Conformal Prediction for Partial Label Learning

Xiuwen Gong [1], Nitin Bisht [1], Guandong Xu[1, 2] *

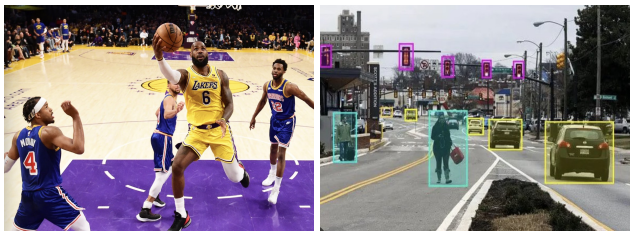
[1] University of Technology Sydney
[2] The Education University of Hong Kong
Xiuwen.Gong@uts.edu.au, Nitin.Bisht@student.uts.edu.au, guandong.xu@uts.edu.au, Gdxu@eduhk.hk

Abstract

Partial label learning (PLL) allows each instance to be annotated with a set of candidate labels, but only one is the ground-truth label. Although the state-of-the-art (SOTA) PLL models have shown competitive performance, they cannot get rid of the negative influence from the noisy false-positive labels during the training process. This leads to a large extent of uncertainty of PLL models' prediction, and it becomes unreliable to trust a PLL model's performance only by its prediction accuracy. To bridge this gap, we develop a new framework to quantify the uncertainty for PLL models with valid confidence guarantee, which is named as Conformal Prediction for Partial Label Learning (CP-PLL). This framework can be implemented on top of any PLL method to quantify their predictive confidence in terms of average prediction set size with a use-specified error rate or coverage/confidence level (i.e., probability). We prove that the coverage guarantee in PLL still holds, that is, the ground-truth label can be covered in the constructed prediction set with the user pre-defined error rate α when we use the noisy calibration data to carlibrate the PLL models, which yields to a probability interval of [1- α, 1- α + 1/n+1 + ε]. Extensive experiments are conducted on SOTA PLL methods and benchmark datasets to verify the effectiveness of the proposed framework.

Introduction

Partial label learning (PLL) (Cour, Sapp, and Taskar 2011; Chen et al. 2014; Yu and Zhang 2017) is an important weakly-supervised learning problem that allows each instance to be annotated with a set of candidate labels, with only one being the true label. PLL has attracted increasing attention in real-world applications due to its lower labeling cost, such as, automatic face recognition, automatic object detection, etc. For example, an image from NBA news contains several players (Fig.1 (a)), with each player being an instance and names extracted from the news constituting a candidate label set; it is required to automatically predict the name of each player in automatic face naming (Su et al. 2018). For another example, a street image contains many objects (Fig.1 (b)), with each object cropped out being an instance, and all labels annotated online constituting a candidate label set; it is required to automatically annotate the true label for each object in automatic object detection (Chen et al. 2020).



(a) Automatic face naming (b) Automatic image annotation

Figure 1: (a) Image of NBA players from sportnews. Each face has a true name in the candidate label set: {LeBron James, Karl Malone, Carmelo Anthony, Steve Kerr, Malik Monk, Austin Reeves, Stephen Curry, Jonathan Kuminga, Russell Westbrook, Klay Thompson, and Andrew Wiggins}. (b) A street image of objects. Each object has a true label in the candidate label set: {car, person, lights, tree}.

PLL research has achieved a big progress by evolving from conventional PLL models (linear/kernel-based) to deep PLL models (DNN-based). However, it remains a fundamental problem that PLL models cannot get rid of the negative influence from the noisy false-positive labels in the candidate label set during training process, which leads to a large extent of uncertainty of a PLL model's prediction. It becomes unreliable to trust a PLL model's performance only by its prediction accuracy. Instead, we should ask for valid confidence in its prediction. How to provide a valid confidence guarantee for the PLL models? Ideally, we expect the valid confidence to be independent of the PLL models, or PLL models without being trained to predict confidence. However, little research has been done to explore this. To bridge this gap, we are motivated to seek for new evaluation mechanisms to quantify the uncertainty for PLL models, which is able to offer the guaranteed confidence.

Inspired by conformal prediction (CP) (Vovk, Gammerman, and Shafer 2005), a powerful (e.g., model-agnostic, distribution-free) statistical tool, we provide a novel insight into PLL with guaranteed confidence in terms of coverage guarantee for the constructed prediction sets to include the

*Corresponding Author.

ground-truth label with a user-specified confidence level or error rate (i.e., probability). However, conformal prediction cannot be applied to PLL directly as it is impossible to get access to the ground-truth label in the held-out calibration dataset; that is, the labels in calibration dataset are false positive labels instead of clean labels, which poses a major challenge to apply CP for quantifying the uncertainty of PLL models.

In this paper, we pioneer the application of conformal prediction to partial label learning and propose a novel PLL framework called Conformal Prediction for Partial Label Learning (CP-PLL). We take the SOTA pre-trained PLL models as black-box, and only require that they can output the predicted labels given the probabilities of models' output (e.g., the softmax output in deep neural networks). We use the held-out PLL datasets as the calibartion datasets, and choose the label with the largest probability in the candidate label set generated from the black box as the calibartion label. We then follow the standard procedures of conformal prediction to construct the set predictors. However, it remains a doubt whether the set predictors constructed by the noisy calibartion dataset still meet the coverage guarantee as that in supervised learning setting. To verify this, we provide a theoretical analysis by analyzing the probability of the ground-truth label being contained in the constructed prediction set. Our findings show that the conformal coverage guarantee in PLL still holds. That is, the ground-truth label can be covered in the constructed prediction set with the user pre-defined error rate α or confidence level $1-\alpha$ when we use the noisy calibration data to carlibrate the PLL models, which yields to a probability interval of $[1-\alpha, 1-\alpha+\frac{1}{n+1}+\epsilon]$. We use the average prediction set size to evaluate the uncertainty (i.e., performance) of PLL models. The proposed framework (CP-PLL) is implemented on top of SOTA PLL models and benchmark datasets to verify the effectiveness. Our main contributions are summarized as follows:

- we propose a novel framework, called Conformal Prediction for Partial Label Learning (CP-PLL), to quantify the uncertainty of PLL models with confidence guarantee;
- bridge the gap between conformal prediction and PLL by calibrating models with noisy labeled calibration data;
- provide a theoretical analysis to prove the validity of prediction set, which is guaranteed to cover the ground-truth label with a user-specifed error rate/confidence level (i.e., probability), and the coverage guarantee yields to a probability interval of $[1-\alpha, 1-\alpha+\frac{1}{n+1}+\epsilon]$;
- and conduct extensive experiments to demonstrate the validity of our proposed framework (CP-PLL), namely, quantifying the uncertainty of SOTA PLL models in terms of average set size that includes the ground-truth label on benchmark datasets with user-defined confidence level.

Related Work

Partial Label Learning. Partial label learning (PLL), also known as ambiguous-label learning (Hüllermeier and Beringer 2006; Zeng et al. 2013) or superset-label learning (Gong et al. 2018), is a weakly supervised learning problem (Zhou 2017; Liu and Tsang 2015, 2017; Liu, Tsang, and Müller 2017), which differs from (semi-)supervised learning (Berthelot et al. 2019; Liu et al. 2019; Mao et al. 2022, 2020). In PLL, each instance has a collection of candidate labels, only one of which is the ground-truth label while the others are false positive labels, resulting in ambiguity while training classification models. Conventional disambiguation methods for PLL can be broadly divided into two categories (Lyu et al. 2021; Zhou, He, and Gu 2017): disambiguation by candidate label average methods, or disambiguation by ground-truth label identification methods. For the average-based methods (Cour, Sapp, and Taskar 2011; Hüllermeier and Beringer 2006; Zhang and Yu 2015), all candidate labels of each instance are treated equally as the ground-truth label, and the prediction is made by averaging the modeling outputs. For the identification-based methods (Liu and Dietterich 2012; Yu and Zhang 2017; Chai, Tsang, and Chen 2020), the ground-truth label is regarded as a latent variable and identified through iterative refining procedures. In recent years, deep learning-based PLL methods are widely developped due to their remarkable training ability on large-scale datasets. For example, PRODEN (Lv et al. 2020) proposed a progressive identification method named for approximately minimizing the proposed risk estimator, which updates the model and the identification of true labels in a seamless manner. VALEN (Xu et al. 2021) assumes the candidate labels are instance-dependent and recovers the latent label distributions by a Bayesian parametrization model. PICO (Wang et al. 2022b) proposed a contrastive learning-based method by embedding class prototype-based label disambiguation strategy. Recently, long-tailed PLL has become a hot topic due to the consideration of long-tailed data distribution in real-world scenarios. For example, SOLAR (Wang et al. 2022a) alleviates the pseudo label bias towards head classes by constraining the pseudo labels to satisfythe estimated class distribution priors. RECORDS (Hong et al. 2023) proposes a dynamic rebalancing auxiliary strategy, which dynamically recovers the class distribution priors by maintaining prototypes, and performs logit adjustment on the output of model. CWE (Jia et al. 2024) constructs a head classifier for dominant classes to keep performance and a tail classifier to improve the performance of tail classes and apply a classifier weight estimation module to automatically estimate and allocate the weights for the head classifier and tail classifier. Although deep PLL and long-tailed PLL methods have shown their competitive performance, they are still negatively and heavily impacted by the noisy false-positive labels during training process. This is because DNNs are largely dependent on the precisely labeled data to guarantee effectiveness of training, and are over-confident on any noisy fed example, which leads to the severe uncertainty of the models' prediction. It is in urgent need to have an evaluation mechanism to quantify uncertainty of the existing PLL models with confidence guarantee.

Conformal Prediction. Conformal prediction (Vovk, Gammerman, and Shafer 2005; Angelopoulos and Bates 2021), also known as conformal inference, is powerful statistical tool for quantifying models' uncertainty by generating prediction sets that are guaranteed to contain the ground-truth label with a user-specified coverage probability. Due to its

distribution-free and model-agnostic properties, conformal prediction has been widely applied to improve reliability of machine learning models by quantifying their uncertainty in real-world applications, such as, computer vision (Angelopoulos et al. 2021), natural language processing (Fisch et al. 2021), medical imaging analysis (Lu, Angelopoulos, and Pomerantz 2022), LLMs (Quach et al. 2023), etc. We use the prediction set size to quanitify a model's uncertainty. For example, if a PLL model is more uncertain, the larger size of the prediction set will be. Conformal prediction post-processes the models to construct prediction sets based on a held-out calibration dataset with coverage guarantee as follows:

$$\mathbb{P}(Y_{test} \in \mathcal{C}_\alpha(X_{test})) \geq 1 - \alpha \tag{1}$$

where (X_{test}, Y_{test}) is a new testing point, which is only assumed to have the same distribution with the calibration dataset (exchangebility); α is a user-specified error rate, and $1 - \alpha$ is the coverage or confidence level, indicating the probability of the ground-truth label being included in the prediction set. $\mathcal{C}$ is the set predictor constructed on the held-out calibration dataset.

Consequently, the paradigm of conformal prediction can also be built on top of any pre-trained partial label learning models by taking them as 'black-box' to improve their reliability. We will introduce it in the next section.

The Proposed Approach

To start with, we present notations used in this paper. Let $\{(x_i, Y_i)\}_{i=1}^n$ denote the partial label dataset drawn i.i.d. n times from some unkown distribution P. For each example, we have an instance $x_i \in \mathbb{R}^d$ with d features and a corresponding candidate label set $Y_i \subseteq \mathcal{Y} = \{1, \cdots, K\}$. Let y_i denote the ground-truth label of x_i, which is known residing in the candidate label set Y_i, i.e., $y_i \in Y_i$. Let $\hat{y}_i$ denote the calibration label, which is generated by any pre-trained prediction model with the maximum value from the candidate label set Y_i. Let $X, Y, \hat{Y}$ denote the random variables corresponding to the instance, the ground-truth label, and the calibration label respectively, and $(X, Y, \hat{Y})$ is assumed to be drawn from their joint distribuition $\mathbb{P}$. Moreover, we use $f_y(x) = f(Y = y|X = x)$ to denote any pre-trained prediction model, which ranks the classes in a descending order when performing predictions.

Problem Setup

In this subsection, we employ conformal prediction to formalize our framework, Conformal Prediction for Partial Label Learning (CP-PLL).

One challenge of applying conformal prediction to the PLL is that the ground-truth label is unkown, so the original partial label dataset cannot be used to carlibrate the prediction models. Thus, we create the PLL calibration dataset $\{(x_i, \hat{y}_i)\}_{i=1}^n$ from the samples $\{(x_i, Y_i)\}_{i=1}^n$ with the following strategy:

$$\hat{y}_i = argmax_{y \in Y_i} f_y(x_i) \tag{2}$$

We consider this strategy in that top-1 prediction of a pre-trained model f may correspond to the ground-truth label

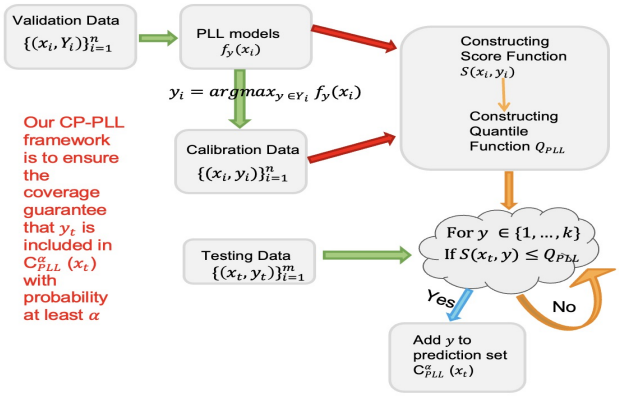


Figure 2: The Proposed CP-PLL Framework.

in the candiate label set most of the time. Although this calibration dataset is noisy, it guarantees the optimal choice of calibration label by considering highest possibility of the ground-truth label. Given the PLL calibartion dataset and any testing instance x_t, our goal is to contruct a set predictor $\mathcal{C}_{PLL}^{\alpha}(x_t)$ that covers the ground-truth label y_t with a high probability, which can be formulated as:

$$\mathbb{P}(y_t \in \mathcal{C}_{PLL}^{\alpha}(x_t)) \geq 1 - \alpha \tag{3}$$

Here, α refers to a user pre-defined error rate; $1 - \alpha$ means coverage or confidence level of the ground-truth label being covered in the prediction set.

In PLL, we cannot get access to the clean ground-truth label Y, and can only get the noisy calibration label $\hat{Y}$. As a result, we can define a score function based on the noisy calibration dataset for partial label learning following the adaptive prediction sets score strategy (Romano, Sesia, and Candès 2020) below:

$$S(x_i, \hat{y}_i) = \sum_{y \in \mathcal{Y}} f_y(x_i) \mathbb{I}\{f_y(x_i) \geq f_{\hat{y}}(x_i)\} \tag{4}$$

Based on the above PLL score function, we further define the quantile function for partial label learning following the general conformal prediction recipe (Angelopoulos and Bates 2021) below:

$$Q_{PLL} = Quantile\Big(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}, \{S(x_i, \hat{y}_i)\}\Big) \tag{5}$$

To this end, we can construct the set predictor $\mathcal{C}_{PLL}^{\alpha}(X)$ with the following strategy:

$$\mathcal{C}_{PLL}^{\alpha}(x_t) = \{y \in \mathcal{Y} : S(x_t, y) \leq Q_{PLL}\} \tag{6}$$

The overall CP-PLL framework is illustrated in Figure 2, and complete procedures of constructing the prediction sets are summarized in Algorithm 1.

Theoretical Analysis

The coverage guarantee of conformal prediction in multiclass classification has been well-studied, however, it has been remaining unexplored in partial label learning setting. In this

Algorithm 1: CP-PLL Algorithm

Goal: Constructing PLL set predictor $\mathcal{C}^{\alpha}_{PLL}(X)$.
Input: PLL calibration dataset $\{(x_i, \hat{y}_i)\}_{i=1}^{n}$, pre-trained model f, a testing instance x_t;

Output: the prediction set $\mathcal{C}^{\alpha}_{PLL}(x_t)$.

1: Compute the partial label score function $S_{PLL}(x_i, \hat{y}_i)$ given Eq. (4);
2: Compute the PLL quantile function Q_{PLL} given Eq. (5);
3: Generate the prediction set $\mathcal{C}^{\alpha}_{PLL}(x_t)$ given Eq. (6).

subsection, we provide a theoretical analysis on the coverage guarantee in PLL setting, that is, whether Eq. (3) holds.

In the problem setup, it is known that conformal prediction utilizes a calibration dataset $\{(x_i, \hat{y}_i)\}_{i=1}^{n}$ and a pre-trained model $f_y(x)$ to build a set predictor for a new testing instance x_t. Specifically, the pre-trained model is used to formulate a scoring function which is intentionally designed to produce large values when the model is uncertainty and small values when the model is confident about its prediction. In multi-class, the held-out calibration dataset $\{(x_i, y_i)\}_{i=1}^{n}$ is clean, and when the clean quantile function Q_{clean} is calculated with the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the clean calibration scores $\{S(x_i, y_i)\}_{i=1}^{n}$, then the constructed set predictor $\mathcal{C}^{\alpha}_{clean}(X)$ does cover the true label with a probability of at least $1-\alpha$. Moreover, Angelopoulos and Bates (2021) provides both the lower and upper bound of the conformal coverage guarantees for multiclass setting:

$$1-\alpha \leq \mathbb{P}(y_t \in \mathcal{C}^{\alpha}_{clean}(x_t)) \leq 1-\alpha+\frac{1}{n+1} \tag{7}$$

When the calibration dataset is noisy in PLL, can we also provide the conformal coverage guarantee of whether the ground-truth label of a tesing point is covered by the prediction set in PLL (i.e., Eq. (3) holds)? To answer this question, we start from the candidate label generation process. Existing PLL works (Lv et al. 2020; Xu et al. 2021; Wang et al. 2022b) are commited to simulating the real-world scenarios by generating candidate labels with flipping strategies. In this paper, we follow the random flipping strategy with each label being flipped to a false positive label with probability ϵ and the flipping model can be formulated as follows:

$$g(y) = \begin{cases} \hat{y}, & \epsilon \\ y, & 1-\epsilon \end{cases} \quad \text{where } \hat{y} \in \mathcal{Y} = \{1, \cdots, K\} \setminus y \tag{8}$$

To this end, we can derive the conformal coverage guarantees for partial label learning setting with the following theorem:

Theorem 0.1. *Let $\mathcal{C}^{\alpha}_{clean}(x_t)$ and $\mathcal{C}^{\alpha}_{PLL}(x_t)$ denote the prediction sets contructed on the clean calibration dataset and the partial calibration dataset, respectively. Let $p_k = \mathbb{P}(Y = k|X = x)$ and $f_k = \mathbb{P}(\hat{Y} = k|X = x)$ denote the clean and parital prediction models that rank the classes both in a descending order, where $k \in \mathcal{Y}$. Let $S(X, Y)$ and $S(X, \hat{Y})$ denote the score functions calculated based on ground-truth label and noisy calibration label respectively. Then we can get the conformal coverage guarantee for PLL as follows:*

$$1-\alpha \leq \mathbb{P}(y_t \in \mathcal{C}^{\alpha}_{PLL}(x_t)) \leq 1-\alpha+\frac{1}{n+1}+\epsilon \tag{9}$$

Proof. Given the flipping model in Eq. (8), we can get

$$\begin{aligned}
&f_k \\
=&\mathbb{P}(\hat{Y}=k|X=x) \\
=&\mathbb{P}(\hat{Y}=k, Y=k|X=x)+\mathbb{P}(\hat{Y}=k, Y\neq k|X=x) \\
=&\frac{\mathbb{P}(\hat{Y}=k, Y=k, X=x)}{\mathbb{P}(Y=k, X=x)} \cdot \frac{\mathbb{P}(Y=k, X=x)}{\mathbb{P}(X=x)} \\
&+\frac{\mathbb{P}(\hat{Y}=k, Y\neq k, X=x)}{\mathbb{P}(Y\neq k, X=x)} \cdot \frac{\mathbb{P}(Y\neq k, X=x)}{\mathbb{P}(X=x)} \\
=&\mathbb{P}(\hat{Y}=k|Y=k, X=x) \cdot \mathbb{P}(Y=k|X=x) \\
&+\mathbb{P}(\hat{Y}=k|Y\neq k, X=x) \cdot \mathbb{P}(Y\neq k|X=x) \\
=&(1-\epsilon)p_k+\frac{\epsilon}{K-1}(1-p_k) \\
=&(1-\frac{\epsilon K}{K-1})p_k+\frac{\epsilon}{K-1}
\end{aligned} \tag{10}$$

Thereby, for any testing instance x_t, we have

$$\begin{aligned}
&\mathbb{P}(S(x_t, \hat{Y}) \leq Q_{PLL}|X=x_t) \\
&=\sum_{k\in\mathcal{Y}} f_k \mathbb{I}\{S(x_t, k) \leq Q_{PLL}\} \\
&=\sum_{k:S(x_t,k)\leq Q_{PLL}} f_k \\
&=\sum_{k=1}^{k^*} f_k \\
&=\sum_{k=1}^{k^*} \Big((1-\frac{\epsilon K}{K-1})p_k+\frac{\epsilon}{K-1}\Big) \\
&=\sum_{k=1}^{k^*} p_k+\frac{\epsilon}{K-1}(k^*-K\sum_{k=1}^{k^*} p_k)
\end{aligned} \tag{11}$$

Since the item $\sum_{k=1}^{k^*} p_k$ can be formulated as

$$\begin{aligned}
&\sum_{k=1}^{k^*} p_k \\
&=\sum_{k:S(x_t,k)\leq Q_{PLL}} p_k \\
&=\sum_{k\in\mathcal{Y}} p_k \mathbb{I}\{S(x_t, k) \leq Q_{PLL}\} \\
&=\mathbb{P}(S(x_t, Y) \leq Q_{PLL}|X=x_t)
\end{aligned} \tag{12}$$

Eq. (11) can be further derived as

$$\begin{aligned}
&\mathbb{P}(S(x_t, \hat{Y}) \leq Q_{PLL}|X=x_t) = \\
&\mathbb{P}(S(x_t, Y) \leq Q_{PLL}|X=x_t)+\frac{\epsilon}{K-1}(k^*-K\sum_{k=1}^{k^*} p_k)
\end{aligned} \tag{13}$$

In addition, we have $1 - K \leq k^* - K\sum_{k=1}^{k^*} p_k \leq 0$ (this is beause $k^* \geq 1$, $\sum_{k=1}^{k^*} p_k \leq \frac{k^*}{K} \leq 1$).

As a result, we have

$$\begin{aligned} &\mathbb{P}(S(x_t, \hat{Y}) \leq Q_{PLL} | X = x_t) \\ &\leq \mathbb{P}(S(x_t, Y) \leq Q_{PLL} | X = x_t) \\ &\leq \mathbb{P}(S(x_t, \hat{Y}) \leq Q_{PLL} | X = x_t) + \epsilon \end{aligned} \tag{14}$$

Given the standard conformal arguement in Vovk, Gammerman, and Shafer (2005), it can be easily learned that $1 - \alpha \leq \mathbb{P}(S(x_t, \hat{Y}) \leq Q_{PLL} | X = x_t) \leq 1 - \alpha + \frac{1}{n+1}$. Thus, we have

$$1-\alpha \leq \mathbb{P}\big(S(x_t, Y) \leq Q_{PLL} | X = x_t\big) \leq 1-\alpha+\frac{1}{n+1}+\epsilon \tag{15}$$

This implies that $y_t \in \mathcal{C}^{\alpha}_{PLL}(x_t)$ with probability of at least $1 - \alpha$ and at most $1 - \alpha + \frac{1}{n+1} + \epsilon$, which concludes the proof. □

Remark. From the above theoretical analysis, we can see that the conformal coverage guarantee in PLL still holds. That is, the ground-truth label can be covered in the constructed prediction set when we use the noisy calibration labels to carlibrate the PLL models, which yields to a probability interval of $[1 - \alpha, 1 - \alpha + \frac{1}{n+1} + \epsilon]$.

Experiments

In this section, we empirically test the validity of the proposed framework CP-PLL in quantifying the uncertainty (i.e., predictive confidence) of partial label learning models by implementing it on top of the state-of-the-art PLL models and various datasets in terms of average set size (the smaller the better). Code is publicly available at https://github.com/kalpiree/CP-PLL.

Datasets

We evaluate CP-PLL on various benchmark datasets, including CIFAR-10 and CIFAR-100 (Krizhevsky and Hinton 2009), and their corresponding long-tailed versions, i.e., CIFAR-10-LT, CIFAR-100-LT.

Following the uniform noise generation process in previous work (Lv et al. 2020; Wen et al. 2021), we generate the PLL datasets by flipping negative labels to false positive labels with a probability ϵ, and then form the candidate set by aggregating the $|Y| - 1$ flipped labels with the ground-truth label. Specifically, for CIFAR-10, we consider the flipping rate $\epsilon = 0.1, 0.3, 0.5, 0.7, 0.9$, while $\epsilon = 0.01, 0.03, 0.05, 0.1, 0.5$ for CIFAR-100. For the long-tailed datasets, we follow the strategies in previous work (Wang et al. 2022a; Jia et al. 2024) to generate the long-tailed versions of CIFAR-10, CIFAR-100. We use imbalance ratio ρ to denote the ratio between sample sizes of the most frequent and least frequent class, i.e., $\rho = \max n_i / \min n_i$. Long-tailed imbalance follows an exponential decay in sample sizes across different classes. Here, $n_1 = 5000$ for CIFAR10; $n_1 = 500$ for CIFAR-100; $n_C = \frac{n_1}{\rho}$, which is the number of samples of C-th class. We use different imbalance ratios to evaluate the performance, with $\rho = 50, 100, 200$ for CIFAR10-LT and $\rho = 10, 20, 50$ for CIFAR100-LT. We then generate the PLL datasets following the uniform noise generation process with $\epsilon = 0.1, 0.3, 0.5$ for CIFAR-10-LT and $\epsilon = 0.01, 0.05, 0.1$ for CIFAR-100-LT.

#Set Size / Flip Rate / Models		ϵ = 0.1	ϵ = 0.3	ϵ = 0.5	ϵ = 0.7	ϵ = 0.9
CP-PLL +	PRODEN	2.96	2.02	1.57	1.25	1.06
	PICO	2.93	2.03	1.52	1.24	1.06
	SOLAR	2.12	1.62	1.53	1.22	1.03
	RECORDS	**1.08**	**1.03**	**1.01**	**1.00**	**1.00**
	CWE	<u>1.19</u>	<u>1.13</u>	<u>1.09</u>	<u>1.01</u>	<u>1.00</u>

Table 1: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-10 dataset with varying flipping rate $\epsilon = 0.1, 0.3, 0.5, 0.7, 0.9$ under error rate $\alpha = 0.1$. Bold indicates the best results; underline indicates the second best.

#Set Size / Flip Rate / Models		ϵ = 0.01	ϵ = 0.03	ϵ = 0.05	ϵ = 0.1	ϵ = 0.5
CP-PLL +	PRODEN	23.70	16.37	12.75	7.81	1.85
	PICO	26.80	18.60	13.64	7.98	1.82
	SOLAR	25.76	16.41	12.37	7.73	1.84
	RECORDS	<u>12.03</u>	<u>3.64</u>	<u>3.32</u>	<u>2.49</u>	<u>1.60</u>
	CWE	**3.00**	**2.77**	**2.55**	**2.06**	**1.16**

Table 2: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-100 dataset with varying flipping rate $\epsilon = 0.01, 0.03, 0.05, 0.1, 0.5$ under error rate $\alpha = 0.1$. Bold indicates the best results; underline indicates the second best.

Base Models

We consider the state-of-the-art PLL methods to validate our proposed framework, including deep PLL models: PRODEN (Lv et al. 2020), PICO (Wang et al. 2022b), and long-tailed PLL models: SOLAR (Wang et al. 2022a), RECORDS (Hong et al. 2023), CWE (Jia et al. 2024). We implement our framework on top of these base models to quantify their uncertainty and evaluate their performance by the average prediction set size on the above datasets.

Implementation Details

We use 18-layer ResNet as the backbone. The mini-batch size is set to 256 and all the methods are trained using SGD with momentum of 0.9 and weight decay of 0.001 as the optimizer. The initial learning rate is set to 0.01. We train the model for 800 epochs. We split the held-out training data with 60% as the calibartion data and 40% as the testing data on all datasets. The hyper-parameters of the base PLL models were configured according to their original papers.

Experimental Results

We evaluate all models' performance in terms of average set size with varying flipping rate $\epsilon = 0.1, 0.3, 0.5, 0.7, 0.9$ for CIFAR-10 dataset, $\epsilon = 0.01, 0.03, 0.05, 0.1, 0.5$ for CIFAR-100 dataset under error rate $\alpha = 0.1$, while flipping rate

#Set Size / Models \ Parameters		$\rho = 50$			$\rho = 100$			$\rho = 200$		
		$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$
CP-PLL +	PRODEN	3.60	2.30	1.64	3.61	2.30	1.63	3.66	2.30	1.65
	SOLAR	2.55	1.89	1.55	2.70	1.97	1.54	2.81	2.05	1.56
	PICO	3.39	2.23	1.62	3.38	2.23	1.64	3.45	2.25	1.63
	RECORDS	1.61	1.21	1.08	2.09	1.53	1.16	2.38	1.61	1.31
	CWE	**1.05**	**1.04**	**1.03**	**1.33**	**1.10**	**1.05**	**1.51**	**1.20**	**1.16**

Table 3: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-10-LT dataset with varying imbalance ratio $\rho = 50, 100, 200$ and flipping rate $\epsilon = 0.1, 0.3, 0.5$ under error rate $\alpha = 0.1$. Bold indicates the best results; underline indicates the second best.

#Set Size / Models \ Parameters		$\rho = 10$			$\rho = 20$			$\rho = 50$		
		$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.1$
CP-PLL +	PRODEN	33.79	14.23	8.25	33.49	14.32	8.11	33.82	14.58	8.26
	SOLAR	24.24	12.27	7.48	25.32	12.58	7.49	25.40	12.56	7.49
	PICO	25.65	12.60	7.75	24.59	12.79	7.63	24.09	13.07	7.62
	RECORDS	**5.25**	**4.11**	**3.58**	**7.05**	**5.12**	4.26	15.31	**9.30**	**5.28**
	CWE	9.80	6.50	4.55	11.18	5.57	**1.00**	**14.41**	9.66	6.47

Table 4: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-100-LT dataset with varying imbalance ratio $\rho = 10, 20, 50$ and flipping rate $\epsilon = 0.01, 0.05, 0.1$ under error rate $\alpha = 0.1$. Bold indicates the best results; underline indicates the second best.

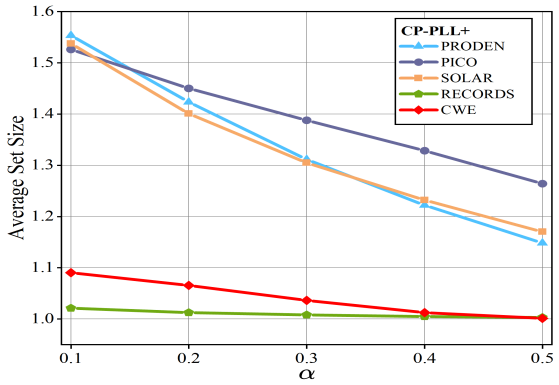


Figure 3: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-10 dataset with varying error rate $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ under flipping rate $\epsilon = 0.5$.

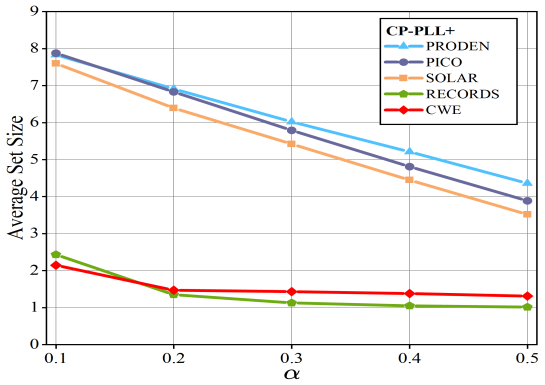


Figure 4: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-100 dataset with varying error rate $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ under flipping rate $\epsilon = 0.1$.

$\epsilon = 0.1, 0.3, 0.5$ and imbalance ratio $\rho = 50, 100, 200$ for CIFAR10-LT dataset, $\epsilon = 0.01, 0.05, 0.1$ and $\rho = 10, 20, 50$ for CIFAR-100-LT dataset. Results are shown in Table 1 to Table 4 respectively. From these results, we can make the following interesting observations:

- All base models after applying the CP-PLL framework can produce valid prediction sets that meet the coverage guarantee of 90% (i.e., $1 - \alpha$) on all datasets. This indicates that CP-PLL is data- and model-agnostic, which can be applied on top of any PLL models and datasets.
- The average set sizes of all models show a decreasing trend when the flipping rate ϵ increases on all datasets. This aligns with our expectation that larger flipping rate leads to larger candidate label set, which leads to a larger range of candidate labels where the calibration label is chosen from by Eq. (2); as a result, label whose value is larger than the value of calibartion label in the whole label space will become fewer; thus the score calculated by Eq. (4) will have smaller value, so is the quantile value calculated by Eq. (5), which results in smaller set size.
- On all datasets, RECORDS and CWE are the best two models with smaller average set size while the remaining models have comparable performance with much larger average set size. For example, when CWE per-

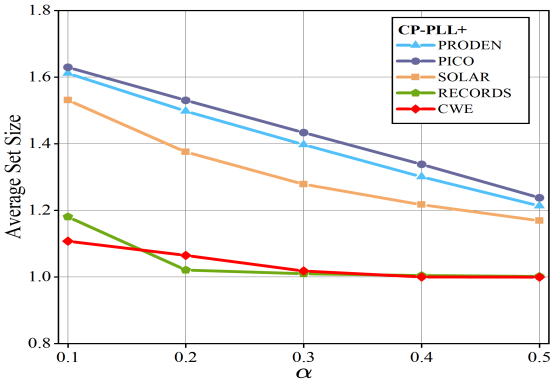


Figure 5: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-10-LT dataset with varying error rate $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ under flipping rate $\epsilon = 0.5$ and imbalance ratio $\rho = 100$.

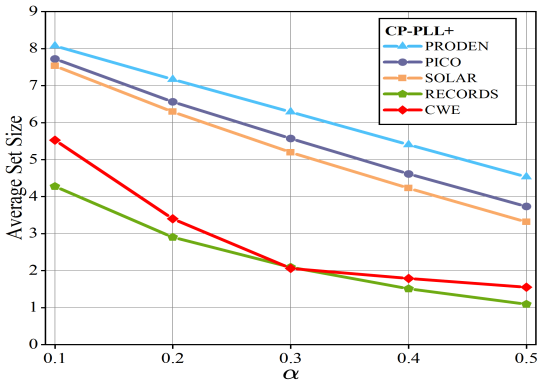


Figure 6: Average Set Size of all PLL models after being applied our conformal framework on CIFAR-100-LT dataset with varying error rate $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ under flipping rate $\epsilon = 0.1$ and imbalance ratio $\rho = 20$.

forms the best on CIFAR-100 and CIFAR-10-LT datasets, RECORDS comes the second best and others performs the worst; when RECORDS performs the best on CIFAR-10 and CIFAR-100-LT datasets, CWE comes the second best and the rest come as the last. This is because the defined score function relies on the output of PLL models, and the model that can make more accurate prediction will also have the better performance in our framework in terms of average set size. In this way, our framework provides an effective method to evaluate the confidence of PLL models in prediction.

- When comparing the performance on long-tailed CIFAR-10-LT and CIFAR-100-LT datasets in Fig. 3 and Fig. 4 with that on partially-labeled CIFAR-10 and CIFAR-100 datasets in Fig. 1 and Fig. 2, we can see that the average set sizes of all models are larger on long-tailed datasets. This is because imbalanced partially-labeled data leads to more ambiguity and worse prediction accuracy, which results in the larger set size to meet the high probability (i.e., 90%) to cover the ground-truth label.
- In addition, the average set size of each model doesn't vary much with different imbalance ratio ρ, which indicates the robustness of the proposed framework for long-tailed datasets, and its potentiality in real-world applications.

Parameter Analysis

We empirically analyze the influence of error rate α on base models' performance in terms of average set size by varying error rate $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ under flipping rate $\epsilon = 0.5$ for CIFAR-10, and $\epsilon = 0.1$ for CIFAR-100, while under flipping rate $\epsilon = 0.5$, imbalance ratio $\rho = 100$ for CIFAR-10-LT, and $\epsilon = 0.1$, $\rho = 20$ for CIFAR-100-LT dataset. Results are reported in Fig. 3 to Fig. 6 respectively. From these results, we can observe that:

- The average set sizes of all models show a decreasing trend when error rate α increases on all datasets. This is because higher error rate means more tolerance for PLL model when required to produce prediction set to cover the ground-truth label; thus, smaller size of prediction sets will be automatically generated corresponding to the more tolerance.
- Moreover, RECORDS and CWE are still the best two models with smaller set size while the others come the last under different error rate α. This result is consistent with that observed in the above tables under different flipping rate ϵ and imbalance ratio ρ, which demonstrates the steady property of our framework when being applied on top of SOTA PLL models.
- It can be concluded that the proposed CP-PLL is an effective framework in generating valid prediction sets that meet the coverage guarantee, which can be used as an effective metric to evaluate the predictive confidence of PLL models in terms of average set size.

Conclusion

This paper investigates the uncertainty quantification problem of partial label learning methods caused by the noisy false-positive labels in the candidate label set. We study whether conformal prediction, a statistical tool, can be applied to quantify the uncertainty of PLL models with valid confidence (i.e., coverage guarantee), and develop a novel PLL framework called Conformal Prediction for Partial Label Learning (CP-PLL). Theoretically, we prove that the conformal coverage guarantee in PLL still holds. That is, the ground-truth label can be covered in the constructed prediction set when we use the noisy calibration labels to carlibrate the PLL models, which yields to a probability interval of $[1-\alpha, 1-\alpha+\frac{1}{n+1}+\epsilon]$. Empirically, we implement CP-PLL on top of five typical state-of-the-art PLL models and four benchmark datasets, the results of which demonstrate the validity of our proposed framework to quantify the uncertainty of PLL models by evaluating their predictive confidence in terms of average set size. This research studies the naive split CP recipe for PLL, and more advanced conformal approaches are encouraged to explore by researchers in future.

Acknowledgments

This work is partially supported by the Australian ResearchCouncil (ARC) Under Grants DP220103717 andLE220100078, and the National Natural Science Foundation of China under Grants No.62072257.

References

Angelopoulos, A. N.; and Bates, S. 2021. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *CoRR*, abs/2107.07511.

Angelopoulos, A. N.; Bates, S.; Jordan, M. I.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *ICLR*.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *NeurIPS*, 5050–5060.

Chai, J.; Tsang, I. W.; and Chen, W. 2020. Large Margin Partial Label Machine. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7): 2594–2608.

Chen, Y.; Patel, V. M.; Chellappa, R.; and Phillips, P. J. 2014. Ambiguously Labeled Learning Using Dictionaries. *IEEE Transactions on Information Forensics and Security*, 9(12): 2076–2088.

Chen, Y.; Zeng, X.; Chen, X.; and Guo, W. 2020. A survey on automatic image annotation. *Applied Intelligence*, 50(10): 3412–3428.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from Partial Labels. *Journal of Machine Learning Research*, 12: 1501–1536.

Fisch, A.; Schuster, T.; Jaakkola, T. S.; and Barzilay, R. 2021. Efficient Conformal Prediction via Cascaded Inference with Expanded Admission. In *ICLR*.

Gong, C.; Liu, T.; Tang, Y.; Yang, J.; Yang, J.; and Tao, D. 2018. A Regularization Approach for Instance-Based Superset Label Learning. *IEEE Transactions on Cybernetics*, 48(3): 967–978.

Hong, F.; Yao, J.; Zhou, Z.; Zhang, Y.; and Wang, Y. 2023. Long-Tailed Partial Label Learning via Dynamic Rebalancing. In *ICLR*.

Hüllermeier, E.; and Beringer, J. 2006. Learning from Ambiguously Labeled Examples. *Intelligent Data Analysis*, 10(5): 419–439.

Jia, Y.; Peng, X.; Wang, R.; and Zhang, M. 2024. Long-Tailed Partial Label Learning by Head Classifier and Tail Classifier Cooperation. In *AAAI*, 12857–12865.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario.

Liu, L.; and Dietterich, T. G. 2012. A Conditional Multinomial Mixture Model for Superset Label Learning. In *NeurIPS*, 557–565.

Liu, W.; and Tsang, I. W. 2015. On the Optimality of Classifier Chain for Multi-label Classification. In *NeurIPS*, 712–720.

Liu, W.; and Tsang, I. W. 2017. Making Decision Trees Feasible in Ultrahigh Feature and Label Dimensions. *The Journal of Machine Learning Research*, 18: 81:1–81:36.

Liu, W.; Tsang, I. W.; and Müller, K. 2017. An Easy-to-hard Learning Paradigm for Multiple Classes and Multiple Labels. *Journal of Machine Learning Research*, 18: 94:1–94:38.

Liu, W.; Xu, D.; Tsang, I. W.; and Zhang, W. 2019. Metric Learning for Multi-Output Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 408–422.

Lu, C.; Angelopoulos, A. N.; and Pomerantz, S. R. 2022. Improving Trustworthiness of AI Disease Severity Rating in Medical Imaging with Ordinal Conformal Prediction Sets. In *MICCAI*, 545–554.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive Identification of True Labels for Partial-Label Learning. In *ICML*, volume 119, 6500–6510.

Lyu, G.; Feng, S.; Wang, T.; Lang, C.; and Li, Y. 2021. GM-PLL: Graph Matching Based Partial Label Learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(2): 521–535.

Mao, Y.; Wang, Z.; Liu, W.; Lin, X.; and Xie, P. 2022. MetaWeighting: Learning to Weight Tasks in Multi-Task Learning. In *ACL*, 3436–3448. Association for Computational Linguistics.

Mao, Y.; Yun, S.; Liu, W.; and Du, B. 2020. Tchebycheff Procedure for Multi-task Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 4217–4226.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal Language Modeling. *CoRR*, abs/2306.10193.

Romano, Y.; Sesia, M.; and Candès, E. J. 2020. Classification with Valid and Adaptive Coverage. In *NeurIPS 2020*.

Su, X.; Zhou, H.; Draghici, V. P.; and Rätsch, M. 2018. Face naming in news images via multiple instance learning and hybrid recurrent convolutional neural network. *Journal of Electronic Imaging*, 27(03): 033–036.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer US. ISBN 0387001522.

Wang, H.; Xia, M.; Li, Y.; Mao, Y.; Feng, L.; Chen, G.; and Zhao, J. 2022a. SoLar: Sinkhorn Label Refinery for Imbalanced Partial-Label Learning. In *NeurIPS*.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022b. PiCO: Contrastive Label Disambiguation for Partial Label Learning. In *ICLR*.

Wen, H.; Cui, J.; Hang, H.; Liu, J.; Wang, Y.; and Lin, Z. 2021. Leveraged Weighted Loss for Partial Label Learning. In *ICML*, volume 139, 11091–11100. PMLR.

Xu, N.; Qiao, C.; Geng, X.; and Zhang, M. 2021. Instance-Dependent Partial Label Learning. In *NeurIPS*, 27119–27130.

Yu, F.; and Zhang, M. 2017. Maximum margin partial label learning. *Machine Learning*, 106(4): 573–593.

Zeng, Z.; Xiao, S.; Jia, K.; Chan, T.; Gao, S.; Xu, D.; and Ma, Y. 2013. Learning by Associating Ambiguously Labeled Images. In *CVPR*, 708–715.

Zhang, M.; and Yu, F. 2015. Solving the Partial Label Learning Problem: An Instance-Based Approach. In *IJCAI*, 4048–4054.

Zhou, Y.; He, J.; and Gu, H. 2017. Partial Label Learning via Gaussian Processes. *IEEE Transactions on Cybernetics*, 47(12): 4443–4450.

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53.