Towards Enabling Binary Decomposition for Partial Label Learning∗

Xuan Wu1,2, Min-Ling Zhang1,2,3

1School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
2Key Laboratory of Computer Network and Information Integration (Southeast University), MOE, China
3Collaborative Innovation Center for Wireless Communications Technology, China

{wuxuan, zhangml}@seu.edu.cn

Abstract

The task of partial label (PL) learning is to learn a multi-class classifier from training examples each associated with a set of candidate labels, among which only one corresponds to the ground-truth label. It is well known that for inducing multi-class predictive model, the most straightforward solution is binary decomposition which works by either one-vs-rest or one-vs-one strategy. Nonetheless, the ground-truth label for each PL training example is concealed in its candidate label set and thus not accessible to the learning algorithm, binary decomposition cannot be directly applied under partial label learning scenario. In this paper, a novel approach is proposed to solving partial label learning problem by adapting the popular one-vs-one decomposition strategy. Specifically, one binary classifier is derived for each pair of class labels, where PL training examples which have distinct relevancy to the label pair are used to generate the corresponding binary training set. Furthermore, an auxiliary set

1 Introduction

In partial label (PL) learning, each training example is represented by a single instance while associated with a set of candidate labels, among which only one label is valid [Cour et al., 2011; Chen et al., 2014; Yu and Zhang, 2017]. Partial label learning has manifested its capability in solving real-world problems where only weakly-supervised information can be acquired, such as web mining [Jie and Orabona, 2010], multimedia contents analysis [Zeng et al., 2013; Chen et al., in press], ecoinformatics [Liu and Dietterich, 2012; Tang and Zhang, 2017], etc.

Formally speaking, let \( X = \mathbb{R}^d \) be the d-dimensional instance space and \( Y = \{\lambda_1, \lambda_2, \ldots, \lambda_q\} \) be the label space with q class labels. Given a set of PL training examples \( D = \{(x_i, S_i)|1 \leq i \leq m\} \), partial label learning aims to induce a multi-class classification model \( f : X \rightarrow Y \) from \( D \). Here, \( x_i \in X \) is a d-dimensional feature vector \( (x_{i1}, x_{i2}, \ldots, x_{id})^T \) and \( S_i \subseteq Y \) is the associated candidate label set. Following the key assumption of partial label learning, the ground-truth label \( y_i \) for \( x_i \) is concealed in its candidate label set (i.e. \( y_i \in S_i \)) and therefore cannot be accessed by the learning algorithm.

To accomplish the task of learning from partial label data, a number of approaches have been proposed by fitting widely-used learning techniques to partial label data, such as k-nearest neighbor [Hullermeier and Beringer, 2006; Zhang and Yu, 2015; Gong et al., 2018], maximum margin [Nguyen and Caruana, 2008; Yu and Zhang, 2017], maximum likelihood [Jin and Ghahramani, 2003; Liu and Dietterich, 2012], boosting [Tang and Zhang, 2017], etc. However, other than fitting existing learning techniques to data, a natural postulation is that whether partial label learning problem can be solved by fitting data to existing learning techniques. Specifically, considering that the ultimate goal of partial label learning is to induce a multi-class classifier, binary decomposition should serve as the most straightforward solution for fulfilling multi-class classification. Unfortunately, due to the fact that ground-truth label is not accessible from the PL training example, neither the one-vs-rest nor the one-vs-one binary decomposition strategy can be directly employed under partial label learning scenario.

In this paper, we aim to enable binary decomposition for partial label learning by adapting the popular one-vs-one decomposition strategy. Accordingly, a novel partial label learning algorithm named PALOC, i.e. Partial Label learning via One-vs-one deComposition, is proposed. During the training phase, one binary classifier is derived for each pair of class labels, where PL training examples which have distinct relevancy to the label pair are utilized to instantiate the corresponding binary training set. Furthermore, an auxiliary set
of binary classifiers (one per class label) are derived by stacking over the predictive outputs of existing binary classifiers to help improve generalization. During the test phase, classification results yielded by all binary classifiers are synergized to make prediction on unseen instance. Extensive experiments on artificial as well as real-world PL data sets clearly show that PALOC achieves highly competitive performance against state-of-the-art partial label learning approaches.

The rest of this paper is organized as follows. Section 2 presents technical details of the proposed PALOC approach. Section 3 discusses existing works related to PALOC. Section 4 reports detailed results of comparative experiments. Finally, Section 5 concludes.

2 The PALOC Approach

Following the notations given in Section 1, partial label learning aims to induce a multi-class classifier \( f : \mathcal{X} \mapsto \mathcal{Y} \) from \( \mathcal{D} \) which maps from the instance space to the label space. Generally, binary decomposition serves as an intuitive solution which transforms multi-class learning problem into a number of binary learning problems. The main difficulty of applying binary decomposition techniques to partial label learning lies in that the ground-truth label \( y_i \) of each PL training example \((x_i, S_i)\) is concealed in its candidate label set \( S_i \). Thereafter, for the one-vs-rest binary decomposition strategy, it is unclear whether \((x_i, S_i)\) should be regarded as a positive or negative example w.r.t. one specific class label. For the one-vs-one binary decomposition strategy, it is unclear which class label should \((x_i, S_i)\) belong to w.r.t. a pair of class labels.

PALOC enables binary decomposition for partial label learning by fitting PL data to the one-vs-one (OVO) decomposition strategy. Specifically, for each pair of class labels \((\lambda_j, \lambda_k)\), the relevancy of \((x_i, S_i)\) to \(\lambda_j\) and \(\lambda_k\) is determined via their assignment to the candidate label set instead of their equivalence to the (unknown) ground-truth label. Let \(\mathcal{S}_i = \mathcal{Y}\setminus S_i\) denote the complementary label set of \(S_i\) in \(\mathcal{Y}\), a binary training set is constructed w.r.t. each pair of class labels \((\lambda_j, \lambda_k)\) as follows:

\[
D_{jk} = \{ (x_i, \psi(S_i, \lambda_j, \lambda_k)) | \phi(S_i, \lambda_j) \neq \phi(S_i, \lambda_k), 1 \leq i \leq m \},
\]

where \(\phi(S_i, \lambda) = \begin{cases} +1, & \text{if } \lambda \in S_i \\ -1, & \text{otherwise} \end{cases}\), and \(\psi(S_i, \lambda_j, \lambda_k) = \begin{cases} +1, & \text{if } \lambda_j \in S_i \text{ and } \lambda_k \in \mathcal{S}_i \\ -1, & \text{if } \lambda_j \in \mathcal{S}_i \text{ and } \lambda_k \in S_i. \end{cases}\)

In other words, only PL training example \((x_i, S_i)\) where \(\lambda_j\) and \(\lambda_k\) have distinct assignment to \(S_i\) is utilized to instantiate the binary training set \(D_{jk}\). Accordingly, a binary classifier \(g_{jk} : \mathcal{X} \mapsto \mathbb{R}\) is derived by invoking some binary training algorithm \(\mathcal{B}\) on \(D_{jk}\), i.e. \(g_{jk} \leftarrow \mathcal{B}(D_{jk})\). Without loss of generality, a total of \(\binom{\mathcal{Y}}{2}\) can be derived from the above one-vs-one binary decomposition procedure (by assuming \(j < k\)). Furthermore, each PL training example \((x_i, S_i)\) will contribute to the learning procedure of \(|S_i||\mathcal{S}_i|\) binary classifiers.

Given the set of \(\binom{\mathcal{Y}}{2}\) binary classifiers, the class label for any instance \(x\) can be predicted by counting the votes yielded by all classifiers:

\[
y = \arg\max_{\lambda_j \in \mathcal{Y}} \ V_{\text{ovo}}(x, \lambda_j) = \arg\max_{\lambda_j \in \mathcal{Y}} \sum_{j=1}^{m-1} \mathbb{I}(g_{jk}(x) < 0) + \sum_{k=j+1}^{m} \mathbb{I}(g_{jk}(x) > 0)
\]

Although it is feasible to make final prediction based on Eq.(2), PALOC employs stacking strategy [Zhou, 2012] to further improve generalization performance. For each PL training example \((x_i, S_i)\), its candidate label set \(S_i\) can be refined to \(\hat{S}_i\) as follows:

\[
\hat{S}_i = \{ \hat{y}_i \}, \text{ if } \hat{y}_i \in S_i \\
S_i, \text{ if } \hat{y}_i \notin S_i
\]

Here, \(\hat{y}_i = \arg\max_{\lambda_j \in \mathcal{Y}} V_{\text{ovo}}(x_i, \lambda_j)\) corresponds to the disambiguation prediction for \(x_i\) based on Eq.(2). Therefore, the candidate label set is refined to be \(\hat{S}_i\) if the disambiguation prediction falls into \(S_i\). Otherwise, the candidate label set remains unchanged.

Given the set of \(\binom{\mathcal{Y}}{2}\) binary classifiers \(g_{jk} (1 \leq j < k \leq q)\), an auxiliary set of \(q\) binary classifiers (one per class label) are further derived via stacked generalization. For each class label \(\lambda_j (1 \leq j \leq q)\), a corresponding binary training set is constructed as follows:

\[
D_j = \{ (\hat{x}_i, \varphi(\hat{S}_i, \lambda_j)) | 1 \leq i \leq m \},
\]

where \(\hat{x}_i = [g_{12}(x_i), g_{13}(x_i), \ldots, g_{(q-1)q}(x_i)]\),

and \(\varphi(\hat{S}_i, \lambda_j) = \begin{cases} +1, & \text{if } \lambda_j \in \hat{S}_i \\ -1, & \text{if } \lambda_j \notin \hat{S}_i. \end{cases}\)

In other words, each PL training example in \(D_j\), \(\hat{x}_i\) is formed by augmenting the original feature vector \(x_i\) with the predictive outputs of all \(\binom{\mathcal{Y}}{2}\) classifiers. Furthermore, the binary label \(\varphi(\hat{S}_i, \lambda_j)\) is determined by the assignment of \(\lambda_j\) w.r.t. the refined candidate label set \(\hat{S}_i\). Accordingly, a binary classifier \(g_j : \mathcal{X} \mapsto \mathbb{R}\) is further derived by invoking some binary training algorithm \(\mathcal{B}\) on \(D_j\), i.e. \(g_j \leftarrow \mathcal{B}(D_j)\).

During the testing phase, the unseen instance \(x^*\) is firstly fed to the \(\binom{\mathcal{Y}}{2}\) classifiers to generate the augmented feature vector:

\[
\hat{x}^* = [x^*, g_{12}(x^*), g_{13}(x^*), \ldots, g_{(q-1)q}(x^*)].
\]

After that, the set of \(\binom{\mathcal{Y}}{2}\) classifiers \(g_{jk} (1 \leq j < k \leq q)\) and the other set of \(q\) classifiers \(g_j (1 \leq j \leq q)\) are synergized to make prediction on \(x^*\):

\[
y^* = f(x^*) = \arg\max_{\lambda_j \in \mathcal{Y}} \ V_{\text{ovo}}(x^*, \lambda_j) + \mu \cdot V_{\text{stack}}(\hat{x}^*, \lambda_j)
\]

\[
= \arg\max_{\lambda_j \in \mathcal{Y}} \sum_{h=1}^{j-1} \mathbb{I}(g_{hk}(x^*) < 0) + \sum_{k=j+1}^{q} \mathbb{I}(g_{jk}(x^*) > 0) + \mu \cdot \mathbb{I}(g_j(\hat{x}^*) > 0)
\]

Here, \(\mu\) is the balance parameter which controls the relative importance of \(V_{\text{ovo}}(x^*, \lambda_j)\), i.e. the votes yielded by one-vs-one binary classifiers, and \(V_{\text{stack}}(x^*, \lambda_j)\), i.e. the votes yielded by stacking binary classifiers.
Semi-supervised learning [Zhu and Goldberg, 2009] aims to induce a classifier \( f : \mathcal{X} \rightarrow \mathcal{Y} \) from both labeled and unlabeled examples, where the ground-truth label assumes the whole label space for unlabeled example while assumes the candidate label set for PL example. Multi-instance learning [Amores, 2013] aims to induce a classifier \( f : 2^\mathcal{X} \rightarrow \mathcal{Y} \) from examples each represented as a labeled bag of instances, where a single label is assigned to a set of instances for multi-instance example while a set of labels are assigned to a single instance for PL example. Multi-label learning [Zhang and Zhou, 2014] aims to induce a classifier \( f : \mathcal{X} \rightarrow 2^\mathcal{Y} \) from training examples each associated with multiple labels, where the associated labels are all valid ones for multi-label example while the associated labels are only candidate ones for PL example.

Most existing algorithms learn from PL examples by fitting widely-used learning techniques to partial label data. For maximum likelihood techniques, the likelihood of observing each PL training example is defined over its candidate label set instead of the unknown ground-truth label [Jin and Ghahramani, 2003; Liu and Dietterich, 2012]. For \( k \)-nearest neighbor techniques, the candidate label sets of neighboring instances are merged via weighted voting for making prediction [Hüllermeier and Beringer, 2006; Zhang and Yu, 2015]. For maximum margin techniques, the classification margin over each PL training example is defined by discriminating modeling outputs from candidate labels and non-candidate labels [Nguyen and Caruana, 2008; Yu and Zhang, 2017]. For boosting techniques, the weight over each PL training example and the confidence of each candidate label being the ground-truth label are updated in each boosting round [Tang and Zhang, 2017].

Other than fitting existing learning techniques to PL data, there are few works which work by fitting PL data to existing learning techniques. The CLPL approach [Cour et al., 2011] maps a \( d \)-dimensional instance in \( \mathcal{X} \) into a \( d \times q \)-dimensional feature vector for each class label in \( \mathcal{Y} \). For each PL training example \( (x_i, S_i) \), one positive example is generated by averaging mapped feature vectors w.r.t. candidate labels in \( S_i \) and \( q - |S_i| \) negative examples are generated by taking the mapped feature vector w.r.t. each non-candidate label in \( \mathcal{Y} \setminus S_i \). The PL-ECOC approach [Zhang et al., 2017] transforms each instance into a binary example based on the utilization of ECOC coding matrix [Dietterich and Bakiri, 1995; Zhou, 2012]. For each PL training example \( (x_i, S_i) \), it is regarded as a positive or negative example if its candidate label set \( S_i \) entirely falls into the column dichotomy of the coding matrix. Although both CLPL and PL-ECOC work by transforming the partial label learning problem into binary learning problem, PALOC enables binary decomposition for PL data in a more concise manner without relying on extra manipulations such as feature mapping or coding matrix.

### 4 Experiments

#### 4.1 Comparing Algorithms

To evaluation the performance of PALOC, five state-of-the-art partial label learning approaches with suggested parameter configurations have been employed for comparative studies:
● CLPL [Cour et al., 2011] which transforms partial label learning problem to binary learning problem via feature mapping with convex loss optimization [suggested configuration: SVM with squared hinge loss];

● PL-KNN [Hüllermeier and Beringer, 2006] which adapts $k$-nearest neighbor technique to learn from PL data via weighted voting [suggested configuration: $k = 10$];

● PL-SVM [Nguyen and Caruana, 2008] which adapts maximum margin technique to learn from PL data via $l_2$ regularization [suggested configuration: regularization parameter pool with $\{10^{-5}, \ldots, 10^3\}$];

● LSB-CMM [Liu and Dietterich, 2012] which adapts maximum likelihood to learn from PL data via mixture models [suggested configuration: 5$q$ mixture components];

● PL-ECOC [Zhang et al., 2017] which transforms partial label learning problem to binary learning problem via E-COC coding matrix [suggested configuration: codeword length $L = \lfloor 10\log_2(q) \rfloor$].

As shown in Table 1, the only parameter to be set for PALOC is $\mu$, which controls relative importance of the generated one-vs-one classifiers and stacking classifiers. In the rest of this paper, $\mu$ is fixed to be 10 for performance evaluation. Furthermore, similar to CLPL and PL-ECOC, SVM [Chang and Lin, 2011] is utilized to instantiate the binary base learner $B$ for PALOC.

Next, two series of experiments are conducted on controlled UCI data sets as well as real-world partial label data sets. For each data set, ten-fold cross-validation is performed where the mean predictive accuracies and standard deviations are recorded for all comparing approaches.

### 4.2 Controlled UCI Data Sets

Table 2 summarizes the characteristics of six controlled UCI data sets. Conventionally, following the widely-used controlling protocol, an artificial partial label data set is generated from one multi-class UCI data set under specified configuration of three controlling parameters $p$, $r$, and $\epsilon$ [Cour et al., 2011; Liu and Dietterich, 2012; Chen et al., 2014; Zhang et al., 2017]. Here, $p$ controls the proportion of examples which are partially labeled (i.e. $|S_i| > 1$), $r$ controls the number of false positive labels in the candidate label set (i.e. $|S_i| = r + 1$), and $\epsilon$ controls the co-occurring

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#Examples</th>
<th>#Features</th>
<th>#Class Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>glass</td>
<td>214</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>ecoli</td>
<td>336</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>vehicle</td>
<td>846</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>abalone</td>
<td>4,177</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>usps</td>
<td>9,298</td>
<td>256</td>
<td>10</td>
</tr>
<tr>
<td>letter</td>
<td>20,000</td>
<td>16</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2: Characteristics of the controlled UCI data sets.
Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

(a) glass

(b) ecoli

(c) vehicle

(d) abalone

(e) usps

(f) letter

Figure 2: Classification accuracy of each comparing algorithm changes as $\epsilon$ (co-occurring probability of the coupling label) increases from 0.1 to 0.7 (with 100% partially labeled examples $[p = 1]$ and one false positive candidate label $[r = 1]$).

<table>
<thead>
<tr>
<th>PALOC against</th>
<th>CLPL</th>
<th>PL-KNN</th>
<th>PL-SVM</th>
<th>LSB-CMM</th>
<th>PL-ECOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>varying $p$ $[r=1]$</td>
<td>35/7/0</td>
<td>28/14/0</td>
<td>34/8/0</td>
<td>25/16/1</td>
<td>7/30/5</td>
</tr>
<tr>
<td>varying $p$ $[r=2]$</td>
<td>30/12/0</td>
<td>29/13/0</td>
<td>32/10/0</td>
<td>22/20/0</td>
<td>7/30/5</td>
</tr>
<tr>
<td>varying $p$ $[r=3]$</td>
<td>35/7/0</td>
<td>31/11/0</td>
<td>37/5/0</td>
<td>21/21/0</td>
<td>9/29/4</td>
</tr>
<tr>
<td>varying $\epsilon$ $[p,r=1]$</td>
<td>22/15/5</td>
<td>30/9/3</td>
<td>29/13/0</td>
<td>22/18/2</td>
<td>10/26/6</td>
</tr>
<tr>
<td>In Total</td>
<td>122/41/5</td>
<td>118/47/3</td>
<td>132/36/0</td>
<td>90/75/3</td>
<td>33/115/20</td>
</tr>
</tbody>
</table>

Table 3: Win/tie/loss counts (pairwise $t$-test at 0.05 significance level) on the classification performance of PALOC against each comparing approach.

probability between one extra candidate label and the ground-truth label. As shown in Table 2, a total of 28 (4x7) parameter configurations are considered for each controlled UCI data set.

Figure 1 illustrates the classification accuracy of each comparing algorithm as $p$ increases from 0.1 to 0.7 with step-size 0.1 ($r = 1$). Along with the ground-truth label, one class label in $\mathcal{Y}$ will be randomly picked up to constitute the candidate label set. Due to page limit, figures for the cases of $r = 2$ and $r = 3$ are not illustrated here while similar results to Figure 1 can be observed as well. Figure 2 illustrates the classification accuracy of each comparing algorithm as $\epsilon$ increases from 0.1 to 0.7 with step-size 0.1 ($p = 1, r = 1$). Given any label $\lambda \in \mathcal{Y}$, one extra label $\lambda' \in \mathcal{Y}$ is designated as the coupling label which co-occurs with $\lambda$ in the candidate label set with probability $\epsilon$. Otherwise, any other class label would be randomly chosen to co-occur with $\lambda$.

As illustrated in Figures 1 and 2, the performance of PALOC is highly competitive to other comparing algorithms in most cases. Furthermore, pairwise $t$-test at 0.05 significance level is conducted based on the results of ten-fold cross-validation. Table 3 reports the win/tie/loss counts between PALOC and each comparing approach. Specifically, out of the 168 statistical tests (28 configurations x 6 UCI data sets), it is shown that: 1) PALOC achieves superior or at least comparable performance against CLPL and PL-SVM in most cases; 2) PALOC achieves superior performance against PL-KNN and LSB-CMM in 76.2% and 53.6% cases while has been outperformed by either of them in only 1.8% cases; 3) PALOC achieves superior performance against CLPL and PL-ECOC in 72.6% and 19.6% cases while has been outperformed by them in only 3.0% and 11.9% cases respectively.
4.3 Real-World Data Sets

Table 4 summarizes the characteristics of real-world partial label data sets, which are collected from several application domains including FG-NET [Panis and Lanitis, 2015] for facial age estimation, Lost [Cour et al., 2011], Soccer Player [Zeng et al., 2013] and Yahoo! News [Guillaumin et al., 2010] for automatic face naming from images or videos, MSRCv2 [Liu and Dietterich, 2012] for object classification, and BirdSong [Briggs et al., 2012] for bird song classification. The average number of candidate labels (avg. #CLs) for each real-world partial label data set is also recorded in Table 4.

Table 5 reports the mean classification accuracy as well as standard deviation of each comparing algorithm on the real-world partial label data sets. In addition, •/○ indicates whether PALOC is statistically superior/inferior to the comparing algorithm on each data set (pairwise t-test at 0.05 significance level).

4.4 Sensitivity Analysis

As shown in Eq.(6), PALOC employs parameter $\mu$ to balance the voting predictions from one-vs-one classifiers and stacking classifiers respectively. To investigate the performance sensitivity of PALOC w.r.t. $\mu$, Figure 3 shows how the classification accuracy of PALOC changes as $\mu$ varies from 0 to 10 with step-size 1. Here, three real-world PL data sets Lost, MSRCv2 and BirdSong are employed for illustrative purpose. As shown in Figure 3, it is obvious that the performance of PALOC improves as $\mu$ increases from 0 and becomes relatively stable as $\mu$ reaches 4. Note that $\mu = 0$ corresponds to the case where only one-vs-one classifiers contribute to the final prediction, and these observations indicate that the stacking classifiers do help improve the performance of PALOC.

5 Conclusion

In this paper, the problem of partial label learning is studied where a novel approach based on binary decomposition is proposed. Specifically, one-vs-one decomposition strategy is enabled to deal with partial label learning problem by considering the relevancy of each label pair w.r.t. the candidate label set of PL training examples. Effectiveness of the pro-
posed approach is validated via comprehensive experiments on both controlled UCI data sets and real-world PL data sets. As shown in Eq.(1), not all PL training examples will contribute to the construction of binary training set w.r.t. each label pair. Therefore, it is interesting to explore possible ways to make full use of the excluded PL training examples to further enhance the proposed binary decomposition approach.

References


