Learning with Biased Complementary Labels

Xiyu Yu[1], Tongliang Liu[1], Mingming Gong[2,3], and Dacheng Tao[1]

[1] UBTECH Sydney AI Centre, SIT, FEIT, The University of Sydney
{xiyu0300@uni., tongliang.liu@, dacheng.tao@}sydney.edu.au
[2] Department of Philosophy, Carnegie Mellon University
[3] Department of Biomedical Informatics, University of Pittsburgh
mig73@pitt.edu

ECCV 2018

Introduction

Sometimes, precisely labeling an example is costly…

True Label

Meerkat

Prairie Dog

Not dog

Monkey

Labeling the image requires:

- expert knowledge
- time-consuming work

…

Compared to "which class it belongs"

it is much more easy to see "which class it does not belong"

Review: Learning from Complementary Label

- Problem Formulation

$x \longleftrightarrow \cancel{y} \quad \bar{y} \qquad \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n} \xrightarrow{?}$ learn a multi-class classifier

- A basic assumption $\quad \bar{p}(\boldsymbol{x}, \bar{y}) = \frac{1}{K-1} \sum_{y \neq \bar{y}} p(\boldsymbol{x}, y)$. ?

All $p(x, y)$ for $y \neq \bar{y}$ equally contribute to $\bar{p}(x, \bar{y})$

$$\bar{\eta}_k(x) := \mathbb{P}(\bar{Y} = k | X = x)$$
$$\eta_k(x) := \mathbb{P}(Y = k | X = x)$$

$\Rightarrow \quad \bar{\boldsymbol{\eta}}(x) = \boldsymbol{T}\boldsymbol{\eta}(x)$

T=

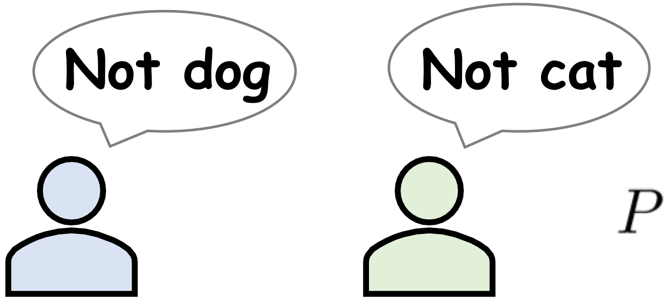
0	1/2	1/2
1/2	0	1/2
1/2	1/2	0

- The unbiased estimate: $R(f) = (K-1)\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\bar{\mathcal{L}}(f(\boldsymbol{x}), \bar{y})\right] - M_1 + M_2$

Motivation

Revisiting the example…

Prairie Dog



Labeler A Labeler B

$$P(\bar{Y} = i | Y = j) \neq 1/(c\text{-}1)$$
$$i \neq j$$

Labelers provide complementary labels based on both the **observation** and their **experience**. → **Biasedness**

How to capture such biasedness? $P(\bar{Y} = i | Y = j)$

Method

$P(Y|X)$? $P(\bar{Y} = i|Y = j)$ $P(\bar{Y}|X)$ Unknown but can be estimated

$$P(\bar{Y} = j|X) = \sum_{i \neq j} P(\bar{Y} = j, Y = i|X)$$

$\mathbf{q}(X)$

$$= \sum_{i \neq j} P(\bar{Y} = j|Y = i, X) P(Y = i|X)$$

$$= \sum_{i \neq j} P(\bar{Y} = j|Y = i) P(Y = i|X)$$

$Q_{ij} = P(\bar{Y} = j|Y = i)$ matrix $\mathbf{Q}$ $\mathbf{g}(X)$

Assumption:

$$P(\bar{Y} = \bar{y}|X = \mathbf{x}, Y = y)$$
$$= P(\bar{Y} = \bar{y}|Y = y)$$

$$\mathbf{q}(X) = \mathbf{Q}^\top \mathbf{g}(X)$$

$$\bar{\ell}(f(X), \bar{Y}) = \ell(\mathbf{q}(X), \bar{Y})$$

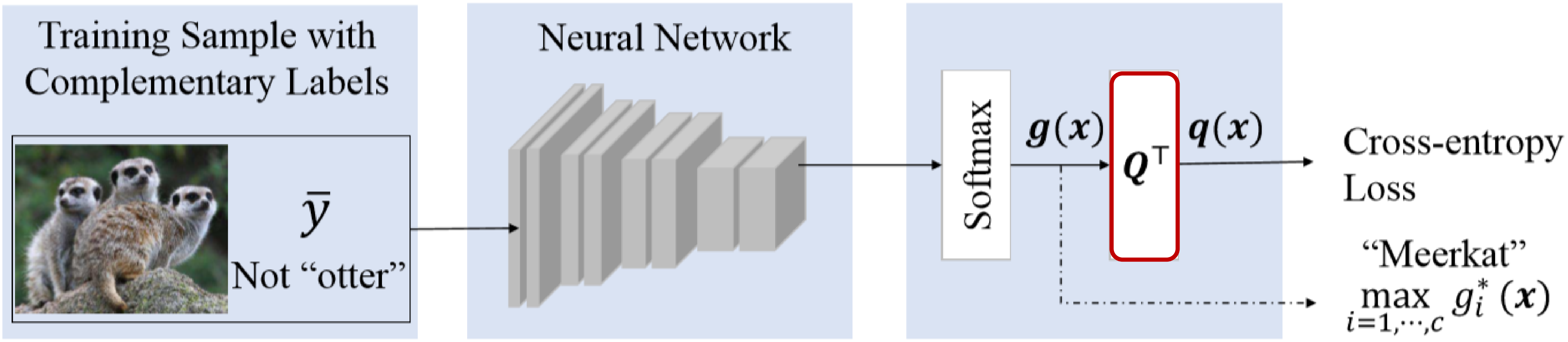
Prairie Dog

Not dog

Method

$$\mathbf{q}(X) = \mathbf{Q}^{\top}\mathbf{g}(X) \quad \Rightarrow \quad \mathbf{Q}^{-\top}\mathbf{q}$$

Learning with DNN



$$\mathbf{q}^* \quad \Rightarrow \quad q_i^*(X) = P(\bar{Y} = i | X), \forall i \in [c] \quad \Rightarrow \quad \mathbf{g}^* \text{ and } f^*$$

Method

Estimate Q

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

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* $P(Y=y'|X=\mathbf{x})=0$, $\forall y'\in\mathcal{Y}\setminus\{y\},\mathbf{x}\in\mathcal{S}_{\mathbf{x}|y}$.

Find some $\mathbf{x}\in\mathcal{S}_{\mathbf{x}|y}$ $\quad P(Y=y|X=\mathbf{x})=1 \quad P(Y=y'|X=\mathbf{x})=0$

⬇

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

Theoretical Analysis

Optimality of the classifier

Assumption 1 *By minimizing the expected risk $R(f)$, the optimal mapping $\mathbf{g}^*$ satisfies $g_i^*(X) = P(Y = i|X), \forall i \in [c]$.*

Theorem 1 *Suppose that $\mathbf{Q}$ is invertible and Assumption 1 is satisfied, then the minimizer $\bar{f}^*$ of $\bar{R}(f)$ is also the minimizer f^* of $R(f)$; that is, $\bar{f}^* = f^*$.*

$$\bar{f}^* = \arg\min_{f \in \mathcal{F}} \bar{R}(f) \qquad \mathbf{q}(X) = \mathbf{Q}^\top \mathbf{g}(X)$$

Convergence analysis

Corollary 1 *Suppose $\bar{\pi}_i = P(\bar{Y} = i)$ is given. Let the loss function be upper bounded by M. Then, for any $\delta > 0$, with the probability $1 - c\delta$, we have*

$$\bar{R}(\bar{f}_n) - \bar{R}(\bar{f}^*) \leq \sum_{i=1}^{c} \left(4c\bar{\pi}_i \mathfrak{R}_{n_i}(\mathcal{H}) + 2\bar{\pi}_i M \sqrt{\frac{\log 1/\delta}{2n_i}} \right). \tag{13}$$

$$\bar{f}_n = \arg\min_{f \in \mathcal{F}} \bar{R}_n(f)$$

Experiments

Table 1. Classification accuracy on USPS and UCI datasets: the means and standard deviations of classification accuracy over 20 trials in percentages are reported. "#train" is the number of training and validation examples in each class. "#test" is the number of test examples in each class.

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)**

Experiments

Table 2. Classification accuracy on MNIST: the means and standard deviations of classification accuracy over five trials in percentages are reported. "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.

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

Table 3. Classification accuracy on CIFAR10: the means and standard deviations of classification accuracy over five trials in percentages are reported. "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.

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

Thanks