Confidence-Rated Discriminative Partial Label Learning

Cai-Zhi Tang,[1,2] **Min-Ling Zhang**[2,3,*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Computer Network and Information Integration
(Southeast University), Ministry of Education, China
[3]Collaborative Innovation Center of Wireless Communications Technology, China
220141515@seu.edu.cn, zhangml@seu.edu.cn* (corresponding author)

Abstract

Partial label learning aims to induce a multi-class classifier from training examples where each of them is associated with a set of *candidate* labels, among which only one label is valid. The common discriminative solution to learn from partial label examples assumes one parametric model for each class label, whose predictions are aggregated to optimize specific objectives such as likelihood or margin over the training examples. Nonetheless, existing discriminative approaches treat the predictions from all parametric models in an equal manner, where the confidence of each candidate label being the ground-truth label is not differentiated. In this paper, a boosting-style partial label learning approach is proposed to enabling confidence-rated discriminative modeling. Specifically, the ground-truth confidence of each candidate label is maintained in each boosting round and utilized to train the base classifier. Extensive experiments on artificial as well as real-world partial label data sets validate the effectiveness of the confidence-rated discriminative modeling.

Introduction

Partial label learning deals with the problem where each training example is associated with a set of candidate labels, among which only one label corresponds to the ground-truth one (Cour, Sapp, and Taskar 2011; Zhang 2014). Formally, let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional instance space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$ denote the label space consisting of q class labels. The task of partial label learning is to induce a multi-class classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ from the partial label training set $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) \mid 1 \leq i \leq m\}$. Here, $\boldsymbol{x}_i \in \mathcal{X}$ is a d-dimensional feature vector and $S_i \subseteq \mathcal{Y}$ is the set of candidate labels associated with $\boldsymbol{x}_i$. Particularly, the ground-truth label y_i for $\boldsymbol{x}_i$ is confined within S_i but not directly accessible to the learning algorithm.

The need of partial label learning arises in a number of real-world scenarios where only weak labeling information can be acquired during training data collection, such as automatic face naming (Cour et al. 2009; Zeng et al. 2013), web mining (Jie and Orabona 2010), ecoinformatics (Liu and Dietterich 2012), etc. In some literature, partial label learning is also termed as *ambiguous label learning* (Hüllermeier and Beringer 2006; Chen et al. 2014) or *superset label learning* (Liu and Dietterich 2014).

To learn from partial label examples, the common discriminative solution is to assume one parametric model $g(y_j \mid \boldsymbol{x}; \boldsymbol{\theta})$ for each class label y_j, whose modeling outputs are aggregated to optimize specific objectives such as likelihood or margin over the training examples (Jin and Ghahramani 2003; Nguyen and Caruana 2008; Cour, Sapp, and Taskar 2011; Liu and Dietterich 2012; Chen et al. 2014; Yu and Zhang 2016). Existing discriminative approaches conduct aggregation by treating the modeling outputs from all parametric models in an equal manner, where the confidence of each candidate label being the ground-truth label is not differentiated. This strategy might be suboptimal as each candidate label should contribute differently to the learning process, especially the contribution from the ground-truth label (i.e. y_i) against those from the false positive labels (i.e. $S_i \setminus \{y_i\}$) (Zhang, Zhou, and Liu 2016).

To overcome the potential drawback of existing strategy, a novel partial label learning approach named CORD, i.e. *COnfidence-Rated Discriminative partial label learning*, is proposed in this paper. CORD learns from partial label examples by adapting the popular boosting techniques, where the weights over training examples and the ground-truth confidences of candidate labels are maintained in each boosting round. Accordingly, the discriminative base classifier is trained by utilizing the currently-available weight and ground-truth confidence information. Empirical studies on a broad range of controlled UCI data sets and real-world partial label data sets clearly verify the effectiveness of the proposed confidence-rated discriminative learning approach.

We start the rest of this paper by briefly reviewing related work on partial label learning. Then, we present technical details of the proposed CORD approach and report experimental results of the comparative studies. Finally, we conclude the paper and indicate future research issues.

Related Work

In partial label learning, the labeling information conveyed by the training examples is weak as the ground-truth label is not accessible to the learning algorithm. It is worth noting that partial label learning is related to other well-studied weakly-supervised learning frameworks including *semi-supervised learning* (Zhu and Goldberg 2009), *multi-*

instance learning (Amores 2013) and *multi-label learning* (Zhang and Zhou 2014), while the weak supervision scenarios to be dealt with are different.

Semi-supervised learning aims to induce a classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ from a few labeled examples along with abundant unlabeled examples, where the ground-truth label assumes the whole label space for unlabeled example while the candidate label set for partial label example. Multi-instance learning aims to induce a classifier $f : 2^{\mathcal{X}} \mapsto \mathcal{Y}$ from training examples each represented by a bag of instances, where the label is assigned at the bag level for multi-instance example while at the instance level for partial label example. Multi-label learning aims to induce a classifier $f : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ from examples each associated with multiple labels, where the associated labels are all valid ones for multi-label example while only candidate ones for partial label example.

Discriminative modeling is the most common solution to learn from partial label examples, where one parametric model $g(y_j \mid \boldsymbol{x}; \boldsymbol{\theta})$ is assumed for each class label y_j $(1 \leq j \leq q)$. Correspondingly, model parameters are trained by optimizing specific objectives $J(\mathcal{D}; \boldsymbol{\theta})$ over the training examples. One popular instantiation of the objective function is to aggregate the modeling output of each parametric model via the maximum likelihood criterion (Jin and Ghahramani 2003; Liu and Dietterich 2012):

$$J(\mathcal{D}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \log \left(\sum_{j=1}^{q} \mathbb{I}(y_j \in S_i) \cdot g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right) \quad (1)$$

Here, $\mathbb{I}(\cdot)$ corresponds to the indicator function. It is obvious that maximizing $J(\mathcal{D}, \boldsymbol{\theta})$ is equivalent to maximizing the following objective function:

$$\tilde{J}(\mathcal{D}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \log \left(\sum_{y_j \in S_i} \frac{1}{|S_i|} \cdot g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right) \quad (2)$$

As shown in Eq.(2), modeling outputs of the parametric models contribute equally to the objective function, i.e. with uniform weight $\frac{1}{|S_i|}$ over each candidate label.

Another popular instantiation of the objective function is to aggregate the modeling output of each parametric model via the maximum margin criterion, such as (Cour, Sapp, and Taskar 2011; Zhang, Zhou, and Liu 2016):

$$J(\mathcal{D}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \left(\sum_{y_j \in S_i} \frac{1}{|S_i|} \cdot g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \sum_{y_k \in \hat{S}_i} \frac{1}{|\hat{S}_i|} \cdot g(y_k \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right) \quad (3)$$

or (Nguyen and Caruana 2008; Yu and Zhang 2016):

$$J(\mathcal{D}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \left(\max_{y_j \in S_i} \frac{1}{|S_i|} \cdot g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \max_{y_k \in \hat{S}_i} \frac{1}{|S_i|} \cdot g(y_k \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right) \quad (4)$$

Here, $\hat{S}_i$ corresponds to the complementary set of S_i in $\mathcal{Y}$. As shown in Eq.(3) and Eq.(4), modeling outputs of the parametric models also contribute equally to the objective function, i.e. with uniform weight $\frac{1}{|S_i|}$ over each candidate label.

In other words, for either maximum likelihood or maximum margin instantiation, the confidence of each candidate label being the ground-truth label is not differentiated. In the next section, a novel partial label learning approach will be introduced. Different from existing discriminative partial label learning approaches, the ground-truth confidence of each candidate label is estimated and utilized to facilitate the learning procedure.

The CORD Approach

Boosting is one of the widely-used machine learning techniques, which builds learning system with strong generalization ability by iteratively combining multiple weak learners. CORD learns from partial label examples by adapting the general boosting procedure, where in each boosting round the weights over training examples as well as ground-truth confidences of candidate labels are maintained simultaneously.

Given the partial label training set $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) \mid 1 \leq i \leq m\}$, in the t-th boosting round, let $\boldsymbol{w}^{(t)} = [w_1^{(t)}, w_2^{(t)}, \ldots, w_m^{(t)}]^\top$ be the weight vector over training examples, and $\mathbf{P}^{(t)} = [p_{ij}^{(t)}]_{m \times q}$ be the confidence matrix over candidate labels respectively. Specifically, $\boldsymbol{w}^{(t)}$ and $\mathbf{P}^{(t)}$ satisfy the non-negativity constraints: $w_i^{(t)} \geq 0$ and $p_{ij}^{(t)} \geq 0$, as well as the normalization constraints: $\sum_{i=1}^{m} w_i^{(t)} = 1$ and $\sum_{j=1}^{q} p_{ij}^{(t)} = 1$.

To train the base classifier $g(y \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ in the t-th boosting round, CORD chooses to maximize the following *confidence-rated* objective function:

$$J(\mathcal{D}, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{m} w_i^{(t)} \log \left(\sum_{y_j \in S_i} p_{ij}^{(t)} \cdot g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}) \right) \quad (5)$$

As shown in Eq.(5), the modeling output $g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})$ of each candidate label is weighted by $p_{ij}^{(t)}$, i.e. the confidence of y_j being the ground-truth label of $\boldsymbol{x}_i$. In this way, the ground-truth confidence of each candidate label is utilized to train the base classifier, reflecting the fact that different candidate labels should contribute differently to the learning process.

As per canonical boosting procedure, the empirical performance of the trained base classifier is evaluated as the *classification accuracy* over the (weighted) training examples. Nonetheless, for partial label learning, the performance of base classifier cannot be evaluated in this way as the ground-truth label of each training example is not directly accessible. In this paper, CORD makes use of the predictive difference between the maximum output of candidate and

Table 1: The pseudo-code of CORD.

Inputs:

$\mathcal{D}$: the partial label training set $\{(\boldsymbol{x}_i, S_i) \mid 1 \le i \le m\}$ $(\boldsymbol{x}_i \in \mathcal{X}, S_i \subseteq \mathcal{Y}, \mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{y_1, y_2, \ldots, y_q\})$

β: the confidence updating parameter

T: the maximum number of boosting rounds

$\boldsymbol{x}^*$: the unseen instance ($\boldsymbol{x}^* \in \mathcal{X}$)

Outputs:

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

Process:

1: Initialize the weight vector $\boldsymbol{w}^{(1)}$ as: $w_i^{(1)} = \frac{1}{m}$ ($\forall i \in \{1, \ldots, m\}$);

2: Initialize the confidence matrix $\mathbf{P}^{(1)}$ as: $p_{ij}^{(1)} = \frac{1}{|S_i|} \cdot \mathbb{I}(y_j \in S_i)$ ($\forall\, i \in \{1, \ldots, m\},\ j \in \{1, \ldots, q\}$);

3: **for** $t = 1$ **to** T **do**

4: Train the base classifier $g(y \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ by maximizing the confidence-rated objective function in Eq.(5);

5: Evaluate the performance of current base classifier $g(y \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ according to Eq.(6);

6: Set $\alpha^{(t)}$ according to Eq.(8);

7: Update $\boldsymbol{w}^{(t+1)}$ and $\mathbf{P}^{(t+1)}$ according to Eq.(7) and Eq.(9) respectively;

8: **end for**

9: **return** $y^* = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha^{(t)} \cdot g(y \mid \boldsymbol{x}^*; \boldsymbol{\theta}^{(t)})$.

non-candidate labels for performance evaluation (Nguyen and Caruana 2008; Yu and Zhang 2016):

$$r^{(t)} = \sum_{i=1}^{m} w_i^{(t)} \cdot \gamma_i^{(t)} \quad \text{where}$$
$$\gamma_i^{(t)} = \max_{y_j \in S_i} g(y_j \mid \boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}) - \max_{y_k \in \hat{S}_i} g(y_k \mid \boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}) \quad (6)$$

Accordingly, the weight vector $\boldsymbol{w}^{(t+1)}$ for the next boosting round is updated as:

$$\forall\, i \in \{1, \ldots, m\}: \quad w_i^{(t+1)} = \frac{w_i^{(t)} \cdot \exp\left(-\alpha^{(t)} \gamma_i^{(t)}\right)}{Z^{(t+1)}} \quad (7)$$

Here, $\alpha^{(t)}$ corresponds to the coefficient of the t-th boosting round to be used for classifier combination:[1]

$$\alpha^{(t)} = \frac{1}{2} \cdot \log\left(\frac{1 + r^{(t)}}{1 - r^{(t)}}\right) \quad (8)$$

and $Z^{(t+1)}$ corresponds to the normalization constant ensuring that $\sum_{i=1}^{m} w_i^{(t+1)} = 1$.

In addition, the confidence matrix $\mathbf{P}^{(t+1)}$ for the next boosting round is updated as:

$$\forall\ i \in \{1, \ldots, m\},\ j \in \{1, \ldots, q\}:$$
$$p_{ij}^{(t+1)} = \frac{p_{ij}^{(t)} \cdot \exp\left(\beta \cdot \mathbb{I}\left(y_j = y_i^{(t)}\right)\right)}{R_i^{(t+1)}} \quad \text{where}$$
$$y_i^{(t)} = \arg\max_{y \in S_i} g(y \mid \boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}) \quad (9)$$

Here, $\beta > 0$ is the confidence updating parameter and $y_i^{(t)}$ is the candidate label of $\boldsymbol{x}_i$ which has the largest modeling output at the t-th boosting round. Similarly, $R_i^{(t+1)}$ corresponds to the normalization constant ensuring that $\sum_{j=1}^{q} p_{ij}^{(t+1)} = 1$. In this way, the ground-truth confidence for the candidate label which coincides with $y_i^{(t)}$ will be increased.

Table 1 summarizes the boosting procedure of CORD.[2] Given the partial label training set, CORD initializes uniform weight over each training example and identical ground-truth confidence (i.e. $\frac{1}{|S_i|}$) for each candidate label of the training example (Steps 1-2). Then, in each boosting round the base classifier is trained w.r.t confidence-rated objective function (Step 4), the performance and coefficient for the base classifier are evaluated (Steps 5-6), and the weight vector and confidence matrix are updated accordingly (Step 7). Finally, the prediction on unseen instance is made by consulting the combined outputs of all base classifiers.

Experiments

Comparing Algorithms

In this paper, the effectiveness of CORD is evaluated against several state-of-the-art partial label learning algorithms, each configured with suggested parameters in the literature:

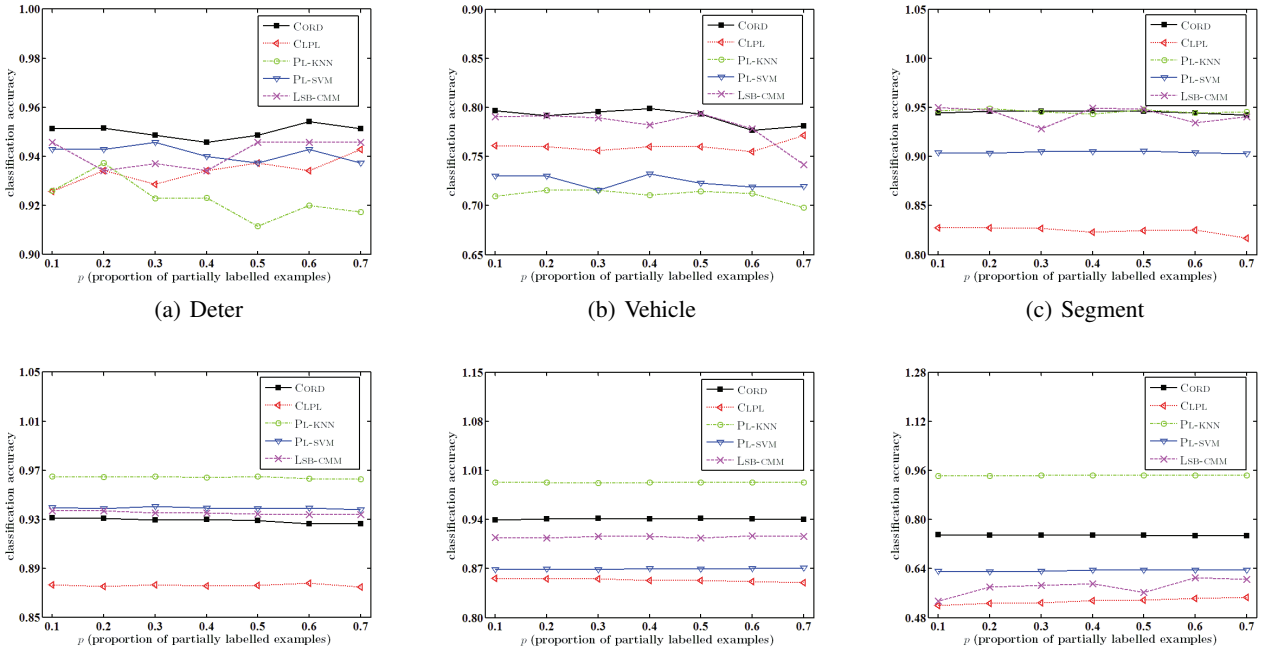
- CLPL (Cour, Sapp, and Taskar 2011): A convex optimization approach which learns from partial label examples by degenerating to binary classification problem [suggested configuration: SVM with squared hinge loss];
- PL-KNN (Hüllermeier and Beringer 2006): A k-nearest neighbor approach which learns from partial label examples by reasoning with the labeling information of neighboring examples [suggested configuration: $k = 10$];
- PL-SVM (Nguyen and Caruana 2008): A maximum-margin approach which learns from partial label examples by regularizing margin-based objective function [suggested configuration: regularization parameter pool with $\{10^{-3}, \ldots, 10^{3}\}$];
- LSB-CMM (Liu and Dietterich 2012): A maximum-likelihood approach which learns from partial label examples by maximizing mixture-based likelihood function [suggested configuration: q mixture components].

As shown in Table 1, the proposed CORD approach employs two parameters β and T for iterative training. In this paper, the confidence updating parameter β is set to be 0.5[3]

[1] Similar to canonical boosting procedure, the boosting rounds of CORD terminate if $\alpha^{(t)} < 0$.

[2] Code package for CORD is publicly-available at: http://cse.seu.edu.cn/PersonalPage/zhangml/Resources.htm#aaai17

[3] Preliminary experiments show that CORD performs stably with β taking values within [0.1, 1].



(a) Deter (b) Vehicle (c) Segment

(d) Usps (e) Pendigits (f) Letter

Figure 1: Classification accuracy of each comparing algorithm changes as p (proportion of partially labeled examples) increases (with one false positive candidate label $[r = 1]$).

Table 2: Characteristics of the controlled UCI data sets.

Data Set	#Examples	#Features	#Class Labels
Deter	358	23	6
Vehicle	846	18	4
Segment	2,310	18	7
Usps	9,298	256	10
Pendigits	10,992	16	10
Letter	20,000	16	26

Configurations
(I) $r = 1, p \in \{0.1, 0.2, \ldots, 0.7\}$
(II) $r = 2, p \in \{0.1, 0.2, \ldots, 0.7\}$
(III) $r = 3, p \in \{0.1, 0.2, \ldots, 0.7\}$
(IV) $p = 1, r = 1, \epsilon \in \{0.1, 0.2, \ldots, 0.7\}$

and the maximum boosting rounds T is set to be 10. Furthermore, maximum entropy model (Jin and Ghahramani 2003; Della Pietra, Della Pietra, and Lafferty 1997) is employed to serve as the base classifier which is trained with gradient-based optimization (Table 1, Step 4).

Two series of comparative studies are conducted among the comparing algorithms, with one series on controlled UCI data sets (Bache and Lichman 2013) and another series on real-world partial label data sets. Ten-fold cross-validation is performed on each data set, and the mean predictive accuracies (as well as the standard deviations) of all comparing algorithms are reported in the rest of this section.

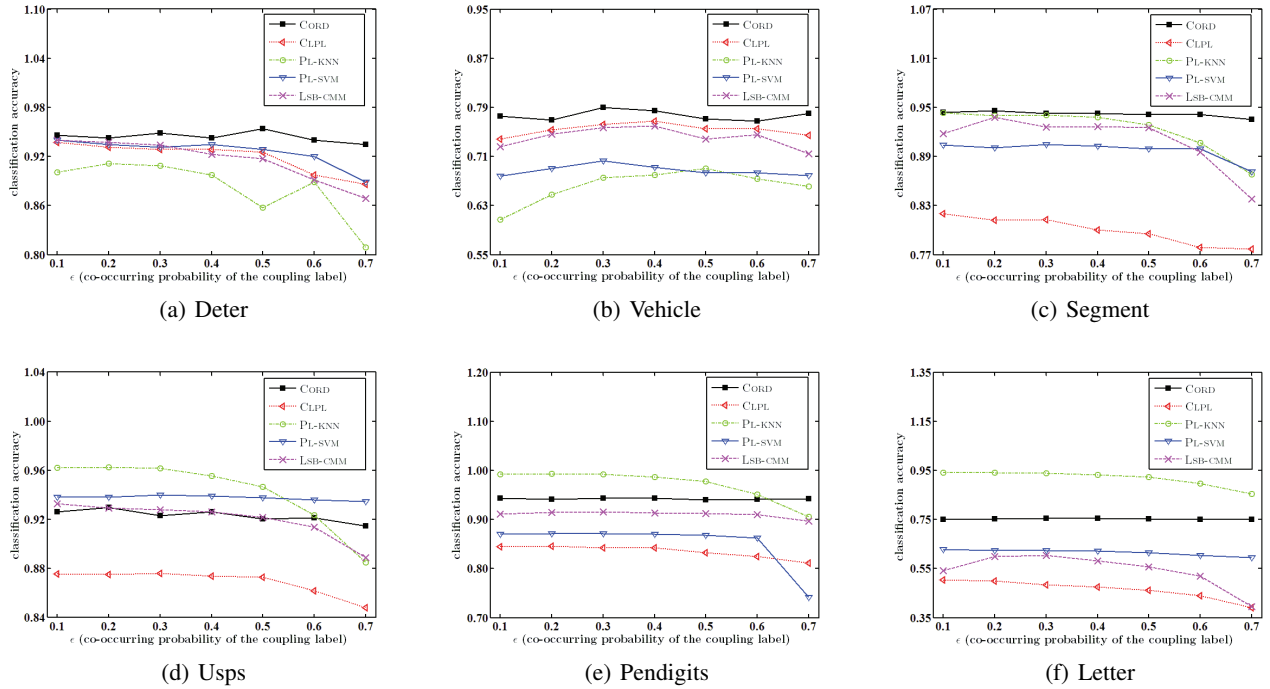
Table 3: Win/tie/loss counts (pairwise t-test at 0.05 significance level) on the classification performance of CORD against each comparing algorithm.

	CORD **against**			
	CLPL	PL-KNN	PL-SVM	LSB-CMM
varying p [r=1]	38/4/0	10/11/21	28/7/7	18/20/4
varying p [r=2]	32/10/0	12/9/21	28/7/7	18/21/3
varying p [r=3]	33/9/0	14/7/21	28/7/7	23/15/4
varying ϵ [p, r=1]	32/10/0	17/7/18	30/5/7	29/12/1
In Total	**135/33/0**	**53/34/81**	**114/26/28**	**88/68/12**

Controlled UCI Data Sets

Table 2 summarizes the characteristics of controlled UCI data sets. Specifically, 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 ϵ (Cour, Sapp, and Taskar 2011; Chen et al. 2014; Liu and Dietterich 2012; Zhang, Zhou, and Liu 2016). Here, p controls the proportion of examples being partially labeled (i.e. $|S_i| > 1$), r controls the number of false positive candidate labels (i.e. $|S_i| = r + 1$), and ϵ controls the co-occurring probability between the ground-truth label and one coupling candidate label. As shown in Table 2, a total of 28 (4x7) controlling parameter configurations are specified here.

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 chosen to constitute the



(a) Deter (b) Vehicle (c) Segment

(d) Usps (e) Pendigits (f) Letter

Figure 2: Classification accuracy of each comparing algorithm changes as ϵ (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 4: Characteristic of the real-world partial label data sets.

Data Set	#Examples	#Features	#Class Labels	avg. #CLs	Task Domain
Lost	1,122	108	16	2.23	*automatic face naming* (Cour, Sapp, and Taskar 2011)
MSRCv2	1,758	48	23	3.16	*object classification* (Liu and Dietterich 2012)
BirdSong	4,998	38	13	2.18	*bird song classification* (Briggs, Fern, and Raich 2012)
Soccer Player	17,472	279	171	2.09	*automatic face naming* (Zeng et al. 2013)
Yahoo! News	22,991	163	219	1.91	*automatic face naming* (Guillaumin, Verbeek, and Schmid 2010)

candidate label set of each partially labeled example. 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 ϵ increases from 0.1 to 0.7 with step-size 0.1 ($p = 1, r = 1$). Given the ground-truth label $y \in \mathcal{Y}$, another label $y' \in \mathcal{Y}$ designated as the coupling label will co-occur with y in the candidate label set with probability ϵ.

As shown in Figures 1 to 2, CORD performs favorably against the comparing algorithms in most cases. Furthermore, Table 3 reports the win/tie/loss counts between CORD and each comparing algorithm based on pairwise t-test at 0.05 significance level.

Out of the 168 statistical tests (28 configurations $\times$ 6 UCI data sets), it is shown that: 1) CORD achieves superior or at least comparable performance against CLPL in all cases; 2) CORD achieves superior performance against PL-KNN in 31.5% cases while has been outperformed by PL-KNN in 49.7% cases; 3) CORD achieves superior performance against PL-SVM and LSB-CMM in 67.8% and 52.3% cases and has been outperformed by them in only 16.7% and 7.1% cases. Generally, CORD is highly competitive to the comparing algorithms w.r.t. controlled UCI data sets, especially performs favorably against the discriminative partial label learning counterparts CLPL, PL-SVM and LSB-CMM.

Real-world Data Sets

Table 4 summarizes the characteristics of real-world partial label data sets, which have been collected from several task domains.[4] For `Lost` (Cour, Sapp, and Taskar 2011), `Soccer Player` (Zeng et al. 2013) and `Yahoo! News` (Guillaumin, Verbeek, and Schmid 2010) from the task of *automatic face naming*, faces cropped from an image or a video frame are represented as instances while names extracted from the associated image captions or video subtitles are regarded as candidate labels. For `MSRCv2` (Liu and

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

Table 5: Classification accuracy (mean±std) of each comparing algorithm on the real-world partial label data sets. In addition, •/◦ indicates whether CORD is statistically superior/inferior to the comparing algorithm on each data set (pairwise t-test at 0.05 significance level).

	Lost	MSRCv2	BirdSong	Soccer Player	Yahoo! News
CORD	0.806±0.026	0.474±0.040	0.712±0.008	0.457±0.013	0.624±0.010
CLPL	0.742±0.038•	0.413±0.039•	0.632±0.017•	0.368±0.010•	0.462±0.009•
PL-KNN	0.424±0.041•	0.448±0.037•	0.614±0.024•	0.497±0.014◦	0.457±0.010•
PL-SVM	0.729±0.040•	0.482±0.043	0.673±0.018•	0.443±0.014•	0.636±0.010
LSB-CMM	0.707±0.055•	0.456±0.031	0.717±0.024	0.525±0.015◦	0.648±0.007◦

Table 6: Transductive accuracy (mean±std) of each comparing algorithm on the real-world partial label data sets. In addition, •/◦ indicates whether CORD is statistically superior/inferior to the comparing algorithm on each data set (pairwise t-test at 0.05 significance level).

	Lost	MSRCv2	BirdSong	Soccer Player	Yahoo! News
CORD	0.925±0.006	0.667±0.007	0.843±0.002	0.764±0.002	0.873±0.001
CORD†	0.925±0.006	0.667±0.007	0.843±0.002	0.763±0.002	0.873±0.001
CLPL	0.894±0.005•	0.656±0.010	0.822±0.004•	0.680±0.010•	0.834±0.002•
PL-KNN	0.615±0.036•	0.616±0.006•	0.772±0.021•	0.492±0.015•	0.692±0.010•
PL-SVM	0.887±0.012•	0.653±0.024	0.825±0.012•	0.688±0.014•	0.871±0.002
LSB-CMM	0.721±0.010•	0.524±0.007•	0.716±0.014•	0.704±0.002•	0.872±0.001

Dietterich 2012) from the task of *object classification*, image segmentations are represented as instances while objects appearing within the image are regarded as candidate labels. For `BirdSong` (Briggs, Fern, and Raich 2012) from the task of *bird song classification*, singing syllables of the birds are represented as instances while bird species jointly singing during the same period are regarded as candidate labels. In addition, the average number of candidate labels (avg. #CLs) for each data set is also recorded in Table 4.

Table 5 reports the mean predictive accuracy as well as standard deviation of each comparing algorithm. Pairwise t-test at 0.05 significance level is conducted based on the ten-fold cross-validation, where the test outcomes between CORD and the comparing algorithms are also recorded.

As shown in Table 5, it is impressive to observe that: 1) On all data sets, CORD significantly outperforms CLPL and achieves superior or at least comparable performance to PL-SVM; 2) CORD achieves superior performance to PL-KNN on the `Lost`, `MSRCv2`, `BirdSong` and `Yahoo! News` data sets, and is only inferior to PL-KNN on the `Soccer Player` data set; 3) CORD is outperformed by LSB-CMM on the `Soccer Player` and `Yahoo! News` data sets, and achieves superior or comparable performance to LSB-CMM on the rest real-world partial label data sets.

In addition to the inductive performance reported in Table 5, it is also interesting to investigate the *transductive* performance of each comparing algorithm on classifying training examples (Cour, Sapp, and Taskar 2011; Zhang, Zhou, and Liu 2016). For each partial label training example $(\boldsymbol{x}_i, S_i)$, its ground-truth label is predicted by confining within the candidate label set, i.e. $y_i = \arg\max_{y \in S_i} g(y \mid \boldsymbol{x}_i; \boldsymbol{\theta})$. Conceptually, transductive performance of each comparing algorithm reflects its *disambiguation* ability in recovering the ground-truth label from candidate label set. Accordingly, Table 6 reports the transductive performance of each comparing algorithm along with the outcomes of pairwise t-tests at 0.05 significance level.

As shown in Table 6, on the `Lost`, `BirdSong` and `Soccer Player` data sets, CORD significantly outperforms all the comparing algorithms in terms of transductive accuracy. Furthermore, on the `MSRCv2` and `Yahoo! News` data sets, the performance of CORD is superior or at least comparable to all the comparing algorithms. As the boosting procedure of CORD terminates, the ground-truth label of each training example can also be predicted from the resulting confidence matrix $\mathbf{P}^{(T)}$, i.e. $y_i = \arg\max_{y_j \in S_i} p_{ij}^{(T)}$. For reference purpose, the corresponding transductive performance is also reported in Table 6 (denoted as CORD†). As shown in Table 6, CORD† and CORD perform almost identically across all data sets, which shows that the confidence matrix serves as a good indicator in disambiguating the ground-truth label.

Conclusion

In this paper, a new solution to partial label learning named CORD is proposed which employs the ground-truth confidence of each candidate label in discriminative modeling. Specifically, boosting techniques are adapted to learn from partial label examples which maintain the weights over training examples as well as the ground-truth confidences over candidate labels in each boosting round. Effectiveness of the proposed approach is clearly verified via extensive experiments on artificial and real-world partial label data sets.

One interesting future work is to explore other ways in instantiating confidence-rated discriminative partial la-

bel learning, such as trying alternative implementations of CORD (e.g. Step 5 in Table 1), adapting other discriminative learning techniques, etc. Furthermore, investigating whether confidence-rated modeling is helpful to improve non-discriminative partial label learning (Hüllermeier and Beringer 2006) is also worth further study.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (61222309, 61573104), the MOE Program for New Century Excellent Talents in University (NCET-13-0130), and partially supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

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

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

Briggs, F.; Fern, X. Z.; and Raich, R. 2012. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 534–542.

Chen, Y.-C.; Patel, V. M.; Chellappa, R.; and Phillips, P. J. 2014. Ambiguously labeled learning using dictionaries. *IEEE Transactios on Information Forensics and Security* 9(12):2076–2088.

Cour, T.; Sapp, B.; Jordan, C.; and Taskar, B. 2009. Learning from ambiguously labeled images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 919–926.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *Journal of Machine Learning Research* 12(May):1501–1536.

Della Pietra, S.; Della Pietra, V.; and Lafferty, J. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(4):380–393.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multiple instance metric learning from automatically labeled bags of faces. In Daniilidis, K.; Maragos, P.; and Paragios, N., eds., *Lecture Notes in Computer Science 6311*. Berlin: Springer. 634–647.

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

Jie, L., and Orabona, F. 2010. Learning from candidate labeling sets. In Lafferty, J.; Williams, C. K. I.; Shawe-Taylor, J.; Zemel, R. S.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 23*. Cambridge, MA: MIT Press. 1504–1512.

Jin, R., and Ghahramani, Z. 2003. Learning with multiple labels. In Becker, S.; Thrun, S.; and Obermayer, K., eds., *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press. 897–904.

Liu, L., and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. In Bartlett, P.; Pereira, F. C. N.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 25*. Cambridge, MA: MIT Press. 557–565.

Liu, L., and Dietterich, T. 2014. Learnability of the superset label learning problem. In *Proceedings of the 31st International Conference on Machine Learning*, 1629–1637.

Nguyen, N., and Caruana, R. 2008. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 381–389.

Yu, F., and Zhang, M.-L. 2016. Maximum margin partial label learning. *Machine Learning,* in press.

Zeng, Z.; Xiao, S.; Jia, K.; Chan, T.-H.; Gao, S.; Xu, D.; and Ma, Y. 2013. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 708–715.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhang, M.-L.; Zhou, B.-B.; and Liu, X.-Y. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1335–1344.

Zhang, M.-L. 2014. Disambiguation-free partial label learning. In *Proceedings of the 14th SIAM International Conference on Data Mining*, 37–45.

Zhu, X., and Goldberg, A. B. 2009. Introduction to semi-supervised learning. In Brachman, R. J., and Dietterich, T. G., eds., *Synthesis Lectures to Artificial Intelligence and Machine Learning*. San Francisco, CA: Morgan & Claypool Publishers. 1–130.