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

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, Tongliang Liu

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Background: Recently, the problem of label-noise learning in multi-label classification has received more and more attention, since it is time-consuming and expensive to collect large-scale accurate labels and the noisy labels are much cheaper and easier to acquire.

Problem Setting: In the setting of **noisy multi-label learning**, the multiple labels assigned to an instance may be corrupted simultaneously with its respective transition matrix:

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

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

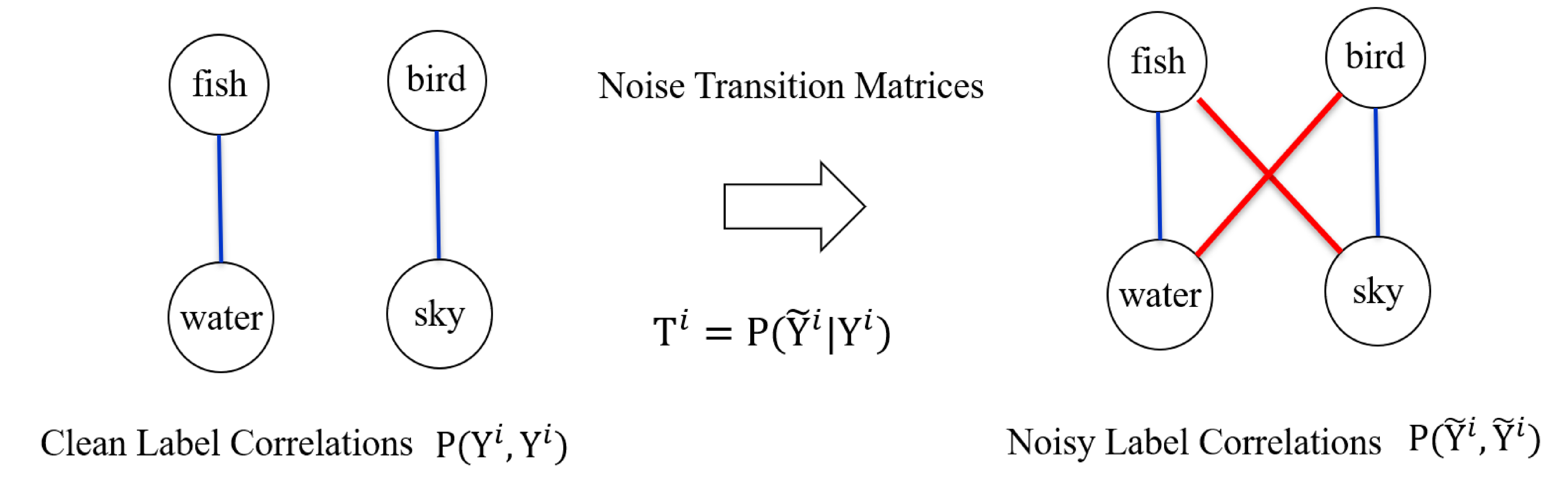
Noisy Data $(\boldsymbol{X}, \bar{\boldsymbol{Y}})$, where $\bar{\boldsymbol{Y}} = \{\bar{Y}^1, \bar{Y}^2, \ldots, \bar{Y}^q\} \in \{0,1\}^q$

Motivation: 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, to study estimating transition matrix in noisy multi-label learning is valuable and important.

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.



Contributions:

- To address the problem of identifying the transition matrices in **noisy multi-label learning**, we **prove some identifiability results** of the class-dependent transition matrix in such setting.
- Inspired by the identifiability results, we propose **a new estimator by exploiting label correlations** without neither anchor points nor accurate fitting of noisy class posterior.

Method

Instance-independent Assumption: 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. $P\left(\bar{Y}^j = k \mid Y^j = i, \boldsymbol{X} = \boldsymbol{x}\right) = P\left(\bar{Y}^j = k \mid Y^j = i\right) = T^j_{ik}$.

Identifiability Results:

> **Theorem:** If $P(\bar{Y}^i \mid Y^i)$ 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.

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

First Stage: We utilize **sample selection** to obtain the extra information that implies clean label correlations.

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

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

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

Official Pytorch Code:

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

Feel Free to Contact Us:

{lishikun,zhanghansong,geshiming}@iie.ac.cn, xxia5420@uni.sydney.edu.au,zhanyibing@jd.com, tongliang.liu@sydney.edu.au

Experiments

Experiments Setting:

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)

Comparison for estimating transition matrices:

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±0.03	10.52±0.58	3.01±0.12	4.47±0.22	3.18±0.22	5.28±0.20	3.99±0.10	6.28±0.44
T-estimator 97%	4.95±0.17	4.42±0.18	1.77±0.03	2.13±0.12	6.99±0.10	6.94±0.17	5.38±0.14	5.17±0.09
Dual T-estimator max	1.94±0.13	7.29±0.16	1.03±0.04	2.68±0.13	**2.13±0.23**	4.02±0.18	**1.71±0.08**	2.67 ±0.27
Dual T-estimator 97%	12.59±0.06	7.43±0.06	1.09±0.03	2.41±0.33	14.39±0.10	11.78±0.06	13.71±0.16	11.15±0.09
Our estimator	**1.51±0.12**	**2.30±0.13**	**0.37±0.08**	**1.34±0.33**	3.06±0.38	**3.21±0.32**	2.03±0.19	**1.84±0.32**

(a) VOC2007 dataset

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	3.90±0.01	10.28±0.33	2.87±0.09	4.55±0.08	3.29±0.07	5.25±0.15	4.05±0.04	6.82±0.20
T-estimator 97%	5.42±0.09	3.98±0.09	1.53±0.06	1.91±0.07	6.43±0.16	6.20±0.17	5.76±0.27	5.16±0.14
Dual T-estimator max	1.02±0.20	5.13±0.26	1.07±0.07	2.06±0.12	1.94±0.05	2.59±0.16	1.17±0.13	1.93±0.08
Dual T-estimator 97%	12.94±0.06	7.49±0.03	1.14±0.04	2.94±0.18	14.23±0.08	11.56±0.05	13.97±0.09	11.10±0.08
Our estimator	**0.83±0.10**	**1.94±0.15**	**0.26±0.03**	**0.91±0.12**	**1.74±0.22**	**1.79±0.17**	**0.94±0.07**	**1.07±0.14**

(b) VOC2012 dataset

Noise rates (ρ_-, ρ_+)	(0,0.2)	(0,0.6)	(0.2,0)	(0.6,0)	(0.1,0.1)	(0.2,0.2)	(0.008,0.2)	(0.015,0.4)
T-estimator max	16.14±0.33	39.09±0.47	10.39±0.21	11.49±0.60	13.95±0.41	20.50±0.04	16.70±0.06	28.16±0.45
T-estimator 97%	50.49±0.01	25.70±0.08	4.04±0.08	3.70±0.02	51.17±0.16	39.45±0.11	49.96±0.18	37.54±0.10
Dual T-estimator max	**5.04±0.04**	**11.22±0.70**	4.65±0.07	9.55±0.84	13.02±0.45	15.79±0.38	**7.04±0.31**	**6.34±0.11**
Dual T-estimator 97%	61.49±0.02	30.97±0.03	1.53±0.00	7.86±0.12	64.20±0.02	48.67±0.01	63.12±0.02	46.91±0.01
Our estimator	7.42±0.38	11.23±0.11	**0.50±0.03**	**0.83±0.06**	**8.88±0.10**	**10.27±0.19**	7.51±0.43	8.77±0.20

(c) MS-COCO dataset

Comparison for classification performance:

Reweight-Ours against	Standard	GCE	CDR	AGCN	CSRA	WSIC	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 [0.26]	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** [0.02]	**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** [0.00]	**win** [0.03]	**win** [0.02]	**win** [0.09]	**win** [0.00]

Summary of the Wilcoxon signed-ranks test for Reweight-Ours against other comparing approaches at 0.1 significance level. The p-values are shown in the brackets.

Welcome to read our paper for ablation study and more experimental results!