ICML | 2021

Discriminative Complementary-Label Learning with Weighted Loss

Yi Gao, Min-Ling Zhang

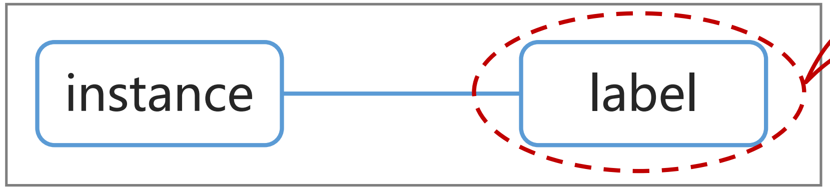
School of Cyber Science and Engineering,
MOE Key Laboratory of Computer Network and Information Integration,
School of Computer Science and Engineering,
Southeast University, China

Outline

ICML | 2021

- **Introduction**
- The Proposed Approach
  - The Discriminative Model
  - The Weighted Loss
- Experiments
- Conclusion

Ordinary Multi-class Classification



Strong supervision information

Strong supervision information

- Sufficient labelled training data
- No ambiguous or incorrect labeling

Annotating is costly and time-consuming!

Weakly supervised learning is frequently encountered in real-world!

Examples for Weakly Supervised Learning

Weakly supervised learning: learning model from data with weak supervision information

semi-supervised learning

instance — label

instance — ???

insufficient labeling

noisy label learning

instance — label

instance — label

noisy labeling

Complementary-label learning (CLL)

instance — label

complementary labeling

Complementary-label Learning

Ordinary multi-class classification: an instance $\boldsymbol{x}$ with a ground-truth label y

CLL: An instance $\boldsymbol{x}$ with a complementary label $\bar{y}$, which is the label that the instance does not belong to

Ground-truth label	Raccoon	Monkey	Marmot
			
Complementary label	not "Monkey"	not "Marmot"	not "Raccoon"

The Problem

Goal: learning a multi-class classifier

Previous work in CLL:

Aiming at modeling the generative relationship between y, i.e., $P(y \mid \boldsymbol{x})$, and $\bar{y}$, i.e., $P(\bar{y} \mid \boldsymbol{x})$

- Unbiased generation: complementary labels are uniformly selected from one of labels other than the ground-truth one
- Biased generation: complementary labels are non-uniformly selected from one of labels other than the ground-truth one, which depends on transition probability $P(\bar{y} \mid y)$

The Problem

Problems :

- Unbiased generation: suffer from overfitting problem, as the empirical gradients may deviate from true gradients during the optimization procedure (Chou et al., 2020)

- Biased generation: need extra conditions, such as the availability of a set of anchor instances , to estimate transition probability

Can we learn from complementary labels without assumption on the generation process?

Our Work

A discriminative solution to directly model $P(\bar{y} \mid \boldsymbol{x})$ from the output of trained classifiers without extra generation assumptions

Contributions

- Deriving a risk estimator with guaranteed estimation error bound at $\mathcal{O}(1/\sqrt{n})$ convergence rate
- Introducing weighted loss to enforce predictive gap between potential ground-truth label and complementary label

Outline

ICML | 2021

- Introduction
- **The Proposed Approach**
  - **The Discriminative Model**
  - **The Weighted Loss**
- Experiments
- Conclusion

Notation

Settings

- $\mathcal{X}$: d-dimensional feature space $\mathbb{R}^d$
- $\mathcal{Y}$: label space with c class labels $\{1, \ldots, c\}$

Data

- $\mathcal{D}$: a set of n training examples $\{(X, Y)\}^n$
- $\bar{\mathcal{D}}$: a set of complementarily labeled training examples $\{(X, \bar{Y})\}^n$

  $X \in \mathcal{X}$ is a d-dimensional feature vector; $Y \in \mathcal{Y}$ is the ground-truth label of X

  $\bar{Y} \in \{\mathcal{Y} \setminus \{Y\}\}$ is the complementary label of X

Outputs

- $\boldsymbol{f}$: multi-class classifier for ordinary multi-class classification
- $\bar{\boldsymbol{f}}$: multi-class classifier for complementary label classification

The Discriminative Model

The ordinary model

For ordinary multi-class classification,

- The predictive probability of the ground-truth label approaches one
- The predictive probability of the complementary label approaches zero

The discriminative model

The prediction probability of complementary label as $\bar{\boldsymbol{f}}(X) = 1 - \boldsymbol{f}(X)$

the complementary loss $\bar{\ell}$

$$\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}}) = \ell(\bar{\boldsymbol{f}}(X), e^{\bar{Y}}) = \ell(1 - \boldsymbol{f}(X), e^{\bar{Y}})$$

where ℓ is the loss function, $e^{\bar{Y}} \in \{0,1\}^c$ is a one-hot vector for label $\bar{Y}$.

Estimation Error Bound

Estimation error bound illustrates that the difference between the risk of the empirical classifier learned by empirical risk minimization and the risk of the optimal CLL classifier can be bounded.

Assumption

The loss function $\ell(\cdot,\cdot)$ satisfies $\ell(1 - f_k(X), 1 - e_k^Y) = \ell(f_k(X), e_k^Y)$.

where e_k^Y and f_k are the k-th element of e^Y and $\boldsymbol{f}$ respectively

Such an assumption holds for some commonly used loss functions, such as MSE (Mean Squared Error) loss and MAE (Mean Absolute Error) loss.

Estimation Error Bound

Theorem

For any $\delta > 0$, with probability at least $1-\delta$,

$$\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}(\bar{\boldsymbol{f}}^*) \leq 4c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M\sqrt{\frac{2log(2/\delta)}{n}},$$

where $\bar{\boldsymbol{f}}_n^* = \mathrm{argmin}_{\boldsymbol{f}\in\mathcal{F}} \bar{R}_n(\boldsymbol{f})$, $\bar{R}_n(\boldsymbol{f})$ is the empirical risk estimator for CLL, $\bar{\boldsymbol{f}}^* = \mathrm{argmin}_{\boldsymbol{f}\in\mathcal{F}} \bar{R}(\boldsymbol{f})$, $\bar{R}(\boldsymbol{f})$ is the expectation risk estimator for CLL.

For all parametric models with a bounded norm, as $n \to \infty$, $\bar{R}(\bar{\boldsymbol{f}}_n^*) \to \bar{R}(\bar{\boldsymbol{f}}^*)$. The theorem shows that the proposed risk estimator exists an estimation error bound and convergence rate is $\mathcal{O}(1/\sqrt{n})$.

Tip

The fewer number of labels, the more effective our proposed CLL method

The Weighted Loss

Motivation

- The estimated posterior probability → measure the prediction uncertainty
- Increasing uncertainty could lead to a deteriorated prediction performance

Our solution

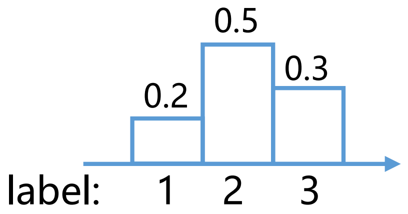
The highly confident predictions during learning can be used to update the model

Introducing a weighted loss term to form the weighted loss:

$$\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}}) = w\ell(1 - \boldsymbol{f}(X), e^{\bar{Y}})$$

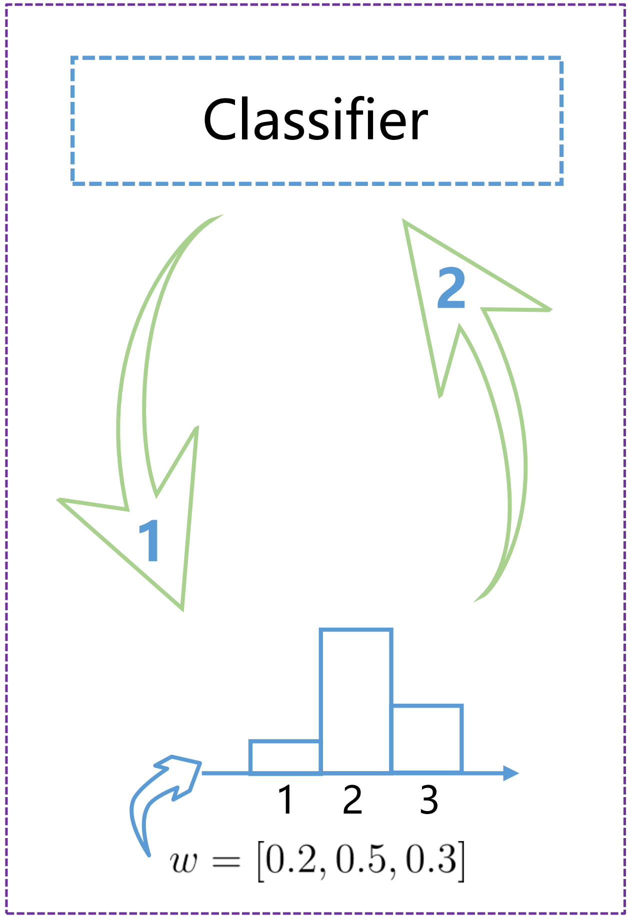
The Weighted Loss – A Case

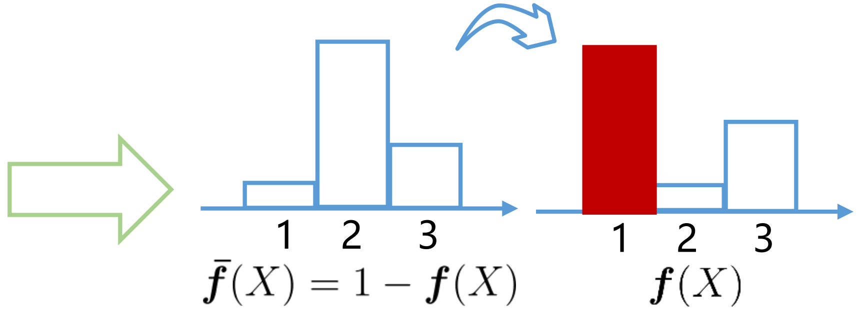
Suppose a three-category CLL task, i.e., c = 3



Predicted probability

$\bar{\boldsymbol{f}}(X) = [0.2, 0.5, 0.3]$





2 Update the model

1 Update the weighted loss term

$$w^k = \frac{1 - f_k(X)}{\sum_{j=1}^{c}(1 - f_j(X))}$$

$$\bar{\boldsymbol{f}} = [0.2, 0.5, 0.3]$$

Yi Gao, Min-Ling Zhang. ICML, 2021.

The Weighted Loss

Targeted loss

Add the weighted loss and the unweighted loss together

$$\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}}) = \sum_{k=1}^{c}(1+\lambda w^k)\ell(1-f_k(X), e_k^{\bar{Y}})$$

The final empirical risk estimator

$$\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}(1+\lambda w_i^k)\ell(1-f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i})$$

The tradeoff parameter

$\lambda = 1$

Outline

ICML | 2021

- Introduction
- The Proposed Approach
  - The Discriminative Model
  - The Weighted Loss
- **Experiments**
- Conclusion

Datasets

01

MNIST (Lecun et al., 1998): a handwritten digits dataset that consists of 10 classes

02

Fashion-MNIST (Fashion) (Xiao et al., 2017): coming from standardized images of fashion items, including 10 classes

03

Kuzushiji-MNIST (Kuzushiji) (Clanuwat et al., 2018): deriving from Kuzushiji which includes 10 classes

Base Models & Baselines

Base models

- linear model
- MLP model (d-500-c)

Baselines

- Pairwise Comparison (PC) with sigmoid loss (Ishida et al., 2017)
- Forward loss correction (Forward) (Yu et al., 2018)
- Gradient Ascent (GA) (Ishida et al., 2019)
- Non-Negative loss (NN) (Ishida et al., 2019)

Comparison on Unbiased Complementary Labels

- Complementary-label generation: unbiased (uniform distribution)

Table 1. Test accuracy (mean±std) out of 10 trials (in %), where data with unbiased complementary labels is used to train. The best performance on each data set is shown in boldface.

Dataset	Model	PC	Forward	GA	NN	L-UW	L-W
MNIST	linear	82.31±0.72	**90.42±0.17**	83.23±0.43	84.56±0.31	89.98±0.20	90.22±0.11
	MLP	84.04±0.55	91.93±0.25	**92.49±0.25**	89.99±0.42	92.45±0.24	92.08±0.28
Fashion	linear	75.29±0.83	81.14±0.20	77.41±0.30	78.32±0.31	81.79±0.22	**82.04±0.21**
	MLP	77.55±0.39	82.31±0.24	81.62±0.19	80.29±0.47	83.15±0.20	**83.40±0.32**
Kuzushiji	linear	54.57±1.13	60.57±0.42	52.52±1.12	55.27±0.85	60.87±0.48	**61.29±0.31**
	MLP	59.32±0.59	65.59±0.54	**69.56±0.53**	65.44±0.51	65.17±1.43	66.98±1.63

The Win/Loss statistics

Baselines	PC	Forward	GA	NN
L-UW	6/0	4/2	4/2	5/1
L-W	6/0	5/1	4/2	6/0

- L-UW (without weighted loss term) achieves comparable test accuracy to baselines
- L-W (with weighted loss term) shows that the weighted loss does help improve the generalization performance

Yi Gao, Min-Ling Zhang. ICML, 2021.

Comparison on Biased Complementary Labels

- Complementary-label generation: biased, where different sets denote the different biased degree of complementary labels

Set 1						
Baselines		PC	Forward	GA	L-UW	L-W
MNIST	linear	**19.66±0.28**	19.54±0.58	9.86±0.15	18.23±0.17	18.57±0.55
	MLP	19.34±0.69	20.44±0.15	9.80±0.00	19.46±0.34	**21.13±2.06**
Set 2						
Baselines		PC	Forward	GA	L-UW	L-W
MNIST	linear	19.69±0.63	20.31±0.10	10.19±0.16	23.55±2.05	**23.67±0.74**
	MLP	22.59±2.32	20.44±0.20	10.09±0.00	23.35±0.66	**26.76±2.00**
Set 3						
Baselines		PC	Forward	GA	L-UW	L-W
MNIST	linear	72.22±1.43	78.53±4.41	78.55±0.80	**81.16±0.12**	79.72±0.27
	MLP	84.46±0.23	80.67±5.34	85.13±0.10	84.98±0.10	**85.91±0.11**

The Win/Loss statistics on three datasets

Baselines	PC	Forward	GA
L-UW	14/4	13/5	15/3
L-W	15/3	16/2	15/3

- The test accuracy of all baselines has improved as the biased degree of complementary labels decreasing
- L-W gets comparable test accuracy to Forward when the biased transition matrix with no additional information is available for Forward

Outline

ICML | 2021

- Introduction
- The Proposed Approach
  - The Discriminative Model
  - The Weighted Loss
- Experiments
- **Conclusion**

Conclusion

- We propose the discriminative model that directly model $P(\bar{y} \mid \boldsymbol{x})$ from the predictive probability of learned classifiers

- A risk estimator with guaranteed estimation error bound based on discriminative model is proposed for CLL

- The weighted loss is further introduced to the classification risk to yield the empirical risk

Thanks ! Q & A