Partial Label Learning with Discrimination Augmentation

Wei Wang
[†]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@seu.edu.cn

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
zhangml@seu.edu.cn

ABSTRACT

Partial label learning is a weakly supervised learning framework where each training example is associated with multiple candidate labels, among which only one is valid. Existing works on partial label learning mainly focus on classification model induction by disambiguating candidate label sets in the output space. Nevertheless, the feature representations of partial label training examples may be less informative of the ground-truth labels, which may result in negative influences on the disambiguation process. To circumvent this difficulty, the first attempt towards discrimination augmentation for partial label learning is investigated in this paper. The feature space is enriched with confidence-rated class prototype features to replenish discriminative characteristics of the underlying ground-truth labels for partial label training examples. Specially, an optimization formulation is proposed to jointly optimize the class prototype and estimate the labeling confidence over partial label training examples, which enforces both global consistency in the feature space and local consistency in the label space. We show that the class prototypes and the labeling confidence can be solved via alternating optimization. Extensive experiments on synthetic as well as real-world data sets validate the effectiveness of the proposed approach for improving the generalization performance of state-of-the-art partial label learning algorithms.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; *Learning paradigms.*

KEYWORDS

Partial label learning; Feature augmentation

ACM Reference Format:
Wei Wang and Min-Ling Zhang. 2022. Partial Label Learning with Discrimination Augmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539363

[*]Corresponding author.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539363

1 INTRODUCTION

In ordinary multi-class classification problems, each training example is represented by a single instance in the input space and associated with a label specifying the class to which the example belongs. Successful supervised learning generally requires large amounts of data with high-quality labels, while the collection of data with accurate annotations is expensive and time-consuming. To overcome this issue, weakly supervised learning has drawn considerable attention recently [47].

In partial label (PL) learning, each training example is represented by a single instance in the input space while associated with multiple candidate labels in the output space. It is assumed that only one label is the unknown ground-truth label, which is not accessible to the learning algorithm during training time [7, 12, 19, 43]. The need to learn from such inaccurate supervision information widely exists in real-world applications, such as web mining [20], image classification [4, 6, 40], natural language processing [25, 26, 46], ecoinformatics [2, 44], etc.

Although existing works on partial label learning have achieved great success by designing effective learning algorithms to identify the ground-truth labels, the limited supervision information of ambiguous annotations usually results in unsatisfactory generalization performance. Essentially, from the perspective of the data generation process, the PL examples are annotated with multiple candidate labels owing to the less informative and distinguishable feature representations. As a consequence, the unsatisfactory feature representations may have negative effects on the disambiguation process of PL learning algorithms to identify the underlying ground-truth labels. Recent works on applying deep learning techniques to partial label learning may alleviate this issue by learning better feature representations for PL training examples via end-to-end training [9, 10, 21, 28, 32, 36]. However, the end-to-end training strategy can't be directly employed by most well-established PL learning algorithms. Recent work on partial label dimension reduction [33] can be directly incorporated into partial label learning algorithms to achieve better generalization performance. However, the effectiveness of partial label dimension reduction techniques may become less satisfactory when the dimension of feature representations isn't high enough.

This paper presents the first attempt towards discrimination augmentation for partial label learning. A novel approach named PLDA, i.e. *Partial Label learning with Discrimination Augmentation*,

is proposed accordingly. The original feature space of PL training examples is enriched with confidence-rated class prototype features, which encode discriminative characteristics of the underlying ground-truth labeling information for PL training examples. An optimization formulation is proposed to jointly optimize the class prototypes and estimate the labeling confidence over PL training examples. This formulation combines local consistency in the label space and global consistency in the feature space in a unified framework, which can be solved via an alternating optimization strategy. Comprehensive experiments conducted over both synthetic and real-world partial label data sets demonstrate that the generalization performance of state-of-the-art partial label learning algorithms can be significantly improved by incorporating PLDA for feature augmentation.

The rest of this paper is organized as follows. Section 2 briefly discusses related works. Section 3 presents technical details of the proposed approach. Section 4 reports experimental results of comparative studies. Finally, Section 5 concludes.

2 RELATED WORK

Partial label learning is an emerging weakly supervised learning framework [47]. The ground-truth label of each PL training example is concealed within its candidate label set and not directly accessible to the learning algorithm. Generally, partial label learning is also related to other well-established weakly supervised learning problems such as semi-supervised learning [30, 48], label-noise learning [23, 31, 34, 45], multi-instance learning [1, 18], complementary-label learning [15] and partial multi-label learning [22, 35, 39].

To learn from PL training examples, a natural strategy is to purify the candidate label set to identify the true label. There are two widely-used disambiguation strategies, i.e. averaging-based disambiguation and identification-based disambiguation. Averaging-based disambiguation strategy considers each candidate label as an equal contribution to model induction. The final model prediction is made by averaging the modeling outputs from all candidate labels. For parametric models, the averaged modeling outputs of candidate labels is distinguished from modeling outputs of non-candidate labels [7]. For non-parametric models, the predicted label for an unseen instance is determined by voting among candidate labels of its neighbor examples [11, 14, 42]. One potential disadvantage of averaging-based disambiguation is that the modeling output of the true label may be overwhelmed by the modeling output of other false positive labels. Identification-based disambiguation strategy disambiguates the candidate label set by identifying the underlying ground-truth labeling information. The ground-truth label is considered as a latent variable and refined iteratively via EM procedure. Accordingly, the objective function for identification-based disambiguation can be defined based on maximum likelihood criterion [16, 19, 21] or maximum margin criterion [3, 24, 29]. One potential disadvantage of identification-based disambiguation lies in that the identified label may not be the ground-truth label, which will have negative effects on model training. Therefore, the effectiveness of disambiguation strategies may be affected by false positive labels in candidate label sets. To overcome this issue, another strategy transforms partial label learning into other well-established machine learning problems [5, 8, 27, 43].

Recent works have employed deep learning techniques to solve partial label learning problems by designing appropriate loss functions and model architectures, which have achieved great success [9, 10, 21, 28, 32, 36, 38]. Thanks to the powerful representation ability of neural networks, better feature representations of PL training examples can be obtained via the end-to-end training strategy. For example, effective learning approaches in deep semi-supervised learning are adapted to the background of partial label learning and have achieved better generalization performance than traditional partial label learning approaches, including consistency regularization [36], entropy minimization [38], generative adversarial networks [37], etc. Furthermore, the data generation process of PL training examples is investigated, and unbiased risk estimators are derived for PL training data [9, 10, 32]. The unbiased risk estimators can be equipped with any model, optimizer and loss function. Recently, other advanced representation learning techniques have been applied to partial label learning and achieved superior performance on benchmark data sets, including contrastive learning [28] and class activation map (CAM) [41].

Although representation learning has been successfully applied to partial label learning, the end-to-end training strategy can't be directly incorporated into most well-established partial label learning approaches equipped with traditional machine learning models. Recent works on partial label dimension reduction can be equipped with any partial label learning algorithm and achieve superior performance on data sets with high dimensions [33]. Nevertheless, the performance will be less satisfactory when the dimension of training examples isn't high enough. In the next section, a novel feature augmentation approach is proposed for partial label learning.

3 THE PROPOSED APPROACH

Let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional feature space and $\mathcal{Y} = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ denote the label space with q class labels. Given the set of PL training examples $\mathcal{D} = \{(\boldsymbol{x}_i, S_i) \mid 1 \le i \le n\}$, 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 $\boldsymbol{x}_i$. It is assumed that the ground-truth label y_i for each instance $\boldsymbol{x}_i$ is concealed within its candidate label set S_i, i.e. $y_i \in S_i$. The task of partial label learning is to learn a multi-class classifier $f : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}$.

Let $\boldsymbol{f}_i = [f_{i1}, f_{i2}, \ldots, f_{iq}]^\top$ denote the labeling confidence vector for $\boldsymbol{x}_i$ $(1 \le i \le n)$ with $f_{il} \in [0, 1]$ and $\sum_{l=1}^{q} f_{il} = 1$. Conceptually, f_{il} represents the probability of λ_l being the ground-truth label of $\boldsymbol{x}_i$. Correspondingly, we have the labeling confidence matrix $\mathbf{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n]^\top \in [0, 1]^{n \times q}$ for PL training examples. The labeling confidence matrix is initialized as follows:

$$\forall 1 \le i \le n: \quad f_{il} = \begin{cases} \dfrac{1}{|S_i|}, & \lambda_l \in S_i \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We choose to estimate the labeling confidence by *local consistency* which has been proven to be effective in identifying the underlying ground-truth label for PL examples [8, 17, 29, 42], where the manifold structure in the feature space should also be preserved in the label space. In this paper, a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{S})$ is constructed over the set of training examples. For each instance $\boldsymbol{x}_i$, let $\mathcal{N}(\boldsymbol{x}_i)$ denote the indexes of its k nearest neighbors identified in $\mathcal{D}$. Then, the similarity matrix $\mathbf{S} = [s_{ij}]_{n \times n}$ over PL

training examples is instantiated as: $s_{ij} = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/\sigma^2\right)$ if $j \in \mathcal{N}(\boldsymbol{x}_i)$ and $s_{ij} = 0$ otherwise. Here, the parameter σ is defined as $\sigma = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_{i_k}\|_2/n$ where $\boldsymbol{x}_{i_k}$ is the k-th nearest neighbor of $\boldsymbol{x}_i$. To ensure that the similarity matrix is symmetric, we set it to be $\mathbf{S} + \mathbf{S}^\top$. Then we introduce the following optimization problem:

$$\begin{aligned} \min_{\mathbf{F}} \ & \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \| \frac{\boldsymbol{f}_i}{\sqrt{d_{ii}}} - \frac{\boldsymbol{f}_j}{\sqrt{d_{jj}}} \|_2^2 \\ \text{s.t.} \ & \sum_{l=1}^{q} f_{il} = 1, \quad 1 \le i \le n \\ & f_{il} \ge 0, \quad 1 \le i \le n,\ y_l \in S_i \\ & f_{il} = 0, \quad 1 \le i \le n,\ y_l \notin S_i. \end{aligned} \tag{2}$$

Here, $d_{ii} = \sum_{j=1}^{n} s_{ij}$ is the degree of vertex $\boldsymbol{x}_i$ in the graph. The first constraint indicates the normalization property of the labeling confidence vector. The second and third constraints imply that the ground-truth label is restricted within the candidate label set.

After that, we introduce the *class prototype* in order to capture the discriminative characteristic of each class label respectively. Concretely, let $\boldsymbol{c}_l \in \mathbb{R}^d$ denote the class prototype for λ_l. Correspondingly, we have the class prototype matrix $\mathbf{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_q]^\top \in \mathbb{R}^{q\times d}$. We consider the *global consistency*, where the feature representation of each PL training example should be closer to the prototype feature representation corresponding to its ground-truth label. Then we get the following optimization problem:

$$\min_{\mathbf{C}} \sum_{i=1}^{n}\sum_{l=1}^{q} f_{il}\|\boldsymbol{x}_i - \boldsymbol{c}_l\|_2^2. \tag{3}$$

As shown in the objective function, the pairwise distance between $\boldsymbol{x}_i$ and $\boldsymbol{c}_j$ is further rated by f_{il} to account for the labeling confidence of λ_l being the ground-truth label for $\boldsymbol{x}_i$. In this paper, we propose a unified framework to jointly optimize the labeling confidence and class prototype by combining Eq.(2) and Eq.(3). Thus the labeling confidence can be optimized by considering both local and global consistency. The optimization problem can be stated as:

$$\begin{aligned} \min_{\mathbf{F},\mathbf{C}} \ & \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} \| \frac{\boldsymbol{f}_i}{\sqrt{d_{ii}}} - \frac{\boldsymbol{f}_j}{\sqrt{d_{jj}}} \|_2^2 + \mu \sum_{i=1}^{n}\sum_{l=1}^{q} f_{il}\|\boldsymbol{x}_i - \boldsymbol{c}_l\|_2^2 \\ \text{s.t.} \ & \sum_{l=1}^{q} f_{il} = 1, \quad 1 \le i \le n \\ & f_{il} \ge 0, \quad 1 \le i \le n,\ y_l \in S_i \\ & f_{il} = 0, \quad 1 \le i \le n,\ y_l \notin S_i \end{aligned} \tag{4}$$

where μ is a trade-off parameter between the preservation error of local consistency and that of global consistency. To solve the derived problem, alternating optimization is employed to iteratively update $\mathbf{F}$ and $\mathbf{C}$.

Fix C, update F When $\mathbf{C}$ is fixed, Eq.(4) corresponds to a quadratic programming (QP) problem. For ease of notations, we introduce a matrix $\mathbf{K} = [k_{il}]_{n\times q}$ and each element k_{il} is defined as $k_{il} = \|\boldsymbol{x}_i - \boldsymbol{c}_l\|_2^2$. Let $\widetilde{\boldsymbol{f}} = \mathrm{vec}(\mathbf{F})$ where $\mathrm{vec}(\cdot)$ is the vectorization operator. Accordingly, we have $\widetilde{\boldsymbol{k}} = \mathrm{vec}(\mathbf{K})$. Thereafter, the optimization problem in Eq.(4) turns out to be:

$$\begin{aligned} \min_{\widetilde{\boldsymbol{f}}} \ & \frac{1}{2}\widetilde{\boldsymbol{f}}^\top \mathbf{H}\widetilde{\boldsymbol{f}} + \widetilde{\boldsymbol{f}}^\top \boldsymbol{p} \\ \text{s.t.} \ & \sum_{l=1}^{q} f_{il} = 1, \quad 1 \le i \le n \\ & f_{il} \ge 0, \quad 1 \le i \le n,\ y_l \in S_i \\ & f_{il} = 0, \quad 1 \le i \le n,\ y_l \notin S_i \end{aligned} \tag{5}$$

where $\boldsymbol{p} = \mu\widetilde{\boldsymbol{k}}$ and $\mathbf{H} \in \mathbb{R}^{nl\times nl}$ is defined as

$$\mathbf{H} = \begin{bmatrix} \mathbf{T} & \mathbf{0}_{n\times n} & \cdots & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times n} & \mathbf{T} & \cdots & \mathbf{0}_{n\times n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \cdots & \mathbf{T} \end{bmatrix}. \tag{6}$$

Here, $\mathbf{T} \in \mathbb{R}^{n\times n}$ is a square matrix defined as $\mathbf{T} = 4(\mathbf{I}_{q\times q} - \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}})$ where $\mathbf{D}$ is a diagonal matrix with its diagonal element defined as $d_{ii} = \sum_{j=1}^{n} s_{ij}$. Thereafter, optimization problem Eq.(5) can be efficiently solved by any off-the-shelf QP toolbox.

Fix F, update C When $\mathbf{F}$ is fixed, optimization problem Eq.(4) reduces to Eq.(3). By setting the gradient of the optimization objective in Eq.(3) w.r.t. $\mathbf{C}$ to zeros, we can obtain the closed-form solution of $\boldsymbol{c}_l$ as

$$\boldsymbol{c}_l = \frac{\sum_{i=1}^{n} f_{il}\boldsymbol{x}_i}{\sum_{i=1}^{n} f_{il}}. \tag{7}$$

As the alternating optimization procedure for $\mathbf{F}$ and $\mathbf{C}$ terminates, the original feature representation of the PL training example is enriched with *confidence-rated class prototype features*. In this way, the feature representations of the PL training example can be more informative of its concealed ground-truth label, which can facilitate the disambiguation process of partial label learning algorithms. Let $\overline{\boldsymbol{f}}_i = [\overline{f}_{i1}, \overline{f}_{i2}, \ldots, \overline{f}_{iq}]^\top$ denote the ultimate labeling confidence vector of $\boldsymbol{x}_i$ at the end of iterations, then the augmented feature vector $\boldsymbol{\Delta}_{\boldsymbol{x}_i}$ for $\boldsymbol{x}_i$ is defined as

$$\boldsymbol{\Delta}_{\boldsymbol{x}_i} = \mathbf{C}^\top \overline{\boldsymbol{f}}_i. \tag{8}$$

After that, the original PL training set $\mathcal{D}$ is transformed into

$$\widehat{\mathcal{D}} = \{(\widehat{\boldsymbol{x}}_i, S_i) \mid 1 \le i \le n\}, \quad \text{where } \widehat{\boldsymbol{x}}_i = [\boldsymbol{x}_i; \boldsymbol{\Delta}_{\boldsymbol{x}_i}]. \tag{9}$$

Here, each instance $\widehat{\boldsymbol{x}}_i$ belongs to the augmented feature space $\widehat{\mathcal{X}}$ which is the concatenation of $\mathcal{X}$ and the confidence-rated class prototype feature space. Thereafter, a multi-class classifier $f : \widehat{\mathcal{X}} \to \mathcal{Y}$ can be induced from $\widehat{\mathcal{D}}$ by applying a partial label learning algorithm $\mathcal{A}$, i.e. $f \leftarrow \mathcal{A}(\widehat{\mathcal{D}})$.

During the testing phase, for an unseen instance $\boldsymbol{x}^*$, the class membership is identified by resorting to the kNN strategy. The k nearest neighbors in the training set, i.e. $\mathcal{N}(\boldsymbol{x}^*)$, are firstly identified. After that, the similarity score w_{*i} between $\boldsymbol{x}^*$ and PL training example $\boldsymbol{x}_i$ is determined as: $w_{*i} = \exp\left(-\|\boldsymbol{x}^* - \boldsymbol{x}_i\|_2^2/\sigma^2\right)$ if $i \in \mathcal{N}(\boldsymbol{x}^*)$ and $w_{*i} = 0$ otherwise. We further normalize the similar score as $h_{*i} = \frac{w_{*i}}{\sum_{i=1}^{n} w_{*i}}$. Thereafter, the class membership for the unseen instance is determined by kNN labeling confidence

Table 1: The pseudo-code of PLDA.

Input:

$\mathcal{D}$: the PL training set $\{(\boldsymbol{x}_i, S_i) \mid 1 \le i \le n\}$ ($\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}, \boldsymbol{x}_i \in \mathcal{X}, S_i \subseteq \mathcal{Y}$)

k: the number of nearest neighbors used for weighted graph construction

μ: the trade-off parameter in objective function (4)

$\mathcal{A}$: the partial label learning algorithm

$\boldsymbol{x}^*$: the unseen instance to be classified

Output:

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

Process:

1: Set the similarity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{S})$ with $\mathcal{V} = \{\boldsymbol{x}_i \mid 1 \le i \le n\}$ and $\mathcal{E} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid j \in \mathcal{N}(\boldsymbol{x}_i)\}$;
2: Initialize the $n \times q$ labeling confidence matrix $\mathbf{F}$ according to Eq.(1);
3: Initialize the $q \times d$ class prototype matrix $\mathbf{C}$ according to Eq.(7);
4: **repeat**
5: Update $\mathbf{F}$ by solving the QP problem in Eq.(4);
6: Update $\mathbf{C}$ according to Eq.(7);
7: **until** convergence or maximum number of iterations being reached
8: Form the transformed PL training set $\widehat{\mathcal{D}} = \{(\widehat{\boldsymbol{x}}_i, S_i) \mid 1 \le i \le n\}$ according to Eq.(9);
9: Induce multi-class classifier f base on $\widehat{\mathcal{D}}$: $f \leftarrow\!\shortmid \mathcal{A}(\widehat{\mathcal{D}})$;
10: Determine the class membership c^* of $\boldsymbol{x}^*$ according to Eq.(10);
11: Generate the transformed feature vector for the unseen instance as $\widehat{\boldsymbol{x}}^* = [\boldsymbol{x}^*; \boldsymbol{c}_{c^*}]$;
12: Return $y^* = f(\widehat{\boldsymbol{x}}^*)$.

aggregation:

$$c^* = \arg\max_{1 \le l \le q} \sum_{i=1}^{n} h_{*i} \overline{f}_{il}. \tag{10}$$

After that, the transformed feature vector $\widehat{\boldsymbol{x}}^*$ of the unseen instance is generated by concatenating the original feature vector $\boldsymbol{x}^*$ and the class prototype $\boldsymbol{c}_{c^*}$ according to its class membership, i.e. $\widehat{\boldsymbol{x}}^* = [\boldsymbol{x}^*; \boldsymbol{c}_{c^*}]$. Finally, the unseen instance can be classified by feeding the transformed feature vector $\widehat{\boldsymbol{x}}^*$ into f.

In summary, Table 1 gives the pseudo-code of PLDA. Firstly, the labeling confidence matrix $\mathbf{F}$ and class prototype matrix $\mathbf{C}$ are initialized (Step 1-3). Then an alternating optimization procedure is invoked to update $\mathbf{F}$ and $\mathbf{C}$ iteratively (Step 4-7). A multi-class classifier is induced by learning from augmented PL training data (Step 8-9). Finally, the label for an unseen instance is predicted based on the augmented feature vector as well (Step 10-12).

4 EXPERIMENTS

4.1 Experimental Setup

To validate the effectiveness of the proposed partial label feature augmentation approach, PLDA is coupled with start-of-the-art partial label learning algorithms for performance evaluation. For any partial label learning algorithm $\mathcal{A}$, its coupling version with PLDA is denoted as $\mathcal{A}$-PLDA which learns from PL training examples with augmented features. The predictive performance of $\mathcal{A}$-PLDA is compared with that of $\mathcal{A}$ to manifest whether the proposed feature augmentation technique does help improve the generalization performance of partial label learning algorithms.

In this paper, four well-established partial label learning algorithms are employed to instantiate $\mathcal{A}$ with suggested configurations in respective literature:

- PL-KNN [14]: an averaging-based partial label learning approach which makes predictions for testing instances via weighted kNN labeling information aggregation [suggested configuration: k=10].
- PL-SVM [24]: an identification-based partial label learning approach which learns the predictive model by maximizing the classification margin between candidate and non-candidate labels [suggested configuration: regularization parameter pool with $\{10^{-3}, \ldots, 10^{3}\}$].
- PL-ECOC [43]: a transformation-based partial label learning approach which learns the predictive model by decomposing the original partial label learning problem into several binary classification problems via error-correcting output codes (ECOC) [suggested configuration: ECOC coding length $\lceil 10 \cdot \log_2(q) \rceil$].
- CLPL [7]: an averaging-based partial label learning approach which learns predictive model by minimizing a convex loss function adapted for PL training examples [suggested setup: SVM with square hinge loss].

For PLDA, the parameters are set as $k = 10$ and $\lambda = 0.01$. The sensitivity analysis of parameter configurations is conducted in

Table 2: Characteristics of the controlled UCI experimental data sets.

Data Set	# Examples	# Features	# Class Labels	# False Positive Labels (r)
glass	214	9	6	$r = 1, 2, 3$
ecoli	336	7	8	$r = 1, 2, 3$
deter	358	23	6	$r = 1, 2, 3$
aggregation	788	2	7	$r = 1, 2, 3$
vowel	871	3	6	$r = 1, 2, 3$
segment	2,310	18	7	$r = 1, 2, 3$
abalone	4,177	7	29	$r = 1, 2, 3$
robot navigation	5,456	24	4	$r = 1, 2$
satimage	6,435	24	6	$r = 1, 2, 3$
usps	9,298	256	10	$r = 1, 2, 3$
pendigits	10,992	16	10	$r = 1, 2, 3$
letter	20,000	16	26	$r = 1, 2, 3$

Table 3: Classification accuracy (mean±std) of each comparing algorithm on controlled UCI data sets (with one false positive candidate label [$r = 1$]). For partial label learning algorithm $\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, CLPL}, the performance of $\mathcal{A}$-PLDA is compared against that of $\mathcal{A}$ where the better performance is shown in boldface.

Data Set	Comparing Algorithm							
	PL-KNN	PL-KNN-PLDA	PL-SVM	PL-SVM-PLDA	PL-ECOC	PL-ECOC-PLDA	CLPL	CLPL-PLDA
glass	0.573±0.071	**0.589±0.060**	0.511±0.054	**0.522± 0.058**	0.584±0.077	**0.603±0.054**	0.520±0.061	**0.615±0.065**
ecoli	0.821±0.028	**0.840±0.032**	0.788±0.048	**0.801±0.045**	0.840±0.033	**0.842±0.032**	0.836±0.033	**0.842±0.033**
deter	0.892±0.022	**0.912±0.023**	0.881±0.033	**0.885±0.046**	0.920±0.033	**0.921±0.026**	**0.917±0.014**	0.913±0.024
aggregation	0.996±0.003	**0.997±0.003**	0.737±0.041	**0.782±0.057**	0.992±0.005	**0.997±0.003**	0.815±0.037	**0.836±0.037**
vowel	0.826±0.017	**0.837±0.022**	0.527±0.051	**0.530±0.049**	0.823±0.026	**0.835±0.023**	0.611±0.033	**0.743±0.036**
segment	0.910±0.009	**0.921±0.008**	0.747±0.023	**0.800±0.033**	0.909±0.010	**0.923±0.008**	0.815±0.007	**0.923±0.009**
abalone	**0.228±0.007**	0.224±0.007	0.097±0.047	**0.159±0.029**	**0.258±0.008**	0.238±0.008	**0.229±0.011**	0.225±0.010
robot navigation	0.773±0.010	**0.797±0.008**	0.597±0.019	**0.668±0.011**	**0.823±0.009**	0.806±0.008	0.622±0.011	**0.806±0.008**
satimage	0.886±0.008	**0.892±0.007**	0.756±0.009	**0.804±0.025**	0.871±0.006	**0.892±0.007**	0.773±0.007	**0.892±0.007**
usps	0.942±0.003	**0.950±0.002**	0.916±0.003	**0.950±0.002**	**0.955±0.003**	0.950±0.002	0.872±0.006	**0.950±0.002**
pendigits	0.984±0.002	**0.989±0.002**	0.804±0.006	**0.949±0.015**	0.987±0.002	**0.989±0.002**	0.845±0.004	**0.989±0.002**
letter	0.898±0.003	**0.921±0.003**	0.553±0.017	**0.710±0.020**	0.851±0.010	**0.923±0.004**	0.495±0.010	**0.859±0.030**

Subsection 4.4. We perform ten runs of 50%/50% random train/test splits on each synthetic and real-world partial label data set, and the mean accuracy as well as standard deviation are recorded.

4.2 Controlled UCI data sets

Following the widely-used experimental protocol in partial label learning [4, 5, 7, 11], synthetic PL data sets are generated from multi-class UCI data sets with controlling parameter r. Here, for any multi-class example $(\boldsymbol{x}_i, y_i)$, one synthetic PL example $(\boldsymbol{x}_i, S_i)$ is generated by replenishing r labels $\Omega_r \subseteq \mathcal{Y} \setminus \{y_i\}$ into S_i at random, i.e. $S_i = \Omega_r \bigcup \{y_i\}$.[1] [2]

Table 2 summarizes the characteristics of twelve controlled UCI data sets with $r = 1, 2, 3$ for performance evaluation which are ordered by the size of the data set. Accordingly, Table 3, 4 and 5 report the detailed experimental results with $r = 1, 2, 3$ respectively. For each partial label learning algorithm $\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, CLPL}, $\mathcal{A}$-PLDA is compared with $\mathcal{A}$ where the best classification performance is shown in boldface.

Furthermore, pairwise t-test at 0.05 significance level is conducted to demonstrate whether the performance difference between $\mathcal{A}$-PLDA and $\mathcal{A}$ is significant, where the resulting win/tie/loss counts are reported in Table 6. Based on the above experimental results, we can draw the following conclusions:

- For the averaging-based disambiguation approach PL-KNN and CLPL, the generalization performance has been greatly improved by incorporating the proposed feature augmentation technique. Especially, PL-KNN-PLDA and CLPL-PLDA achieve better performance than PL-KNN and CLPL in 77.1% and 74.3% cases respectively while have been outperformed by them in none of the cases.

[1] For `robot navigation`, the setting $r = 3$ is not considered as there are only four class labels in the label space.

[2] For several data sets with $r = 2$ or $r = 3$, PL-ECOC can't converge. We set $S_i = \{y_i\}$ for 20% of training examples on these data sets.

Table 4: Classification accuracy (mean±std) of each comparing algorithm on controlled UCI data sets (with two false positive candidate labels [$r = 2$]). For partial label learning algorithm $\mathcal{A} \in \{$PL-KNN, PL-SVM, PL-ECOC, CLPL$\}$, the performance of $\mathcal{A}$-PLDA is compared against that of $\mathcal{A}$ where the better performance is shown in boldface.

Data Set	Comparing Algorithm							
	PL-KNN	PL-KNN-PLDA	PL-SVM	PL-SVM-PLDA	PL-ECOC	PL-ECOC-PLDA	CLPL	CLPL-PLDA
glass	0.519±0.046	**0.544±0.061**	0.443±0.045	**0.490±0.083**	0.312±0.118	**0.346±0.101**	0.427±0.071	**0.551±0.046**
ecoli	0.817±0.038	**0.826±0.038**	0.771±0.037	**0.789±0.029**	0.823±0.030	**0.824±0.028**	0.819±0.044	**0.838±0.025**
deter	0.858±0.042	**0.874±0.029**	0.846±0.052	**0.849±0.053**	0.846±0.039	**0.851±0.043**	**0.875±0.062**	0.873±0.050
aggregation	0.993±0.004	**0.995±0.004**	0.723±0.039	**0.753±0.060**	0.950±0.027	**0.981±0.027**	0.823±0.025	**0.843±0.012**
vowel	0.802±0.011	**0.822±0.013**	0.489±0.027	**0.499±0.048**	0.758±0.050	**0.799±0.025**	0.609±0.048	**0.724±0.066**
segment	0.895±0.009	**0.909±0.006**	0.661±0.043	**0.726±0.053**	0.869±0.010	**0.910±0.008**	0.810±0.010	**0.910±0.007**
abalone	0.222±0.008	**0.230±0.004**	0.072±0.039	**0.133± 0.066**	**0.235±0.015**	0.219±0.012	0.229±0.008	**0.232±0.010**
robot navigation	0.684±0.006	**0.713±0.011**	0.402±0.084	**0.469±0.080**	0.785±0.019	**0.787±0.010**	0.598±0.011	**0.748±0.011**
satimage	0.868±0.006	**0.881±0.005**	0.751±0.004	**0.786±0.009**	0.818±0.041	**0.851±0.021**	0.764±0.006	**0.882±0.006**
usps	0.940±0.003	**0.948±0.003**	0.913±0.005	**0.948±0.003**	0.935±0.005	**0.948±0.003**	0.865±0.004	**0.948±0.003**
pendigits	0.983±0.002	**0.987±0.002**	0.778±0.022	**0.885±0.025**	0.974±0.004	**0.987±0.002**	0.838±0.005	**0.987±0.002**
letter	0.891±0.003	**0.914±0.003**	0.518±0.011	**0.649±0.023**	0.793±0.009	**0.913±0.004**	0.476±0.011	**0.830±0.024**

Table 5: Classification accuracy (mean±std) of each comparing algorithm on controlled UCI data sets (with three false positive candidate labels [$r = 3$]). For partial label learning algorithm $\mathcal{A} \in \{$PL-KNN, PL-SVM, PL-ECOC, CLPL$\}$, the performance of $\mathcal{A}$-PLDA is compared against that of $\mathcal{A}$ where the better performance is shown in boldface.

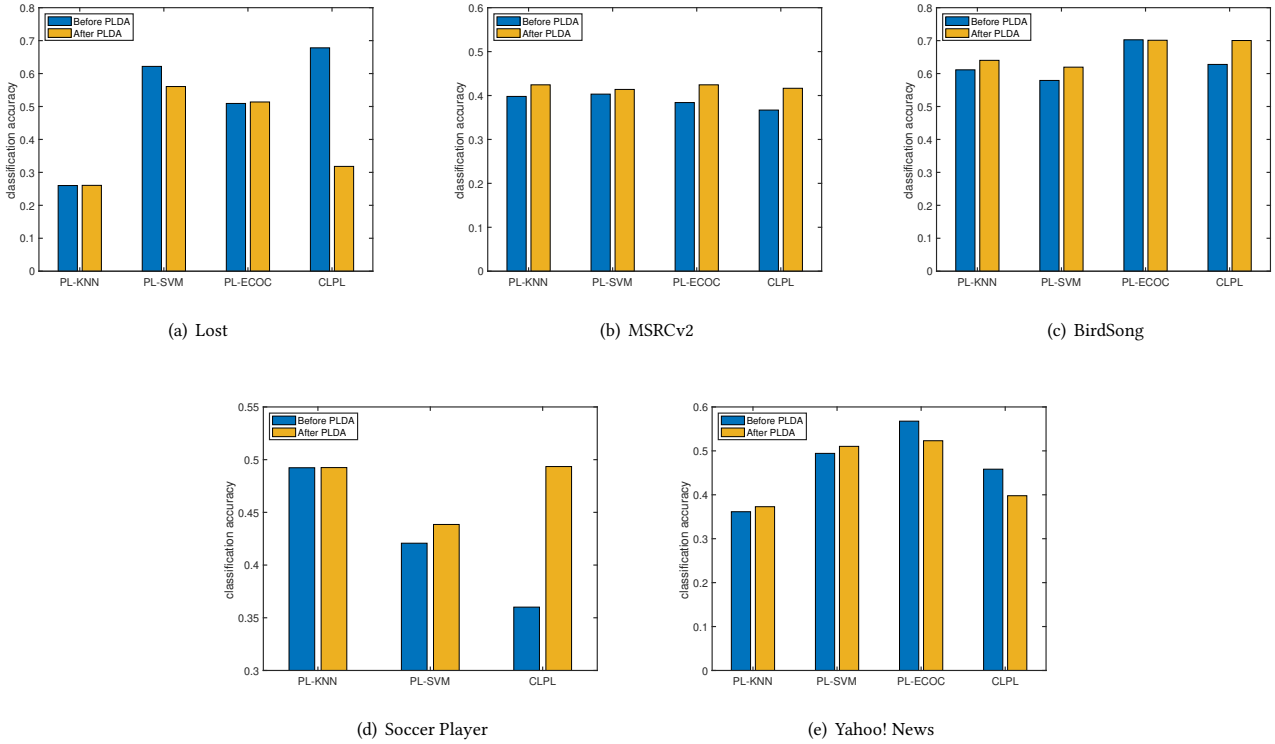
Data Set	Comparing Algorithm							
	PL-KNN	PL-KNN-PLDA	PL-SVM	PL-SVM-PLDA	PL-ECOC	PL-ECOC-PLDA	CLPL	CLPL-PLDA
glass	0.522±0.054	**0.542±0.042**	0.367±0.083	**0.400±0.084**	0.435±0.105	**0.482±0.089**	0.366±0.093	**0.507±0.057**
ecoli	0.760±0.023	**0.788±0.028**	0.742±0.068	**0.752±0.037**	0.814±0.025	**0.824±0.032**	**0.823±0.018**	0.814±0.025
deter	0.820±0.025	**0.869±0.026**	0.804±0.058	**0.832±0.037**	0.858±0.051	**0.864±0.050**	0.792±0.040	**0.903±0.029**
aggregation	0.988±0.007	**0.994±0.004**	0.720±0.057	**0.739±0.040**	0.981±0.026	**0.991±0.013**	0.790±0.051	**0.808±0.054**
vowel	0.742±0.026	**0.789±0.020**	0.440±0.090	**0.454±0.060**	0.814±0.022	**0.834±0.021**	0.606±0.048	**0.678±0.107**
segment	0.873±0.014	**0.892±0.012**	0.581±0.037	**0.600±0.046**	0.893±0.011	**0.911±0.006**	0.809±0.014	**0.898±0.009**
abalone	0.201±0.007	**0.211±0.010**	0.045±0.040	**0.085±0.062**	0.162±0.052	**0.173±0.025**	**0.227±0.010**	0.225±0.009
satimage	0.828±0.009	**0.851±0.006**	0.707±0.087	**0.778±0.012**	0.859±0.007	**0.879±0.006**	0.760±0.008	**0.866±0.008**
usps	0.932±0.004	**0.942±0.003**	0.906±0.008	**0.939±0.003**	0.945±0.004	**0.947±0.003**	0.860±0.006	**0.939±0.003**
pendigits	0.984±0.002	**0.988±0.001**	0.673±0.070	**0.804±0.086**	0.981±0.003	**0.989±0.001**	0.838±0.005	**0.988±0.001**
letter	0.885±0.004	**0.908±0.003**	0.466±0.025	**0.599±0.022**	0.720±0.008	**0.881±0.015**	0.472±0.010	**0.818±0.033**

- For the identification-based disambiguation approach PL-SVM, PL-SVM-PLDA achieves superior performance against PL-SVM in more than 50% cases while has been outperformed by PL-SVM in none of the cases.
- For the transformation-based disambiguation approach PL-ECOC, PL-ECOC-PLDA significantly outperforms PL-ECOC in 51.4% cases while has been outperformed in 11.4% cases.

4.3 Real-World Data Sets

Table 7 summarizes characteristics of real-world PL data sets from different task domains, including `Lost` [7], `Soccer Player` [40] and `Yahoo! News` [13] for automatic face naming from images or videos, `MSRCv2` [19] for object classification and `BirdSong` for bird-song classification. For the task of *automatic face naming*, faces cropped from images or video frames are treated as instances while names extracted from the associated captions or subtitles are regarded as candidate labels. For the task of *object classification*, image segmentation is represented as instances while objects appearing within the same image are regarded as candidate labels. For the task of *bird song classification*, singing syllables of birds are treated as instances while bird species jointly singing during 10 seconds are regarded as candidate labels.

Figure 1 illustrates the predictive accuracy of each partial label learning approach before and after employing the proposed feature augmentation technique. Furthermore, Table 8 reports the



(a) Lost (b) MSRCv2 (c) BirdSong

(d) Soccer Player (e) Yahoo! News

Figure 1: Comparison of the classification accuracy of each partial label learning algorithm on real-world data sets before (*blue* bar) and after (*yellow* bar) employing PLDA.

Table 6: Win/tie/loss counts (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-PLDA and $\mathcal{A}$ in terms of different number of false positive labels ($r = 1, 2, 3$).

	$\mathcal{A}$-PLDA **against** $\mathcal{A}$			
	$\mathcal{A}$=PL-KNN	$\mathcal{A}$=PL-SVM	$\mathcal{A}$=PL-ECOC	$\mathcal{A}$=CLPL
$r = 1$	9/3/0	8/4/0	5/4/3	9/3/0
$r = 2$	8/4/0	6/6/0	7/4/1	9/3/0
$r = 3$	10/1/0	4/7/0	6/5/0	8/3/0
In Total	**27/8/0**	**18/17/0**	**18/13/4**	**26/9/0**

win/tie/loss statistic based on pairwise t-test at 0.05 significance level on each real-world experimental data set. From the above results, we can observe that:

- Out of nineteen statistical comparisons on real-world data sets, $\mathcal{A}$-PLDA achieves superior or comparable performance against $\mathcal{A}$ in sixteen cases and has been outperformed by $\mathcal{A}$ in only three cases.
- For the averaging-based disambiguation approach PL-KNN and the identification-based disambiguation approach PL-SVM, PL-KNN-PLDA and PL-SVM-PLDA achieve better predictive performance in 60% cases respectively, and have been outperformed in none of the cases.

4.4 Further Analysis

Sensitivity Analysis As shown in Table 1, PLDA employ k nearest neighbors to construct the similarity graph and the trade-off parameter μ to balance the preservation error of local consistency in the label space and that of global consistency in the feature space. To investigate the performance sensitivity of PLDA w.r.t. k and μ, Figure 2 shows how the classification accuracy of PLDA changes as k and μ vary. Here, two real-world data sets `MSRCv2` and `BirdSong` and two controlled UCI data sets `deter` and `usps` with $r = 2$ are employed for illustrative purposes.

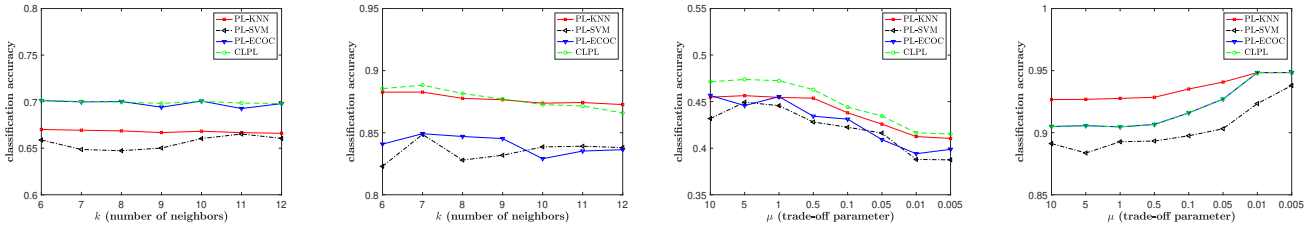
To investigate how the performance changes with different k, we fix $\lambda = 0.01$ and report the classification accuracy of $\mathcal{A}$-PLDA with k varying from 6 to 12 with interval 1 on `BirdSong` and `deter` ($r = 2$). We can observe that the classification performance is relatively stable when k varies. Therefore, we set $k = 10$ in this paper.

We also demonstrate how the predictive accuracy will change by employing different trade-off parameter μ. We fix $k = 10$ and report the classification accuracy of $\mathcal{A}$-PLDA with different $\mu \in [0.005, 1]$. It is observed that the predictive performance is generally stable for each PL learning algorithm with feature augmentation. Therefore, we set $\mu = 0.01$ in this paper for convenience.

Convergence Analysis Figure 3 illustrates the convergence curve of PLDA by calculating the difference of the labeling confidence matrix $\mathbf{F}$ and the class prototype matrix $\mathbf{C}$ between two adjacent iterations. It is observed that the labeling confidence matrix

Table 7: Characteristics of the real-world experimental data sets.

Data Set	# Examples	# Features	# Class Labels	average # Candidate Labels	Task Domain
Lost	1,122	108	16	2.23	*automatic face naming* [7]
MSRCv2	1,758	48	23	3.16	*object classification* [19]
BirdSong	4,998	38	13	2.18	*bird song classification* [2]
Soccer Player	17,472	279	171	2.09	*automatic face naming* [40]
Yahoo! News	22,991	163	219	1.91	*automatic face naming* [13]

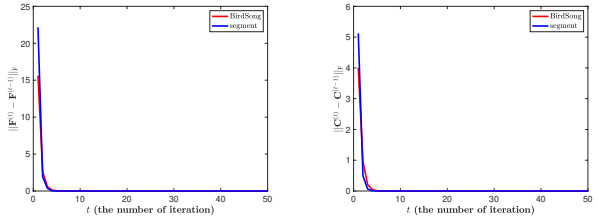


(a) BirdSong (b) deter ($r = 2$) (c) MSRCv2 (d) usps ($r = 2$)

Figure 2: Parameter sensitivity analysis for PLDA. (a) Classification accuracy of $\mathcal{A}$-PLDA ($\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, CLPL}) on `BirdSong` by varying k; (b) Classification accuracy of $\mathcal{A}$-PLDA ($\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, CLPL}) on `deter` with $r = 2$ by varying k; (c) Classification accuracy of $\mathcal{A}$-PLDA ($\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, CLPL}) on `MSRCv2` by varying μ; (d) Classification accuracy of $\mathcal{A}$-PLDA ($\mathcal{A} \in$ {PL-KNN, PL-SVM, PL-ECOC, CLPL}) on `usps` with $r = 2$ by varying μ.

Table 8: Win/tie/loss statistic (pairwise t-test at 0.05 significance level) between $\mathcal{A}$-PLDA and $\mathcal{A}$ on each real-world partial label data set.

	$\mathcal{A}$-PLDA against $\mathcal{A}$			
	$\mathcal{A}$=PL-KNN	$\mathcal{A}$=PL-SVM	$\mathcal{A}$=PL-ECOC	$\mathcal{A}$=CLPL
Lost	tie	tie	tie	loss
MSRCv2	win	tie	win	win
BirdSong	win	win	tie	win
Soccer Player	tie	win	N/A	win
Yahoo! News	win	win	loss	loss
In Total	3/2/0	3/2/0	1/2/1	3/0/2



(a) **F** (b) **C**

Figure 3: Convergence curves of F and C (on `BirdSong` and `segment` with $r = 2$).

and the class prototype matrix converge quickly with increasing number of iterations. Therefore, the convergence of the proposed approach is demonstrated empirically.

5 CONCLUSION

In this paper, the problem of discrimination augmentation for partial label learning is investigated. The original feature space is enriched with confidence-rated class prototype features to replenish discriminative information of underlying ground-truth labels. This paper proposes an optimization problem to jointly estimate the labeling confidence and the class prototypes, which can be solved via alternating optimization. Extensive experimental results clearly validate the effectiveness of the proposed feature augmentation technique for partial label learning.

In future, it is interesting to investigate effective data augmentation techniques for deep learning-based partial label learning approaches. Furthermore, it is desirable to extend our approach to other weakly supervised learning scenarios.

ACKNOWLEDGMENTS

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 (62176055). We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

REFERENCES

[1] Jaume Amores. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201 (2013), 81–105.

[2] Forrest Briggs, Xiaoli Z. Fern, and Raviv Raich. 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*. Beijing, China, 534–542.

[3] Jing Chai, Ivor W. Tsang, and Weijie Chen. 2020. Large margin partial label machine. *IEEE Transactions on Neural Networks and Learning Systems* 31, 7 (2020), 2594–2608.

[4] Ching-Hui Chen, Vishal M. Patel, and Rama Chellappa. 2018. Learning from ambiguously labeled face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 7 (2018), 1653–1667.

[5] Yi-Chen Chen, Vishal M. Patel, Rama Chellappa, and P. Jonathon Phillips. 2014. Ambiguously labeled learning using dictionaries. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2076–2088.

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

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

[8] Lei Feng and Bo An. 2018. Leveraging latent label distributions for partial label learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 2107–2113.

[9] Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. 2020. Learning with Multiple Complementary Labels. In *Proceedings of the 37th International Conference on Machine Learning*. Virtual Event, 3072–3081.

[10] Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. 2020. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems 33*. Virtual Event, 10948–10960.

[11] Chen Gong, Tongliang Liu, Yuanyan Tang, Jian Yang, Jie Yang, and Dacheng Tao. 2018. A regularization approach for instance-based superset label learning. *IEEE Transactions on Cybernetics* 48, 3 (2018), 967–978.

[12] Xiuwen Gong, Dong Yuan, and Wei Bao. 2022. Discriminative metric learning for partial label learning. *IEEE Transactions on Neural Networks and Learning Systems* (2022), in press.

[13] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multiple instance metric learning from automatically labeled bags of faces. In *Lecture Notes in Computer Science 6311*, K. Daniilidis, P. Maragos, and N. Paragios (Eds.). Springer, Berlin, 634–647.

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

[15] Takashi Ishida, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. 2019. Complementary-Label Learning for Arbitrary Losses and Models. In *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, CA, 2971–2980.

[16] Rong Jin and Zoubin Ghahramani. 2002. Learning with multiple labels. In *Advances in Neural Information Processing systems 15*. Vancouver, BC, 897–904.

[17] Changchun Li, Ximing Li, and Jihong Ouyang. 2020. Learning with noisy partial labels by simultaneously leveraging global and local consistencies. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Virtual Event, 725–734.

[18] Xin-Chun Li, De-Chuan Zhan, Jia-Qi Yang, and Yi Shi. 2021. Deep multiple instance selection. *Science China Information Sciences* 64, 3 (2021), Article 130102.

[19] Liping Liu and Thomas G. Dietterich. 2012. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25*. Lake Tahoe, NV, 548–556.

[20] Jie Luo and Francesco Orabona. 2010. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems 23*. Vancouver, BC, 1504–1512.

[21] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. 2020. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning*. Virtual Event, 6500–6510.

[22] Gengyu Lyu, Songhe Feng, and Yidong Li. 2020. Partial multi-label learning via probabilistic graph matching mechanism. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event, 105–113.

[23] Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi N. R. Wijewickrema, and James Bailey. 2018. Dimensionality-Driven Learning with Noisy Labels. In *Proceedings of the 35th International Conference on Machine Learning*. Stockholm, Sweden, 3361–3370.

[24] Nam Nguyen and Rich Caruana. 2008. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, NV, 551–559.

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

[26] Xiang Ren, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, and Jiawei Han. 2016. 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*. San Francisco, CA, 1825–1834.

[27] Deng-Bao Wang, Min-Ling Zhang, and Li Li. 2022. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), in press.

[28] Haobo Wang, Ruixuan Xiao, Sharon Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. 2022. Contrastive label disambiguation for partial label learning. In *Proceedings of the 10th International Conference on Learning Representations*. Virtual Event, in press.

[29] Wei Wang and Min-Ling Zhang. 2020. Semi-supervised partial label learning via confidence-rated margin maximization. In *Advances in Neural Information Processing Systems 33*. Virtual Event, 6982–6993.

[30] Tong Wei, Hai Wang, Wei-Wei Tu, and Yu-Feng Li. 2022. Robust model selection for PU learning under constraint. *Science China Information Sciences* (2022), in press.

[31] Yi Wei, Mei Xue, Xin Liu, and Pengxiang Xu. 2022. Data fusing and joint training for learning with noisy labels. *Frontiers of Computer Science* 16, 6 (2022), Article 166338.

[32] Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. 2021. Leveraged weighted loss for partial label learning. In *Proceedings of the 38th International Conference on Machine Learning*. Virtual Event, 11091–11100.

[33] Jing-Han Wu and Min-Ling Zhang. 2019. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK, 416–424.

[34] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. 2019. Are anchor points really indispensable in label-noise learning?. In *Advances in Neural Information Processing Systems 32*. Vancouver, BC, 6835–6846.

[35] Ming-Kun Xie and Sheng-Jun Huang. 2022. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), in press.

[36] Yan Yan and Yuhong Guo. 2020. Partial label learning with batch label correction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY, 6575–6582.

[37] Yan Yan and Yuhong Guo. 2021. Multi-level generative models for partial label learning with non-random label noise. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. Virtual Event, 3264–3270.

[38] Yao Yao, Jiehui Deng, Xiuhua Chen, Chen Gong, Jianxin Wu, and Jian Yang. 2020. Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY, 12669–12676.

[39] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. 2018. Feature-induced partial multi-label learning. In *2018 IEEE International Conference on Data Mining*. Singapore, 1398–1403.

[40] Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. 2013. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, 708–715.

[41] Fei Zhang, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Tao Qin, and Masashi Sugiyama. 2022. Exploiting class activation value for partial-label learning. In *Proceedings of the 10th International Conference on Learning Representations*. Virtual Event, in press.

[42] Min-Ling Zhang and Fei Yu. 2015. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. Buenos Aires, Argentina, 4048–4054.

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

[44] Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, 1335–1344.

[45] Yivan Zhang, Gang Niu, and Masashi Sugiyama. 2021. Learning Noise Transition Matrix from Only Noisy Labels via Total Variation Regularization. In *Proceedings of the 38th International Conference on Machine Learning*. Virtual Event, 12501–12512.

[46] Deyu Zhou, Zhikai Zhang, Min-Ling Zhang, and Yulan He. 2018. Weakly supervised POS tagging without disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing* 17, 4 (2018), Article 35.

[47] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (2017), 44–53.

[48] Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3, 1 (2009), 1–130.