Semi-Supervised Partial Label Learning via Confidence-Rated Margin Maximization

Wei Wang    Min-Ling Zhang
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
{wang_w, zhangml}@seu.edu.cn

Abstract

Partial label learning assumes inaccurate supervision where each training example is associated with a set of candidate labels, among which only one is valid. In many real-world scenarios, however, it is costly and time-consuming to assign candidate label sets to all the training examples. To circumvent this difficulty, the problem of semi-supervised partial label learning is investigated in this paper, where unlabeled data is utilized to facilitate model induction along with partial label training examples. Specifically, label propagation is adopted to instantiate the labeling confidence of partial label examples. After that, maximum margin formulation is introduced to jointly enable the induction of predictive model and the estimation of labeling confidence over unlabeled data. The derived formulation enforces confidence-rated margin maximization and confidence manifold preservation over partial label examples and unlabeled data. We show that the predictive model and labeling confidence can be solved via alternating optimization which admits QP solutions in either alternating step. Extensive experiments on synthetic as well as real-world data sets clearly validate the effectiveness of the proposed semi-supervised partial label learning approach.

1 Introduction

In partial label (PL) learning, each training example is represented by a single instance while associated with multiple candidate labels. It is assumed that the ground-truth label of PL training example resides in its candidate label set, which is not directly accessible to the training algorithm [8,17,32]. The need to learn from these inaccurate supervision information widely exists in various applications, such as image classification [6,9,30], ecoinformatics [4,17,33], web mining [18], natural language processing [23,24,35], etc.

Most partial label learning approaches work under supervised setting where the candidate labeling information is available for all training examples. In many real-world scenarios, however, the process of acquiring training examples with candidate labels might be demanding while abundant unlabeled data are readily available to facilitate model training. For instance, in crowdsourced image tagging, acquiring candidate annotations from web users for a large number of images would be costly and time-consuming while abundant unlabeled images can be easily collected from the web. Therefore, it is a natural remedy to consider semi-supervised partial label learning which exploits unlabeled data in conjunction with PL training examples to help induce predictive model with strong generalization performance.

*Corresponding author
Correspondingly, a novel approach named PARM, i.e. semi-supervised Partial label learning via confidence-rated margin Maximization, is proposed in this paper. To make use of unlabeled data, PARM chooses to jointly estimate the labeling confidence over unlabeled data and induce the desired multi-class classification model. Specifically, PARM considers confidence-rated margin which is maximized by preserving labeling confidence manifold structure between PL training examples and unlabeled data. PARM tackles the resulting formulation based on alternating optimization, where the predictive model and labeling confidence are updated in either alternating step with QP solutions. Comparative studies on both synthetic and real-world data sets show that PARM achieves favorable performance against state-of-the-art approaches in exploiting unlabeled data for partial label learning.

To the best of our knowledge, SSPL [27] corresponds to the only prior work which considers utilizing unlabeled data for partial label learning. Specifically, SSPL adopts graph-based techniques to disambiguate the labeling information between PL training examples and unlabeled data via label propagation. Due to the transductive nature of graph-based techniques, the resulting algorithm won’t generalize to make prediction on unseen instances. To account for this issue, kNN rule is further employed to enable prediction on unseen instances. Consequently, SSPL has to store all the disambiguated PL training examples as well as unlabeled data during testing phase, which makes SSPL less efficient in terms of storage overhead and prediction time. Due to the inductive nature of maximum margin approach, PARM is capable of making predictions on unseen examples without resorting to extra procedure.

The rest of this paper is organized as follows. Firstly, we briefly review related work on partial label learning. Secondly, technical details of the proposed approach are presented. Thirdly, experimental results of comparative studies are reported. Finally, we conclude this paper.

2 Related Work

Partial label learning corresponds to the weakly supervised learning problem with inaccurate labeling information [36], where the ground-truth label of each PL training example is concealed within its candidate label set and not directly accessible to the learning algorithm. To learn from PL training examples, a natural strategy is trying to disambiguate the candidate label set. One way to instantiate the disambiguation strategy is to treat the ground-truth label as latent variable, and then identify its value via iterative optimization procedure such as EM. Accordingly, the objective function for identification-based disambiguation can be defined based on the maximum likelihood criterion [15] [17] [19], maximum margin criterion [5] [22] [29], etc. Another way to instantiate the disambiguation strategy is to treat all candidate labels in an equal manner, and then make final prediction by averaging the modeling outputs from all candidate labels. Accordingly, the prediction rule for averaging-based disambiguation can be defined based on convex formulation [8], instance-based formulation [11] [14] [31], etc.

For the disambiguation strategy, one potential issue lies in that the effectiveness of disambiguation would be largely affected by the false positive labels within candidate label set. As the size of candidate label set increases, it is highly possible that the identified ground-truth label might turn out to be false positive one for identification-based disambiguation, while the modeling output from ground-truth label would be overwhelmed by those from false positive labels for averaging-based disambiguation. In light of this, another strategy to learn from PL training examples is trying to transform the partial label learning problem into other well-established learning problems. Accordingly, the transformation strategy can be instantiated based on binary decomposition [28] [32], dictionary learning [7], graph matching [20], regression [10] [26] [33], etc.

To help deal with the difficulty brought by weak supervision, one natural choice is to make use of the unlabeled data which are readily available for model training [36]. Semi-supervised learning [39] aims to make use of unlabeled data for training and typically learns from few labeled training examples together with large amount of unlabeled data. There are four major categories of semi-supervised learning methods, including graph-based approaches [2] [34] [38], disagreement-based approaches [3] [37], generative approaches [21] and low-density separation approaches [1] [16]. Graph-based approaches construct a graph and propagate label information to unlabeled data following the cluster assumption or manifold assumption. Disagreement-based approaches build multiple classifiers and make use of the disagreement among them to facilitate the learning process. Generative approaches treat the missing labels as latent variables and estimate them via iterative process. Low-
We also normalize diag\[
\begin{bmatrix}
W_i
\end{bmatrix}
\] associated with \(x_i\) between unlabeled data and PL training examples is instantiated as: \(w_{ij}^P = \exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)\) if \(x_j \in \mathcal{N}_P(x_i)\) and \(w_{ij}^P = 0\) otherwise. We further normalize \(W_P\) by row to yield the propagation matrix \(H = D_P^{-1}W_P\) with \(D_P = \text{diag}[d_1^P, d_2^P, \ldots, d_n^P]\) and \(d_i^P = \sum_{j=1}^n w_{ij}^P\). Accordingly, the initial labeling confidence matrix \(F_P^{(0)}\) is set as:

\[
\forall 1 \leq i \leq p : \quad f_{ij}^{(0)} = \begin{cases} 
\frac{1}{|S_i|}, & j \in S_i \\
0, & \text{otherwise}
\end{cases}
\]  

Thereafter, the following iterative label propagation procedure is invoked to update \(F_P\) until convergence:

\[
\bar{F}_P^{(t)} = \alpha \cdot HF_P^{(t-1)} + (1 - \alpha) \cdot F_P^{(t-1)}
\]  

\[
\forall 1 \leq i \leq p : \quad f_{il}^{(t)} = \begin{cases} 
\frac{f_{il}^{(t)}}{n_i}, & l \in S_i \\
0, & \text{otherwise}
\end{cases}
\]  

Here, \(\alpha \in (0, 1)\) corresponds to the balancing parameter for label propagation.  

For each unlabeled data \(x_i \in \mathcal{D}_U\), let \(\mathcal{N}_U(x_i)\) be the set of \(x_i\)’s \(k\)-nearest neighbours identified in \(\mathcal{D}_P\). Similarly, we set the similarity matrix \(W_{UP} = [w_{ij}^{UP}]_{n \times p}\) between unlabeled data and PL training examples as: \(w_{ij}^{UP} = \exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)\) if \(x_j \in \mathcal{N}_U(x_i)\) and \(w_{ij}^{UP} = 0\) otherwise. We also normalize \(W_{UP}\) by row to yield \(S = \text{diag}[s_{ij}]_{n \times p}\) such that \(S = D_{UP}^{-1}W_{UP}\) with \(D_{UP} = \text{diag}[d_1^{UP}, d_2^{UP}, \ldots, d_n^{UP}]\) and \(d_i^{UP} = \sum_{j=1}^n w_{ij}^{UP}\).

In this paper, \(\sigma, k\) and \(\alpha\) are fixed to be 1, 8 and 0.95 respectively.
For the proposed PARM approach, the predictive model \( \mathbf{w} \in \mathbb{R}^{nq} \) and the labeling confidence matrix \( \mathbf{F}_U \) over unlabeled data are jointly optimized by solving the following confidence-rated margin maximization problem:

\[
\min_{\mathbf{w}, \Xi, \mathbf{F}_U} \frac{1}{2} \| \mathbf{w} \|_2^2 + \lambda \sum_{i=1}^{p} \sum_{l=1}^{q} f_{il} \xi_{il} + \mu \sum_{u=p+1}^{p+u} \sum_{i=1}^{q} f_{il} \xi_{il} + \gamma \sum_{i=1}^{u} \sum_{j=1}^{p} s_{ij} \| \mathbf{f}_{p+i} - \mathbf{f}_j \|_2^2
\]

\[
\text{s.t. } \mathbf{w}^T \Phi(x_i, y_i) - \max_{y_{i', \neq y_i}} \mathbf{w}^T \Phi(x_i, y_{i'}) \geq 1 - \xi_{il}, \quad (1 \leq i \leq p + u, 1 \leq l \leq q)
\]

\[
\xi_{il} \geq 0, \quad (1 \leq i \leq p + u, 1 \leq l \leq q)
\]

\[
f_{il} \geq 0, \quad (p + 1 \leq i \leq p + u, 1 \leq l \leq q)
\]

\[
\sum_{l=1}^{q} f_{il} = 1, \quad (p + 1 \leq i \leq p + u)
\]

Here, \( \Xi = [\xi_{il}]_{(p+u) \times q} \) corresponds to the set of slack variables with \( \xi_{il} \) characterizing the multi-class classification margin. As shown in the second and third terms of the above objective function, \( \xi_{il} \) is further rated by \( f_{il} \) to account for the labeling confidence of \( y_i \) being the ground-truth label for \( x_i \).

To make full use of available supervision information, the estimated labeling confidence \( \mathbf{F}_P \) over PL training examples are utilized in the fourth term to enforce manifold consistency between \( \mathbf{F}_P \) and \( \mathbf{F}_U \). To solve the derived problem, PARM employs alternating optimization to iteratively update \( \mathbf{w} \) and \( \mathbf{F}_U \).

**Fix \( \mathbf{w} \), Optimize \( \mathbf{F}_U \)**

When \( \mathbf{w} \) is fixed, according to the first and second constraints in Eq. (4), we can have the values of slack variables as:

\[
\xi_{il} = \max \left(0, 1 + \max_{y_{i', \neq y_i}} \mathbf{w}^T \Phi(x_i, y_{i'}) - \mathbf{w}^T \Phi(x_i, y_i)\right)
\]

Thereafter, the optimization problem in Eq. (4) turns out to be:

\[
\min_{\mathbf{F}_U} \mu \sum_{u=p+1}^{p+u} \sum_{i=1}^{q} f_{il} \xi_{il} + \gamma \sum_{i=1}^{u} \sum_{j=1}^{p} s_{ij} \| \mathbf{f}_{p+i} - \mathbf{f}_j \|_2^2
\]

\[
\text{s.t. } f_{il} \geq 0, \quad (p + 1 \leq i \leq p + u, 1 \leq l \leq q)
\]

\[
\sum_{l=1}^{q} f_{il} = 1, \quad (p + 1 \leq i \leq p + u)
\]

Note that Eq. (6) corresponds to a quadratic programming (QP) problem with \( uq \) variables and \( u(q+1) \) constraints, whose computational complexity would be demanding if \( uq \) is large. To improve efficiency, we can decompose Eq. (6) into \( u \) QP sub-problems each with \( q \) variables and \( q+1 \) constraints. Without loss of generality, the labeling confidence vector \( \mathbf{f}_i \) for unlabeled data \( x_i \) \( (p + 1 \leq i \leq p + u) \) can be optimized by fixing the values of other elements in \( \mathbf{F}_U \):

\[
\min_{\mathbf{f}_i} \gamma f_i^T f_i + \left( \frac{\mu}{u} \xi_i - 2\gamma \sum_{j=1}^{p} s_{ij} f_j \right)^T f_i
\]

\[
\text{s.t. } f_{il} \geq 0, \quad (1 \leq l \leq q)
\]

\[
\sum_{l=1}^{q} f_{il} = 1
\]

Here, \( \xi_i = [\xi_{i1}, \xi_{i2}, \ldots, \xi_{iq}]^T \) is the slack vector for \( x_i \).

**Fix \( \mathbf{F}_U \), Optimize \( \mathbf{w} \)**

When \( \mathbf{F}_U \) is fixed, the optimization problem in Eq. (4) turns out to be:

\[
\min_{\mathbf{w}, \Xi} \frac{1}{2} \| \mathbf{w} \|_2^2 + \lambda \sum_{i=1}^{p} \sum_{l=1}^{q} f_{il} \xi_{il} + \mu \sum_{u=p+1}^{p+u} \sum_{i=1}^{q} f_{il} \xi_{il}
\]

\[
\text{s.t. } \mathbf{w}^T \Phi(x_i, y_i) - \max_{y_{i', \neq y_i}} \mathbf{w}^T \Phi(x_i, y_{i'}) \geq 1 - \xi_{il}, \quad (1 \leq i \leq p + u, 1 \leq l \leq q)
\]

\[
\xi_{il} \geq 0, \quad (1 \leq i \leq p + u, 1 \leq l \leq q)
\]
For simplicity, the first and second constraints in Eq. (6) can be rewritten as: $w^T \Phi(x_i, y_t) + \delta_{tl} - w^T \Phi(x_i, y_r) \geq 1 - \xi_{il} \leq 1 \leq p + u$, $1 \leq l, r \leq q$. Here, $\delta_{tl} = 1$ if $l = r$ and $\delta_{tl} = 0$ otherwise. Then, the Lagrangian of Eq. (9) corresponds to:

$$
L(w, \Xi, \alpha) = \frac{1}{2} \|w\|^2 + \lambda \sum_{i=1}^{p} \sum_{l=1}^{q} f_{il} \xi_{il} + \mu \sum_{u=0}^{p+u} q \sum_{i=1}^{q} \alpha_{il}^r (w^T \Phi(x_i, y_r) - w^T \Phi(x_i, y_t) - \delta_{tl} + 1 - \xi_{il})
$$

where $\alpha = [\alpha_1^1, \ldots, \alpha_l^r, \ldots, \alpha_n^{p+u}]^T$ correspond to the Lagrangian multipliers with $\alpha_{il}^r \geq 0 \leq 1 \leq p + u$, $1 \leq l, r \leq q)$. By setting the gradient of $L(w, \Xi, \alpha)$ w.r.t. $w$ and $\Xi$ to zero, we can have the dual problem of Eq. (8) as follows:

$$
\min_{\alpha} \frac{1}{2} \sum_{i=1}^{p+u} \sum_{j=1}^{p+u} x_i^T x_j \sum_{l=1}^{q} \left( \sum_{r=1}^{q} \alpha_{il}^r - \sum_{r=1}^{q} \alpha_{il}^r \right) \left( \sum_{r=1}^{q} \alpha_{jl}^r - \sum_{r=1}^{q} \alpha_{jl}^r \right) + \sum_{i=1}^{p+u} \sum_{j=1}^{p+u} \sum_{r=1}^{q} \alpha_{il}^r \delta_{tl} \quad \text{s.t. } \alpha_{il}^r \geq 0, \quad (1 \leq i \leq p + u, 1 \leq l, r \leq q)
$$

Note that Eq. (10) is a QP problem with $(p + u)q^2$ variables and $(p + u)q^2$ constraints, which would be difficult to be efficiently solved when $p + u$ or $q$ is large. Therefore, we decompose Eq. (10) into $p + u$ sub-problems each with $q^2$ variables and $q^2$ constraints. For ease of notation, we group the Lagrangian multipliers w.r.t. $x_i$ into $\alpha^i = [\alpha_{i1}^1, \ldots, \alpha_{iq}^q]$ and introduce the following terms $M \in \{0, 1\}^{q \times q^2}$ and $N \in \{0, 1\}^{q \times q^2}$:

$$
\alpha^i = \begin{bmatrix}
\alpha_{i1}^1 & \alpha_{i2}^1 & \cdots & \alpha_{iq}^1 \\
\alpha_{i1}^2 & \alpha_{i2}^2 & \cdots & \alpha_{iq}^2 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_{i1}^q & \alpha_{i2}^q & \cdots & \alpha_{iq}^q
\end{bmatrix}, \quad M = [I_{q \times q}, \ldots, I_{q \times q}], \quad N = \begin{bmatrix}
I_{1 \times q} & 0_{1 \times q} & \cdots & 0_{1 \times q} \\
0_{1 \times q} & I_{1 \times q} & \cdots & 0_{1 \times q} \\
\vdots & \vdots & \ddots & \vdots \\
0_{1 \times q} & 0_{1 \times q} & \cdots & I_{1 \times q}
\end{bmatrix}
$$

Here, $I_{q \times q}$ is the identity matrix. Without loss of generality, $\alpha^i$ can be optimized by fixing the values of other Lagrangian multipliers in $\alpha$:

$$
\min_{\alpha^i} \frac{1}{2} x_i^T X_i \text{vec}(\alpha^i)^T C^T C \text{vec}(\alpha^i) + \left( \sum_{j \neq i} x_j^T x_j C^T C \text{vec}(\alpha^j) + \text{vec}(I_{q \times q}) \right)^T \text{vec}(\alpha^i) \quad \text{s.t. } \alpha_{il}^r \geq 0, \quad (1 \leq l, r \leq q)
$$
Table 2: Characteristics of the experimental data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th># Examples</th>
<th># Features</th>
<th># Class Labels</th>
<th># False Positive Labels (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deter</td>
<td>358</td>
<td>23</td>
<td>6</td>
<td>(r = 1, 2, 3)</td>
</tr>
<tr>
<td>Vehicle</td>
<td>846</td>
<td>18</td>
<td>4</td>
<td>(r = 1, 2)</td>
</tr>
<tr>
<td>Abalone</td>
<td>4,177</td>
<td>7</td>
<td>29</td>
<td>(r = 1, 2, 3)</td>
</tr>
<tr>
<td>Satimage</td>
<td>6,435</td>
<td>36</td>
<td>7</td>
<td>(r = 1, 2, 3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set</th>
<th># Examples</th>
<th># Features</th>
<th># Class Labels</th>
<th>Avg. # CLs</th>
<th>Task Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lost</td>
<td>1,122</td>
<td>108</td>
<td>16</td>
<td>2.23</td>
<td>automatic face naming</td>
</tr>
<tr>
<td>Mirflickr</td>
<td>2,780</td>
<td>1536</td>
<td>14</td>
<td>2.76</td>
<td>web image classification</td>
</tr>
<tr>
<td>BirdSong</td>
<td>4,998</td>
<td>38</td>
<td>13</td>
<td>2.18</td>
<td>bird song classification</td>
</tr>
<tr>
<td>LYN10</td>
<td>16,526</td>
<td>163</td>
<td>10</td>
<td>1.84</td>
<td>automatic face naming</td>
</tr>
<tr>
<td>LYN20</td>
<td>17,511</td>
<td>163</td>
<td>20</td>
<td>1.85</td>
<td>automatic face naming</td>
</tr>
</tbody>
</table>

Figure 1: Classification accuracy of each comparing approach changes as the proportion of PL training examples \(\rho\) increases from 0.05 to 0.7 (Data set: Satimage; \(r = 1, 2, 3\)).

Here, \(\text{vec}(\cdot)\) is the vectorization operator and \(C = M - N\).

As the alternating optimization procedure for \(w\) and \(F_U\) terminates, the predictive model \(w = [w^1; w^2; \ldots; w^q]\) can be obtained based on the KKT condition:

\[
 w_r = \sum_{i=1}^{p+u} \left( \sum_{l=1}^{q} \alpha_{rl}^i - \sum_{l=1}^{q} \alpha_{lr}^i \right) x_i \quad (1 \leq r \leq q) \tag{13}
\]

Accordingly, given the unseen instance \(x^*_s\), it is natural for PARM to predict its class label \(y_s\) as:

\[
 y_s = \arg \max_{y \in Y} \ w^T \Phi(x_s, y) \tag{14}
\]

In summary, Table 1 gives the pseudo-code of PARM. Firstly, the labeling confidence matrix \(F_P\) over PL training examples is estimated (Step 1). After that, an alternating optimization procedure is invoked to update predictive model \(w\) and the labeling confidence matrix \(F_U\) over unlabeled data (Steps 2-7). Finally, the class label for unseen instance is predicted based on the learned classification model (Step 8).

4 Experiments

4.1 Experimental Setup

The performance of PARM is compared against five state-of-the-art partial label learning algorithms, each configured with parameters suggested in respective literatures: 1) SSPL [27]: The only available semi-supervised partial label learning approach which learns from PL training examples and unlabeled data via graph-based label propagation [suggested configuration: \(k = 10, \alpha = 0.7, \beta = 0.25, r = \)]
Table 3: Classification accuracy (mean±std) of each comparing approach on real-world partial label data sets with $\rho \in \{0.05, 0.1, 0.15, 0.3, 0.5, 0.7\}$. In addition, */° indicates whether PARM is statistically superior/inferior on each data set (pairwise t-test at 0.05 significance level).

<table>
<thead>
<tr>
<th>Comparing approach</th>
<th>P = 0.05</th>
<th>P = 0.1</th>
<th>P = 0.15</th>
<th>P = 0.3</th>
<th>P = 0.5</th>
<th>P = 0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-AGGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparing approach</th>
<th>Mirfi&lt;sup&gt;ck&lt;/sup&gt;</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PARM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-AGGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparing approach</th>
<th>BirdSong</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PARM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-AGGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparing approach</th>
<th>LYN10</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PARM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-AGGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparing approach</th>
<th>LYN20</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PARM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL-AGGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

0.7, $T = 100$; 2) PL-KNN: An instance-based partial label learning approach which works by kNN weighted voting [suggested configuration: $k = 10$]; 3) CLPL: A convex partial label learning approach which works by averaging-based disambiguation [suggested configuration: SVM with squared hinge loss]; 4) PL-SVM: A maximum margin partial label learning approach which works by identification-based disambiguation [suggested configuration: regularization parameter pool with $\{10^{-3}, \cdots, 10^3\}$]; 5) PL-AGGD: A transformation-based partial label learning approach which works by manifold regularization [suggested configuration: $k = 10$, $\lambda = 1$, $\mu = 1$, $\gamma = 0.05$].

Table summarizes characteristics of the experimental data sets used in this paper. Following the widely used experimental protocol in partial label learning, synthetic PL data sets are generated from multi-class UCI data sets with controlling parameter $r$. Here, for any multi-class example $(x_i, y_i)$, one synthetic PL example $(x_i, S_i)$ is generated by randomly adding $r$ labels $\Delta_r \subseteq \mathcal{Y} \setminus \{y_i\}$ into $S_i$, i.e., $S_i = \Delta_r \cup \{y_i\}$. Furthermore, five real-world PL data sets from different task domains have also been employed for experimental studies, including Lost [8], LYN10, LYN20 [12] for automatic face naming, Mirfi<sup>ck</sup> [13] for web image classification, and BirdSong [4] for bird song classification.

For vehicle, the setting $r = 3$ is considered as there are only four class labels in the label space.
Table 4: Win/tie/loss counts (pairwise t-test at 0.05 significance level) between PARM and each comparing approach on synthetic as well as real-world partial label data sets. [Controlled UCI data sets: 36 cases (4 data sets × 9 configurations of ρ) for r = 1, 2; 27 cases (3 data sets × 9 configurations of ρ) for r = 3. Real-world data sets: 45 cases (5 data sets × 9 configurations of ρ)]

<table>
<thead>
<tr>
<th></th>
<th>SSPL</th>
<th>PL-KNN</th>
<th>CLPL</th>
<th>PL-SVM</th>
<th>PL-AGGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled UCI data sets (r = 1)</td>
<td>14/21/1</td>
<td>26/9/1</td>
<td>11/16/9</td>
<td>20/14/2</td>
<td>10/19/7</td>
</tr>
<tr>
<td>Controlled UCI data sets (r = 2)</td>
<td>17/15/4</td>
<td>23/13/0</td>
<td>10/17/9</td>
<td>19/14/3</td>
<td>9/22/5</td>
</tr>
<tr>
<td>Controlled UCI data sets (r = 3)</td>
<td>18/7/2</td>
<td>23/4/0</td>
<td>12/14/1</td>
<td>18/7/2</td>
<td>9/16/2</td>
</tr>
<tr>
<td>Real-world data sets</td>
<td>20/21/4</td>
<td>45/0/0</td>
<td>31/14/0</td>
<td>34/11/0</td>
<td>24/16/5</td>
</tr>
<tr>
<td>In Total</td>
<td>69/64/11</td>
<td>117/26/1</td>
<td>64/61/19</td>
<td>91/46/7</td>
<td>52/73/19</td>
</tr>
</tbody>
</table>

On each data set, ten-fold cross validation is performed whose mean accuracy as well as standard deviation are recorded for all comparing approaches. Given the training set $D_{train}$ and test set $D_{test}$, a proportion $\rho \in (0, 1)$ of training examples in $D_{train}$ are sampled to form $D_P$ and the rest training examples are used to form $D_U$ by discarding their candidate labeling information. For thorough performance evaluation, we consider varying proportions of PL training examples in this paper with $\rho \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. For semi-supervised comparing approaches PARM and SSPL which learn from PL training examples and unlabeled data, the classification model is trained in $D_P \cup D_U$ and evaluated on $D_{test}$. For the other four comparing approaches PL-KNN, CLPL, PL-SVM and PL-AGGD which learn from PL training examples, the classification model is trained in $D_P$ and evaluated on $D_{test}$.

As shown in Table 4, the regularization parameters $\lambda$ and $\mu$ for PARM are chosen among \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\} via cross-validation on training set and $\gamma = 0.01$.

### 4.2 Experimental Results

Due to page limit, Figure 1 and Table 3 report the experimental results on synthetic as well as real-world partial label data sets under certain experimental configurations. Specifically, Figure 1 illustrates how the classification accuracy of each comparing approach changes as $\rho$ (proportion of PL training examples) increases on the synthetic data set $SatImage$ (with $r = 1, 2, 3$). In addition, Table 3 gives the classification accuracy of each comparing approach on the real-world partial label data sets (with $\rho \in \{0.05, 0.1, 0.15, 0.3, 0.5, 0.7\}$). Furthermore, Table 4 summarizes the win/tie/loss counts between PARM and each comparing approach (pairwise t-test at 0.05 significance level) across all experimental configurations.

Table 4 reports the win/tie/loss counts between PARM and each comparing algorithm based on pairwise t-test at 0.05 significance level. As shown in the reported results, we can observe that: a) On synthetic data sets, PARM achieves superior or at least comparable performance to SSPL, PL-KNN, CLPL, PL-SVM and PL-AGGD in 92.9%, 99.0%, 80.8%, 92.9% and 85.9% cases respectively; b) On real-world data sets, compared to the semi-supervised partial label learning approach SSPL, PARM achieves superior performance in 44.4% cases and inferior performance in only 8.9% cases; c) On real-world data sets, compared to partial label learning approaches under supervised setting, PARM significantly outperforms PL-KNN in all cases. Furthermore, PARM significantly outperforms CLPL, PL-SVM and PL-AGGD in 68.9%, 75.6% and 53.3% cases respectively, and has been outperformed by CLPL and PL-SVM in none cases; d) As shown in Figure 1 the performance advantage of PARM over comparing approaches is more pronounced under the challenging cases where $\rho$ (i.e. proportion of PL training examples) is small.

Figure 2 gives the parameter sensitivity analysis for PARM on BirdSong data set ($\rho = 0.5$). As shown in Figure 2(a), the performance of PARM is somewhat sensitive w.r.t. $\lambda$ and $\mu$, whose values are chosen via cross-validation on the training set in this paper. As shown in Figure 2(b)-(c), the performance of PARM is relatively stable w.r.t. $\gamma$, whose value is fixed to be 0.01 in this paper.

Figure 3 illustrates how the classification model (i.e. $\|w^{(t)} - w^{(t-1)}\|_2$) and the confidence matrix over unlabeled examples (i.e. $\|F_U^{(t)} - F_U^{(t-1)}\|_F$) converge as the number of optimization iterations $t$
increases. We can see that the classification model and labeling confidence of unlabeled data converge fast with increasing number of iterations.

5 Conclusion

In this paper, the problem of semi-supervised partial label learning is investigated. To learn from both PL training examples and unlabeled data, we introduce confidence-rated margin maximization to jointly optimize predictive model and estimate latent labeling confidence. Comprehensive experiments show that the proposed approach performs favorably against state-of-the-art approaches.

In the future, it would be interesting to investigate ways of enabling the proposed approach to deal with large-scale data sets. Furthermore, other than adopting label propagation to instantiate the labeling confidence of PL examples, it is desirable to explore alternative ways of exploiting the supervision information of PL examples to facilitate model training.

Broader Impact

In this paper, we study the problem of semi-supervised partial label learning which has been less investigated in weakly supervised learning. The developed techniques can be applied to scenarios where the supervision information collected from the environment is accurate. For ethical use of the proposed approach, one should expect proper acquisition of the candidate labeling information (e.g. crowdsourcing) as well as the unlabeled data. We believe that developing such techniques is important to meet the increasing needs of learning from weak supervision in many real-world applications.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key R&D Program of China (2018YFB1004300), the National
Science Foundation of China (61573104), the China University S&T Innovation Plan Guided by the Ministry of Education, and partially supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization. We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References


[23] Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, and Jiawei Han. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1378, Austin, TX, 2016.

[24] Xiang Ren, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, and Jiawei Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1825–1834, San Francisco, CA, 2016.


