Label Correction of Crowdsourced Noisy Annotations with an Instance-Dependent Noise Transition Model

Hui Guo [1] Boyu Wang [1] Grace Y. Yi [1, 2]

[1]Department of Computer Science
[2]Department of Statistical and Actuarial Sciences
University of Western Ontario

NeurIPS 2023, Virtual Talk

Problem Setup

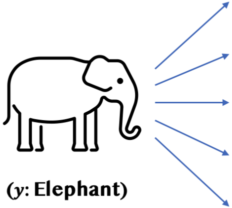
Crowdsourcing:

Each data item is labeled by multiple annotators with diverse expertise

- Noisy training data $\mathcal{D} = \{\mathbf{x}_i, \tilde{\mathrm{y}}_i^{(1)}, \ldots, \tilde{\mathrm{y}}_i^{(R)}\}_{i=1}^{N}$
  - $\mathcal{X} \subset \mathbb{R}^p$: feature space
  - $\mathcal{Y} = \{1, \ldots, K\}$: label space
  - R: number of annotators
  - $\mathbf{x}_i \in \mathcal{X}$: input data
  - $\mathrm{y}_i \in \mathcal{Y}$: unobserved true label
  - $\tilde{\mathrm{y}}_i^{(r)} \in \mathcal{Y}$: the label given by the rth annotator with $r \in \{1, \ldots, R\}$
- Goal: learn a classifier which correctly labels the new input data

Data (x):

Crowdsourced Noisy Labels ($\tilde{\mathrm{y}}$):



Noisy Label Generation Process

- **Assumption:** the R annotators *independently* label the instances
- **Noisy label generation model:**

$$\mathbb{P}(\tilde{y}^{(1)}, \ldots, \tilde{y}^{(R)}|\mathbf{x}) = \prod_{r=1}^{R} \mathbb{P}(\tilde{y}^{(r)}|\mathbf{x}) = \prod_{r=1}^{R} \sum_{k \in \mathcal{Y}} \left\{ \mathbb{P}(\tilde{y}^{(r)}|y = k, \mathbf{x}) \; P(y = k|\mathbf{x}) \right\}$$

instance-dependent noise transition matrix for the rth annotator
$f_0^{k,r}(\mathbf{x})$: distribution of $\tilde{y}^{(r)}|\{y = k, \mathbf{x}\}$, modeled by $f_{\boldsymbol{\theta}}^{k,r}(\mathbf{x})$

base model $h(\cdot; \boldsymbol{\vartheta})$
(true label predictor)

- **Issues about instance-dependent transition matrices:**
  - Most available methods require the *instance independent* assumption: $\mathbb{P}(\tilde{y}^{(r)}|y = k, \mathbf{x}) = \mathbb{P}(\tilde{y}^{(r)}|y = k)$; however, the instance dependent assumption is more realistic
  - Modeling the instance-dependent transition matrix is challenging and remains relatively less explored
  - Theoretical characterization of the distance of the noise model and the true transition matrix remains absent in the literature

Western UNIVERSITY · CANADA

Approximate the Instance-Dependent Noise Transition Matrices

- **Bayesian network:**
  - Deploy a set of (δ-pseudo) anchor points $\overline{\mathcal{D}}_0$ learned from noisy training data
    - An instance $\mathbf{x}$ is defined to be an (δ-pseudo) anchor point of class k if $\mathbb{P}(y = k|\mathbf{x}) = 1$ ($\mathbb{P}(y = k|\mathbf{x}) \geq \delta$)
    - The subsample size n of $\overline{\mathcal{D}}_0$ is relatively small compared to the main sample size N
  - Employ a hierarchical spike and slab prior on the network parameters
    - Sparse Bayesian network $f_{\boldsymbol{\theta}}^{k,r}$ with $\boldsymbol{\theta} \in \Theta$
- **Posterior consistency result:**
  - The sparse noise transition model is close to the underlying true transition matrix with respect to the Hellinger distance under mild conditions

Theorem 1

Let $d(\cdot,\cdot)$ denote the Hellinger distance. Under regularity conditions, there exists a sequence of constants $\{\epsilon_n^2\}_{n=1}^{\infty}$ satisfying $\lim_{n\to\infty} \epsilon_n = 0$ and $\lim_{n\to\infty} n\epsilon_n^2 = \infty$, such that for any $k \in \{1, \ldots, K\}$ and $r \in \{1, \ldots, R\}$, with probability tending to 1, the posterior measure satisfies

$$\Pi\left\{\boldsymbol{\theta} \in \Theta : d(f_{\boldsymbol{\theta}}^{(k,r)}, f_0^{(k,r)}) > M_n\epsilon_n | \overline{\mathcal{D}}_0\right\} \to 0 \text{ as any } M_n \to \infty.$$

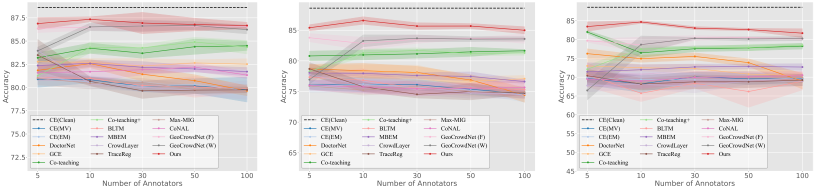
Pairwise Likelihood Ratio Test for Label Correction

- **Reformulate the label correction process:**
  - Selecting the label for the instance $\mathbf{x}_i$ from $\{g, g'\}$, is equivalent to choosing from the two competitors $\mathbb{P}(\tilde{\mathbf{y}}|\mathrm{y}=g,\mathbf{x}_i)$ and $\mathbb{P}(\tilde{\mathbf{y}}|\mathrm{y}=g',\mathbf{x}_i)$, where $1 \leq g < g' \leq K$
  - Hypothesis testing: $H_g : \tilde{\mathbf{y}}_i|\{\mathrm{y}_i,\mathbf{x}_i\} \sim \mathbb{P}(\tilde{\mathbf{y}}|\mathrm{y}=g,\mathbf{x}_i)$ versus $H_{g'} : \tilde{\mathbf{y}}_i|\{\mathrm{y}_i,\mathbf{x}_i\} \sim \mathbb{P}(\tilde{\mathbf{y}}|\mathrm{y}=g',\mathbf{x}_i)$
- **Label correction method:**
  - (Neyman-Pearson Lemma) Set the estimated label of $\mathbf{x}_i$ to be $\overline{\mathrm{y}}_i = g$ if

$$\frac{\hbar_{i,g}\prod_{r=1}^{R}\prod_{l=1}^{K}\left\{\tau_{i,gl}^{(r)}\right\}^{\mathbf{1}(\tilde{y}_i^{(r)}=l)}}{\hbar_{i,g'}\prod_{r=1}^{R}\prod_{l=1}^{K}\left\{\tau_{i,g'l}^{(r)}\right\}^{\mathbf{1}(\tilde{y}_i^{(r)}=l)}} > \Omega \text{ for any } g' \neq g$$

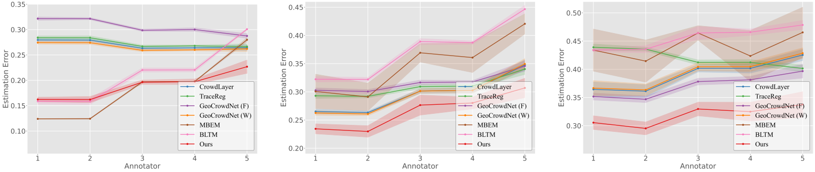
  - $\hbar_{ig}$: class prior for the ground truth label for the ith task for $g \in \{1,\ldots,K\}$
    $\Longrightarrow$ the predictions of base classifiers
  - $\boldsymbol{\tau}_{i,kl}^{(r)}$: the lth element of $f_{\boldsymbol{\theta}}^{(k,r)}(\mathbf{x}_i)$ for $k,l \in \{1,\ldots,K\}$ and $r \in \{1,\ldots,R\}$
    $\Longrightarrow$ the maximum a posteriori (MAP) estimate
  - Ω: pre-specified threshold
- **Theorem 2:**
  - Information-theoretic bounds on the Bayes error

Empirical Results on CIFAR10 with Varying Number of Annotators



(a) IDN-LOW (b) IDN-MID (c) IDN-HIGH

Figure 1: Average accuracy.



(a) IDN-LOW (b) IDN-MID (c) IDN-HIGH

Figure 2: Average estimation error of noise transition matrices.

With varying number of annotators, the proposed method

- achieves **the highest average test accuracy**;
- leads to **smaller estimation error** in most of the cases, especially when the noise rate is high.

Western UNIVERSITY · CANADA

Summary

In this work,

- We explore the challenging problem of learning with instance-dependent crowdsourced noisy annotations
- We formulate the annotator-specific noise transition matrix in the **Bayesian framework**
- We **theoretically characterize the closeness** of the proposed sparse Bayesian model and the underlying annotator confusions with respect to the Hellinger distance
- We develop a novel **label correction algorithm** by aggregating the noisy annotations using the pairwise likelihood ratio test, and identify **information-theoretic bounds** on the Bayes error
- **Numerical experiments** demonstrate that the proposed method outperforms the competing methods

Thank You