	2.23	7.48	3.16	2.18	1.91

Table 1. Characteristics of the experimental datasets

Comparing Algorithms In this paper, the effectiveness of PP-PLL is evaluated against five state-of-the-art partial label learning algorithms, and the recommended parameters for each comparing algorithm in corresponding literature are used in our experiments:

- **PL-KNN** [9]: An k-nearest neighbor approach based on ADS averages the output of respective neighbors to disambiguate the set of candidate labels [Recommended configuration: $k = 10$]
- **PL-SVM** [15]: A maximum margin approach based on IDS incorporates maximum margin to disambiguate the set of candidate labels [Recommended configuration: regularization parameter pool with $\left\{10^{-3}, \cdots, 10^{3}\right\}$]
- **PL-LEAF** [21]: A partial-label learning method disambiguate the set of candidate labels via postulating that the potentially useful information from feature space [Recommended configuration: $k = 10$, $C_1 = 10$, $C_2 = 1$];
- **PL-ECOC** [20]: It learns from partial-label training instances via adapting error-correcting output codes [Recommended configuration: the codeword length $L = \lceil \log_2(q) \rceil$];

– **GM-PLL** [14]: A partial-label learning method disambiguate the set of candidate labels via incorporating the instance relationship and the co-occurrence possibility of varying label basd on Graph Matching (GM) scheme [Recommended configuration:set β among $\{0.3, 0.4, \ldots, 0.8\}$].

The parameters employed by PP-PLL are set as $T = 60$, $k = 10$, $\mu = 1$ and $\lambda = 0.005$, which the analysis of parameter configuration is conducted in Subsection 5.3. In this paper, we perform ten runs of 50%/50% random train/test on four controlled UCI datasets as well as five real-world partial label datasets, and we evaluate comparing algorithms by the mean predictive accuracies (with standard deviation). Furthermore, we adopt pairwise t-test at 0.05 significance level to investigate whether PP-PLL is significantly superior/inferior to the comparing algorithms.

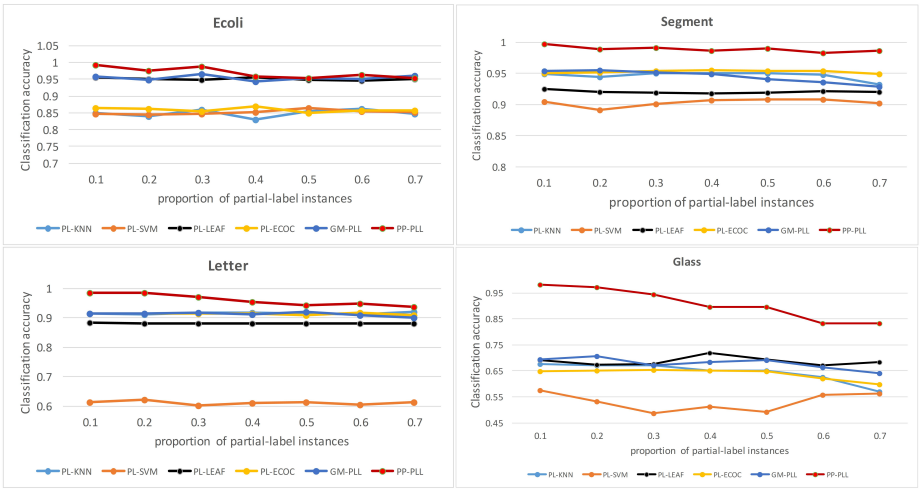

Fig. 1. The classification accuracy of each comparing method on four controlled UCI datasets with stochastic r.

5.2 Experimental Results

Since four controlled UCI datasets are generated manually via two parameters while five real-world datasets are generated via real world scenarios, we perform two series of experiments to evaluate the performance of the proposed method. Meanwhile, the following two subsections exhibit the experimental results separately.

Controlled UCI Datasets In Figure 1, the classification accuracy of each comparing algorithm is illustrated where the probability of generating partial

labeled data p varies from 0.1 to 0.7 with step-size 0.1, while the size of distracting labels set r is randomly selected among $\{1, 2, 3\}$.

From Figure 1, we can see that PP-PLL achieves better classification accuracy than the comparing algorithms in most cases. Table 2 reports the experimental results with fixed value of r, along with the win/tie/loss counts between PP-PLL and other comparing algorithms. The result of statistical comparisons in Table 2 shows that PP-PLL achieves competitive classification performance against other state-of-the-art partial label learning algorithms on most controlled UCI datasets .

	PP-LEAF	PL-KNN	PL-SVM	PL-ECOC	GM-PLL
vary $p(r=1)$	26/1/1	28/0/0	28/0/0	28/0/0	19/7/2
vary $p(r=2)$	26/2/1	26/2/0	28/0/0	26/2/0	14/9/5
vary $p(r=3)$	26/2/0	26/1/1	28/0/0	25/2/1	16/8/4
In Total	**79/3/2**	**80/3/1**	**84/0/0**	**79/4/1**	**49/24/11**

Table 2. Win/tie/loss counts (pairwise t-test at 0.05 significance level) on the controlled UCI datasets between PP-PLL and the comparing algorithms with classification accuracy.

Real-World Datasets We compare the PP-PLL with all above comparing algorithms on the real-world datasets from four task domains mentioned above. The classification performance of each algorithm in terms of accuracy is reported in Table 3. As shown in Table 3, which is classification accuracy of each algorithm on the real-world datasets, it is obvious that PP-PLL achieves superior classification accuracy comparing with all the counterpart algorithms on these real-world datasets except for GM-PLL, PL-SVM and PL-ECOC on Yahoo! News.

	Lost	MSRCv2	Yahoo!News	BirdSong	FG-NET
PP-PLL	0.748 ± 0.031	0.546 ± 0.045	0.554 ± 0.004	0.850 ± 0.24	0.128 ± 0.007
GM-PLL	0.737 ± 0.043 •	0.530 ± 0.019 •	0.629 ± 0.007 ○	0.663 ± 0.010 •	0.065 ± 0.021 •
PL-KNN	0.332 ± 0.030 •	0.417 ± 0.012 •	0.457 ± 0.009 •	0.614 ± 0.024 •	0.037 ± 0.008 •
PL-SVM	0.639 ± 0.056 •	0.417 ± 0.027 •	0.636 ± 0.018 ○	0.662 ± 0.032 •	0.058 ± 0.010 •
PL-ECOC	0.703 ± 0.052 •	0.505 ± 0.027 •	0.662 ± 0.010 ○	0.740 ± 0.016 •	0.040 ± 0.018 •

Table 3. Classification accuracy (mean ± standard deviation) of each algorithm on the real-world datasets. Furthermore, • or ○ is denoted as whether PP-PLL is statistically superior or inferior to the comparing algorithm (pairwise t-test at 0.05 significance level).

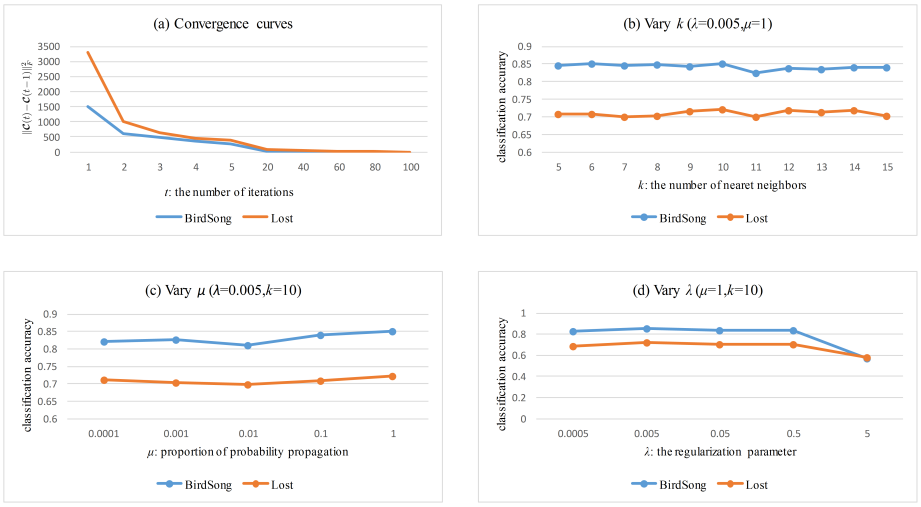

Fig. 2. Parameter sensitivity analysis of PP-PLL on the real-world datasets BirdSong and Lost.

5.3 Sensitivity Analysis

Figure 2 shows the performance of PP-PLL under different parameter configurations, and the convergence of PP-PLL on BirdSong and Lost. As shown in (a), $\|\boldsymbol{\mathcal{C}}(t) - \boldsymbol{\mathcal{C}}(t-1)\|_F^2$ which is the square of Frobenius norm about difference in the conditional probability matrix $\boldsymbol{\mathcal{C}}$ between two continuous iterations gradually approaches 0 as t tends to be infinite. Especially when the number of iterations reaches 20-40 loops, PP-PLL becomes convergent. Therefore, the convergence of PP-PLL is demonstrated, and PP-PLL shows relative stability with the varying of parameters (k, μ, λ) in (b)-(d). In addition, Figure 2 also reports that the parameter configuration specified for Subsection 5.1 $(T = 60, k = 10, \mu = 1, \lambda = 0.005)$ naturally follows from the analysis mentioned above, and makes PP-PLL obtain relatively superior performance compared with other parameter combinations.

6 Conclusion

In this paper, we present a biconvex formulation containing a mapping relationships from input features to the ground-truth label based on manifold assumption, which is optimized by the alternating optimization method, to deal with partial label learning via probability propagation procedure. Extensive experimental results on controlled UCI datasets as well as real-world datasets demonstrate that our proposed method can achieve superior classification performance than the state-of-the-art partial label learning algorithms. However, In terms of weighted graph, how to create a more meaningful weight matrix will be one of

the future directions of partial label learning. It would help all MADS (Manifold Assumption based Disambiguation Strategies) extend to the more special situations, especially when the size of each candidate label set is too large, which causes the information of the ground-truth label in each candidate label set to disappear. For PP-PLL, an important future work is to combine weighted graph with probability distribution of candidate label sets, to improve the availability of candidate label sets.

7 Acknowledgments

This study was supported by National Natural Science Foundation of China (Grant No. 61806033).

References

1. Briggs, F., Fern, X.Z., Raich, R.: Rank-loss support instance machines for miml instance annotation. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 534–542. ACM (2012)
2. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. Journal of Machine Learning Research **12**(May), 1501–1536 (2011)
3. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 919–926. IEEE (2009)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
5. Gong, C., Liu, T., Tang, Y., Yang, J., Yang, J., Tao, D.: A regularization approach for instance-based superset label learning. IEEE transactions on cybernetics **48**(3), 967–978 (2018)
6. Gorski, J., Pfeuffer, F., Klamroth, K.: Biconvex sets and optimization with biconvex functions: a survey and extensions. Mathematical methods of operations research **66**(3), 373–407 (2007)
7. Grandvalet, Y., Bengio, Y.: Learning from partial labels with minimum entropy (2004)
8. Guillaumin, M., Verbeek, J., Schmid, C.: Multiple instance metric learning from automatically labeled bags of faces. In: European Conference on Computer Vision. pp. 634–647. Springer (2010)
9. Hüllermeier, E., Beringer, J.: Learning from ambiguously labeled examples. Intelligent Data Analysis **10**(5), 419–439 (2006)
10. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: Advances in neural information processing systems. pp. 921–928 (2003)
11. Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical programming **45**(1-3), 503–528 (1989)
12. Liu, L., Dietterich, T.G.: A conditional multinomial mixture model for superset label learning. In: Advances in neural information processing systems. pp. 548–556 (2012)
13. Luo, J., Orabona, F.: Learning from candidate labeling sets. In: Advances in neural information processing systems. pp. 1504–1512 (2010)

14. Lyu, G., Feng, S., Wang, T., Lang, C., Li, Y.: Gm-pll: Graph matching based partial label learning. arXiv preprint arXiv:1901.03073 (2019)
15. Nguyen, N., Caruana, R.: Classification with partial labels. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 551–559. ACM (2008)
16. Panis, G., Lanitis, A., Tsapatsoulis, N., Cootes, T.F.: Overview of research on facial ageing using the fg-net ageing database. Iet Biometrics **5**(2), 37–46 (2016)
17. Yi-Chen, C., Patel, V.M., Chellappa, R., Phillips, P.J.: Ambiguously labeled learning using dictionaries. IEEE Transactions on Information Forensics and Security **9**(12), 2076–2088 (2014)
18. Yu, F., Zhang, M.L.: Maximum margin partial label learning. In: Asian Conference on Machine Learning. pp. 96–111 (2016)
19. Zhang, M.L., Yu, F.: Solving the partial label learning problem: An instance-based approach. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
20. Zhang, M.L., Yu, F., Tang, C.Z.: Disambiguation-free partial label learning. IEEE Transactions on Knowledge and Data Engineering **29**(10), 2155–2167 (2017)
21. Zhang, M.L., Zhou, B.B., Liu, X.Y.: Partial label learning via feature-aware disambiguation. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1335–1344. ACM (2016)
22. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in neural information processing systems. pp. 321–328 (2004)
23. Zhou, Y., He, J., Gu, H.: Partial label learning via gaussian processes. IEEE transactions on cybernetics **47**(12), 4443–4450 (2017)
24. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning **3**(1), 1–130 (2009)