Estimating Noise Transition Matrix with Label Correlations for Noisy Multi-Label Learning

Shikun Li[1,2] Xiaobo Xia[3] Hansong Zhang[1,2] Yibing Zhan[4]
Shiming Ge[1,2] Tongliang Liu[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences

[2]School of Cyber Security, University of Chinese Academy of Sciences

[3]Trustworthy Machine Learning Lab, The University of Sydney [4]JD Explore Academy

NeurIPS 2022

Outline

Introduction

Figure: Example of image with noisy multi-labels. (C. O. Pene *et al.*)

Clean Data $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{Y} = \left\{Y^1, Y^2, \ldots, Y^q\right\} \in \{0,1\}^q$

Noisy Data $(\boldsymbol{X}, \bar{\boldsymbol{Y}})$, where $\bar{\boldsymbol{Y}} = \left\{\bar{Y}^1, \bar{Y}^2, \ldots, \bar{Y}^q\right\} \in \{0,1\}^q$

Transition Matrix $T^j_{ik}(\boldsymbol{x}) = P\left(\bar{Y}^j = k \mid Y^j = i, \boldsymbol{X} = \boldsymbol{x}\right), j = 1, 2, ..., q$

Introduction

The transition matrix bridges the class posterior probabilities for noisy and clean data, i.e. $P(\bar{Y} = k \mid \boldsymbol{X} = \boldsymbol{x}) = \sum_{i=0}^{1} T_{ik} P(Y = i \mid \boldsymbol{X} = \boldsymbol{x})$. Thus, it has been exploited to achieve many **statistically consistent algorithms** in noisy multi-class learning, which also can be applied in noisy multi-label learning. As the effectiveness of these algorithms heavily relies on estimating the transition matrix, a series of methods have been proposed to achieve estimation.

Problem:

- Most of estimation methods assume **the existence of anchor points** (T. Liu *et al.*), while the assumption is strong and hard to check.
- The methods need to **accurately fit the noisy class posterior** of anchor points, which is rather difficult in multi-label cases, due to the severe positive-negative imbalance.

Intuitive Solution:
To address the problems, we consider utilizing **label correlations among noisy multiple labels**. At a high level, we can utilize the **mismatch of label correlations** to identify the transition matrix without neither anchor points nor accurate fitting of noisy class posterior.

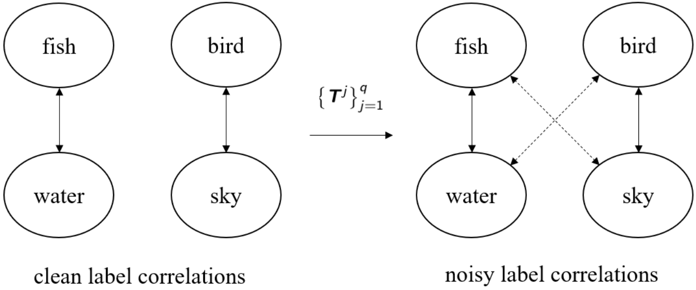

Figure: The mismatch of label correlations.

Our Method

As the instance-dependent transition matrix is non-identifiable without any additional assumption (X. Xia *et al.*), we assume that the transition matrix is class-dependent and instance-independent, i.e. $P\left(\bar{Y}^j = k \mid Y^j = i, \boldsymbol{X} = \boldsymbol{x}\right) = P\left(\bar{Y}^j = k \mid Y^j = i\right) = T^j_{ik}$.

Theorem

If $P(\bar{Y}^i \mid Y^j)$ is known, two noisy labels $\{\bar{Y}^j, \bar{Y}^i\}$ are sufficient to identify $\boldsymbol{T}^j$.

This theorem theoretically guarantees that the identifiability of the class-dependent transition matrix can be achieved by utilizing the occurrence probabilities $P(\bar{Y}^i, \bar{Y}^j)$ and $P(\bar{Y}^i \mid Y^j)$. Note that $P(\bar{Y}^i, \bar{Y}^j)$ can represent **noisy label correlations**, and $P(\bar{Y}^i \mid Y^j) = \sum_{Y^i} P(\bar{Y}^i \mid Y^i) P(Y^i \mid Y^j)$, which can imply **clean label correlations**.

At a high level, **the mismatch of label correlations** implied in these occurrence probabilities can achieve the identifiability.

Our Method

Based on the above discussions, we propose to estimate transition matrices $\left\{\boldsymbol{T}^j\right\}_{j=1}^q$ by following two stages.

First Stage: We utilize **sample selection** to obtain the extra information that implies clean label correlations, which can be used to estimate $\hat{P}\left(\bar{Y}^i \mid Y^j\right)$.

Second Stage: We perform **co-occurrence estimation** ($\hat{P}\left(\bar{Y}^i, \bar{Y}^j\right)$ and $\hat{P}\left(\bar{Y}^i \mid Y^j\right)$) by frequency counting, and then estimate the transition matrix $\boldsymbol{T}^j$ by **solving the following probability equation**.

$$\hat{P}\left(\bar{Y}^j, \bar{Y}^i\right) = \sum_{Y^j} \hat{P}\left(Y^j\right) \hat{P}\left(\bar{Y}^j \mid Y^j\right) \hat{P}\left(\bar{Y}^i \mid Y^j\right),$$

where $\hat{P}\left(\bar{Y}^j \mid Y^j\right)$ represents the transition matrix $\hat{\boldsymbol{T}}^j$.

Experiments

Multi-label classification datasets:

- VOC2007 (20 classes)
- VOC2012 (20 classes)
- MS-COCO (80 classes)

Generate label noise:

$$\boldsymbol{T}^j = \boldsymbol{T} = \begin{pmatrix} 1-\rho_- & \rho_- \\ \rho_+ & 1-\rho_+ \end{pmatrix}$$

where $j = 1, 2, ..., q$.

Metric for estimating transition matrices:

- Estimation Error $= \sum_{j=1}^{q} \| \boldsymbol{T}^j - \hat{\boldsymbol{T}}^j \|_1 / \| \boldsymbol{T}^j \|_1$

Metric for classification performance:

- mean Average Precision (mAP)
- Overall F1-measure (OF1)
- per-Class F1-measure (CF1)

Experiments

Table: Comparison for estimating transition matrices on Pascal-VOC2007 dataset.The best results are in **bold**.

Noise rates (ρ_-, ρ_+)	(0,0.2)	(0,0.6)	(0.2,0)	(0.6,0)	(0.1,0.1)	(0.2,0.2)	(0.017,0.2)	(0.034,0.4)
T-estimator max	3.89±0.0	10.52±0.5	3.01±0.1	4.47±0.2	3.18±0.2	5.28±0.2	3.99±0.1	6.28±0.4
T-estimator 97%	4.95±0.1	4.42±0.1	1.77±0.0	2.13±0.1	6.99±0.1	6.94±0.1	5.38±0.1	5.17±0.0
DualT-estimator max	1.94±0.1	7.29±0.1	1.03±0.0	2.68±0.1	**2.13±0.2**	4.02±0.1	**1.71±0.0**	2.67 ±0.2
DualT-estimator 97%	12.59±0.1	7.43±0.1	1.09±0.0	2.41±0.3	14.39±0.1	11.78±0.1	13.71±0.2	11.15±0.1
Our estimator	**1.51±0.1**	**2.30±0.1**	**0.37±0.1**	**1.34±0.3**	3.06±0.4	**3.21±0.3**	2.03±0.2	**1.84±0.3**

Table: Summary of the Wilcoxon signed-ranks test for **Reweight-Ours** against other baselines at 0.1 significance level.

Reweight-Ours against	Standard	GCE	CDR	AGCN	CSRA	Reweight-T max	Reweight-T 97%	Reweight-DualT max	Reweight-DualT 97%
mAP	tie[0.29]	tie[0.26]	tie[0.32]	tie[0.18]	tie[0.45]	tie[0.32]	tie[0.23]	tie[0.35]	**win**[0.02]
OF1	**win**[0.00]	**win**[0.05]	**win**[0.04]	**win**[0.02]	**win**[0.05]	**win**[0.01]	**win**[0.08]	**win**[0.09]	**win**[0.00]
CF1	**win**[0.02]	**win**[0.01]	**win**[0.01]	**win**[0.01]	**win**[0.02]	**win**[0.03]	**win**[0.02]	**win**[0.09]	**win**[0.00]

Thank you for your hearing!

Code: `https://github.com/ShikunLi/Estimating_T_For_Noisy_Mutli-Labels`