西安交通大學
XI'AN JIAOTONG UNIVERSITY

Learning with Biased Complementary Labels

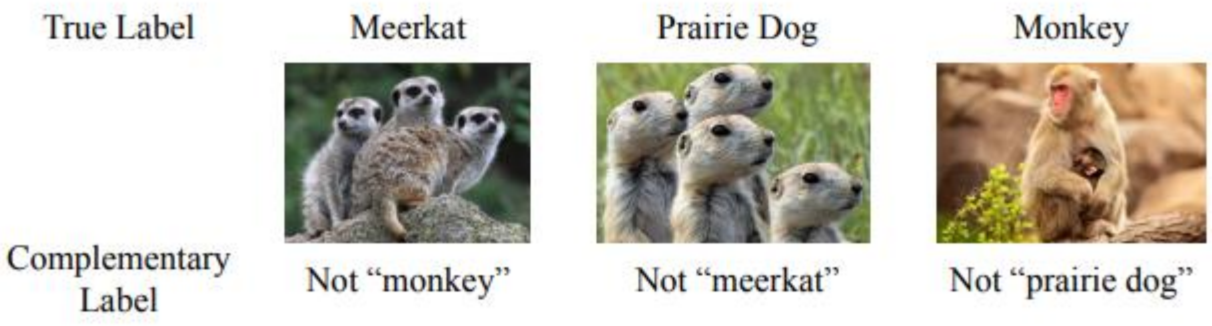
Xiyu Yu* Tongliang Liu* Mingming Gong†‡ Dacheng Tao*

Yuxuan Wu

2021.10.28

Complementary-Label Learning

- Ordinary supervised learning: training samples labeled with true labels.
- Complementary-label learning: weakly supervised learning, training samples labeled with complementary labels which indicate the categories that the samples do not belong to.



A comparison between true labels (top) and complementary labels (bottom).[1]

Background

西安交通大學 XI'AN JIAOTONG UNIVERSITY

- First modeling the annotation of complementary labels via transition probabilities $P\left(\bar{Y}=i|Y=j\right), i \neq j \in \{1,\ldots,c\}$, where c is the number of classes.

- Previous methods implicitly assume that transition probabilities are identical, $P\left(\bar{Y}=i|Y=j\right)=\frac{1}{c-1}, i \neq j \in \{1,\ldots,c\}$

- Labels are often annotated by humans, and humans are biased toward their own experience. Therefore the transition probabilities will be different.

Contributions of the proposed framework

1. It estimates transition probabilities with no bias.

2. It provides a general method to modify traditional loss functions and extends standard deep neural network classifiers to learn with biased complementary labels.

3. It theoretically ensures that the classifier learned with complementary labels converges to the optimal one learned with true labels.

4. Comprehensive experiments on several benchmark datasets validate the superiority of our method to current state-of-the-art methods.

Problem Setup

In multi-class classification, let $\mathcal{X} \in \mathbb{R}^d$ be the feature space and $\mathcal{Y} = [c]$ be the label space, where *d* is the feature space dimension.

For each example $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, a complementary label $\bar{y}$ is selected from the complement set $\mathcal{Y} \backslash \{y\}$ .We assign a probability for each $\bar{y} \in \mathcal{Y} \backslash \{y\}$ to indicate how likely it can be selected, i.e., $P\left(\bar{Y} = \bar{y} | Y = y, X = \boldsymbol{x}\right)$

Assuming that complementray label is independent of feature *X* conditioned on true label *Y*.

$$P\left(\bar{Y} = \bar{y} | Y = y, X = \boldsymbol{x}\right) = P\left(\bar{Y} = \bar{y} | Y = y\right)$$

Problem Setup

Summarizing all the probabilities into a transition matrix $\boldsymbol{Q} \in \mathbb{R}^{c\times c}$

$$Q_{ij} = P\left(\bar{Y} = j | Y = i\right) \text{ and } Q_{ii} = 0, \forall i,j \in [c]$$

If complementary labels are uniformly selected from the complement set

$$\forall i,j \in [c] \text{ and } i \neq j,\ Q_{ij} = \frac{1}{c-1}$$

Learning with True Labels:

$$f(X) = \underset{i \in [c]}{\operatorname{argmax}}\, g_i(X)$$

where $g: \mathcal{X} \rightarrow \mathbb{R}^c$ and $g_i(X)$ is the estimate of $P(Y = i|X)$

The expected risk is defined as: $R(f) = \mathbb{E}_{(X,Y)\sim P_{XY}}[\ell(f(X), Y)]$
The optimal classifier is the one that minimizes the expected risk; that is,

$$f^* = \underset{f \in \mathcal{F}}{\arg\min}\, R(f)$$

We then approximate $R(f)$ by using its empirical counterpart:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)$$

$$f_n = \underset{f \in \mathcal{F}}{\arg\min}\, R_n(f)$$

Learning with Complementary Labels:

we can only learn a mapping $q: \mathcal{X} \to \mathbb{R}^c$ that tries to predict conditional probabilities $P(\bar{Y}|X)$

Therefore, we need to modify these loss functions such that the classifier learned with biased complementary labels can converge to the optimal one learned with true labels.

$$\bar{R}(f) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\bar{\ell}\left(f(X),\bar{Y}\right)\right] \text{ and } \bar{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\bar{\ell}(f(\boldsymbol{x}_i),\bar{y}_i)$$

We hope that the modified loss function $\bar{\ell}$ can ensure that $\bar{f}_n \to f^*$

Learning with Complementary Labels:

Recall that in transition matrix Q:

$$Q_{ij} = P\left(\bar{Y} = j | Y = i\right) \text{ and } Q_{ii} = 0, \forall i,j \in [c]$$

$$\begin{aligned} P\left(\bar{Y} = j | X\right) &= \sum_{i \neq j} P\left(\bar{Y} = j, Y = i | X\right) \\ &= \sum_{i \neq j} P\left(\bar{Y} = j | Y = i, X\right) P(Y = i | X) \\ &= \sum_{i \neq j} P\left(\bar{Y} = j | Y = i\right) P(Y = i | X) \end{aligned}$$

Intuitively, if $q_i(X)$ tries to predict the probability $P\left(\bar{Y} = i | X\right), \forall i \in [c]$, then $Q^{-\mathrm{T}} q$ can predict the probability $P(Y|X)$

Learning with Complementary Labels:

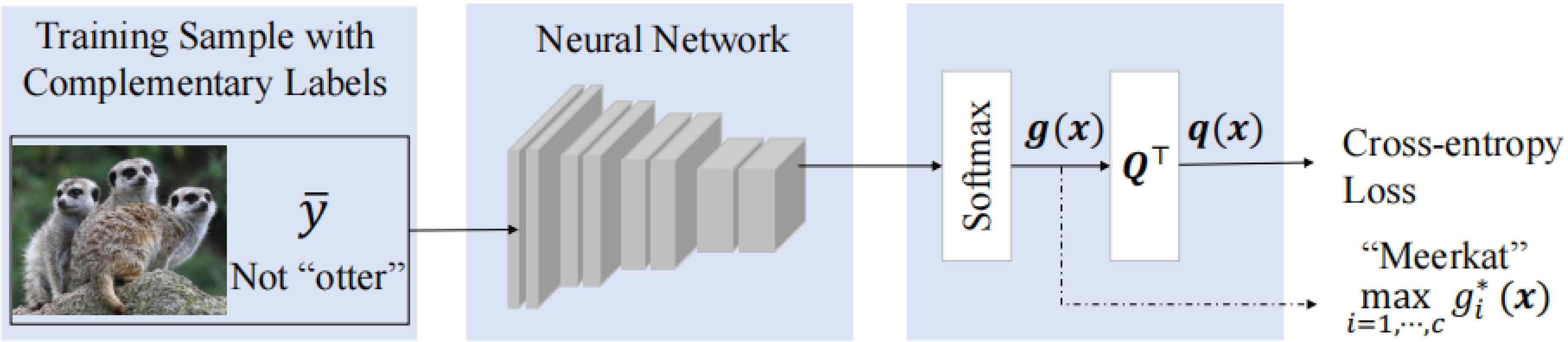
To enable end-to-end learning rather than transferring after training

$$\boldsymbol{q}(X) = \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{g}(X)$$

Then, the modified loss function $\bar{\ell}$ is

$$\bar{\ell}\left(f(X), \bar{Y}\right) = \ell\left(\boldsymbol{q}(X), \bar{Y}\right)$$

Forward loss correction



Estimating Q

Assumption 2 (Anchor Set Condition). *For each class y, there exists an anchor set $\mathcal{S}_{\mathbf{x}|y} \subset \mathcal{X}$ such that $P(Y = y|X = \mathbf{x}) = 1$ and $P(Y = y'|X = \mathbf{x}) = 0, \forall y' \in \mathcal{Y} \setminus \{y\}, \mathbf{x} \in \mathcal{S}_{\mathbf{x}|y}$.*

Here, $\mathcal{S}_{\mathbf{x}|y}$ is a subset of features in class y. Given several observations in $\mathcal{S}_{\mathbf{x}|y}, y \in [c]$, we are ready to estimate the transition matrix $\mathbf{Q}$.

$$P(\bar{Y} = \bar{y}|X) = \sum_{y' \neq \bar{y}} P(\bar{Y} = \bar{y}|Y = y')P(Y = y'|X). \tag{14}$$

Suppose $\mathbf{x} \in \mathcal{S}_{\mathbf{x}|y}$, then $P(Y = y|X = \mathbf{x}) = 1$ and $P(Y = y'|X = \mathbf{x}) = 0, \forall y' \in \mathcal{Y} \setminus \{y\}$. We have

$$P(\bar{Y} = \bar{y}|X = \mathbf{x}) = P(\bar{Y} = \bar{y}|Y = y). \tag{15}$$

Experiments

Classification accuracy on USPS and UCI datasets:

Dataset	c	d	#train	#test	PC/S	PL	ML	LM (ours)
WAVEFORM1	$1 \sim 3$	21	1226	398	**85.8 (0.5)**	**85.7 (0.9)**	79.3 (4.8)	**85.1 (0.6)**
WAVEFORM2	$1 \sim 3$	40	1227	408	**84.7 (1.3)**	**84.6 (0.8)**	74.9 (5.2)	**85.5 (1.1)**
SATIMAGE	$1 \sim 7$	36	415	211	**68.7 (5.4)**	60.7 (3.7)	33.6 (6.2)	**69.3 (3.6)**
PENDIGITS	$1 \sim 5$	16	719	336	**87.0 (2.9)**	76.2 (3.3)	44.7 (9.6)	**92.7 (3.7)**
	$6 \sim 10$		719	335	78.4 (4.6)	71.1 (3.3)	38.4 (9.6)	**85.8 (1.3)**
	even #		719	336	**90.8** (2.4)	76.8 (1.6)	43.8 (5.1)	**90.0 (1.0)**
	odd #		719	335	76.0 (5.4)	67.4 (2.6)	40.2 (8.0)	**86.5 (0.5)**
	$1 \sim 10$		719	335	38.0 (4.3)	33.2 (3.8)	16.1 (4.6)	**62.8 (5.6)**
DRIVE	$1 \sim 5$	48	3955	1326	**89.1 (4.0)**	77.7 (1.5)	31.1 (3.5)	**93.3 (4.6)**
	$6 \sim 10$		3923	1313	88.8 (1.8)	78.5 (2.6)	30.4 (7.2)	**92.8 (0.9)**
	even #		3925	1283	**81.8 (3.4)**	63.9 (1.8)	29.7 (6.3)	**84.3 (0.7)**
	odd #		3939	1278	**85.4 (4.2)**	74.9 (3.2)	27.6 (5.8)	**85.9 (2.1)**
	$1 \sim 10$		3925	1269	40.8 (4.3)	32.0 (4.1)	12.7 (3.1)	**75.1 (3.2)**
LETTER	$1 \sim 5$	16	565	171	**79.7 (5.4)**	75.1 (4.4)	28.3 (10.4)	**84.3 (1.5)**
	$6 \sim 10$		550	178	76.2 (6.2)	66.8 (2.5)	34.0 (6.9)	**84.4 (1.0)**
	$11 \sim 15$		556	177	78.3 (4.1)	67.4 (3.4)	28.6 (5.0)	**88.3 (1.9)**
	$16 \sim 20$		550	184	77.2 (3.2)	68.4 (2.1)	32.7 (6.4)	**85.2 (0.7)**
	$21 \sim 25$		585	167	80.4 (4.2)	75.1 (1.9)	32.0 (5.7)	**82.5 (1.0)**
	$1 \sim 25$		550	167	**5.1 (2.1)**	**5.0 (1.0)**	**5.2 (1.1)**	**7.0 (3.6)**
USPS	$1 \sim 5$	256	652	166	**79.1 (3.1)**	70.3 (3.2)	44.4 (8.9)	**86.4 (4.5)**
	$6 \sim 10$		542	147	69.5 (6.5)	66.1 (2.4)	37.3 (8.8)	**88.1 (2.7)**
	even #		556	147	67.4 (5.4)	66.2 (2.3)	35.7 (6.6)	**79.5 (5.4)**
	odd #		542	147	77.5 (4.5)	69.3 (3.1)	36.6 (7.5)	**86.3 (3.1)**
	$1 \sim 10$		542	127	**30.7 (4.4)**	26.0 (3.5)	13.3 (5.4)	**37.2 (5.4)**

PL: a partial label method
ML: a multi-label method
PC/S: the pairwise-comparison formulation with sigmoid loss

Experiments Classification accuracy on MNIST datasets:

XI'AN JIAOTONG UNIVERSITY

Method	Uniform	Without0	With0
TL	99.12	99.12	99.12
PC/S	86.59 ± 3.99	76.03 ± 3.34	29.12 ± 1.94
LM/T	97.18 ± 0.45	97.65 ± 0.15	98.63 ± 0.05
LM/E	96.33 ± 0.31	97.04 ± 0.31	98.61 ± 0.05

"TL" denotes the result of learning with true labels.
"LM/T" and "LM/E" refer to our method with the true Q and the estimated one, respectively.

(1) for each image in class y, the complementary label is uniformly selected from $\mathcal{Y} \setminus \{y\}$ ("uniform");
(2) the complementary label is non-uniformly selected, but each label in $\mathcal{Y} \setminus \{y\}$ has non-zero probability to be selected ("without0");
(3) The complementary label is non-uniformly selected from a small subset of $\mathcal{Y} \setminus \{y\}$ ("with0").

	0	1	2	3	4	5	6	7	8	9
0	0.0	0.5	0.0	0.0	0.0	0.3	0.0	0.2	0.0	0.0
1	0.2	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.3	0.0
2	0.0	0.5	0.0	0.0	0.3	0.0	0.0	0.0	0.2	0.0
3	0.3	0.5	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
4	0.2	0.0	0.0	0.3	0.0	0.5	0.0	0.0	0.0	0.0
5	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
6	0.0	0.2	0.0	0.0	0.3	0.0	0.0	0.0	0.5	0.0
7	0.3	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.5	0.0
8	0.0	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.3
9	0.0	0.0	0.5	0.3	0.0	0.2	0.0	0.0	0.0	0.0

	0	1	2	3	4	5	6	7	8	9
0	0.001	0.512	0.001	0.001	0.001	0.310	0.000	0.171	0.002	0.001
1	0.191	0.001	0.000	0.003	0.001	0.001	0.519	0.000	0.284	0.000
2	0.002	0.489	0.000	0.000	0.305	0.000	0.001	0.000	0.202	0.000
3	0.307	0.476	0.001	0.001	0.211	0.000	0.000	0.000	0.002	0.002
4	0.201	0.001	0.003	0.261	0.000	0.533	0.000	0.000	0.001	0.000
5	0.296	0.500	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.201
6	0.000	0.162	0.000	0.000	0.319	0.000	0.000	0.000	0.518	0.000
7	0.277	0.000	0.000	0.164	0.000	0.000	0.000	0.000	0.558	0.000
8	0.005	0.462	0.226	0.000	0.001	0.000	0.000	0.000	0.001	0.303
9	0.013	0.003	0.505	0.280	0.000	0.195	0.000	0.000	0.003	0.001

	0	1	2	3	4	5	6	7	8	9
0	0.000	0.200	0.100	0.100	0.200	0.100	0.033	0.200	0.033	0.033
1	0.200	0.000	0.100	0.100	0.100	0.033	0.200	0.033	0.200	0.033
2	0.100	0.200	0.000	0.100	0.200	0.033	0.033	0.033	0.200	0.100
3	0.200	0.200	0.033	0.000	0.200	0.033	0.100	0.100	0.033	0.100
4	0.100	0.100	0.033	0.200	0.000	0.200	0.100	0.033	0.200	0.033
5	0.200	0.200	0.033	0.033	0.200	0.000	0.033	0.100	0.100	0.100
6	0.100	0.200	0.033	0.100	0.200	0.033	0.000	0.200	0.033	0.100
7	0.200	0.033	0.033	0.100	0.200	0.100	0.100	0.000	0.200	0.033
8	0.033	0.200	0.100	0.033	0.100	0.200	0.033	0.200	0.000	0.100
9	0.033	0.100	0.100	0.100	0.033	0.200	0.200	0.033	0.200	0.000

	0	1	2	3	4	5	6	7	8	9
0	0.006	0.202	0.093	0.100	0.197	0.113	0.027	0.211	0.022	0.028
1	0.210	0.006	0.099	0.091	0.108	0.029	0.202	0.038	0.183	0.035
2	0.102	0.203	0.005	0.125	0.184	0.029	0.031	0.027	0.199	0.095
3	0.198	0.183	0.037	0.007	0.220	0.030	0.098	0.093	0.030	0.103
4	0.121	0.090	0.037	0.198	0.013	0.201	0.101	0.034	0.184	0.020
5	0.219	0.213	0.019	0.023	0.205	0.006	0.037	0.091	0.089	0.097
6	0.081	0.204	0.025	0.102	0.253	0.031	0.006	0.166	0.042	0.089
7	0.224	0.043	0.030	0.103	0.256	0.073	0.092	0.005	0.146	0.029
8	0.021	0.194	0.095	0.021	0.077	0.242	0.030	0.207	0.005	0.108
9	0.045	0.102	0.088	0.096	0.019	0.205	0.217	0.036	0.183	0.009

Experiments

Classification accuracy on CIFAR10 datasets:

Method	Uniform	Without0	With0
TL	90.78	90.78	90.78
PC/S	41.19 ± 0.04	42.97 ± 3.00	18.12 ± 1.45
LM/T	73.38 ± 1.06	78.80 ± 0.45	85.32 ± 1.11
LM/E	42.96 ± 0.76	70.56 ± 0.34	84.60 ± 0.14

Classification accuracy on CIFAR100 and Tiny ImageNet under the setting “with0”:

Method	CIFAR100	Tiny ImageNet
TL	69.55	63.26
PC/S	8.95 ± 1.47	N/A
LM/T	62.84 ± 0.30	52.71 ± 0.71
LM/E	60.27 ± 0.28	49.70 ± 0.78

Conclusions

- Addressing the problem of **learning with biased complementary labels.**
- Specifically, considering the setting that the transition probabilities vary and most of them are zeros.
- Devising an effective method to estimate the transition matrix given a small amount of data in the **anchor set**.
- Based on the transition matrix, **modifying traditional loss functions** such that learning with complementary labels can theoretically converge to the optimal classifier learned from examples with true labels.
- Comprehensive experiments on a wide range of datasets verify that the proposed method is superior to the current state-of-the-art methods.