Paper ID: 1691

Beyond Images:
Label Noise Transition Matrix Estimation for Tasks with Lower-Quality Features

Zhaowei Zhu, Jialu Wang, Yang Liu

UC Santa Cruz

{**zwzhu**, faldict, yangliu}@ucsc.edu

Code

A Data-Centric Method

Noise Transition Matrix **T**

- Each element of **T**:

$$T_{ij} := \mathbb{P}(\widetilde{Y} = j | Y = i)$$

- Clean label *i* ➜ Noisy label *j*

Why we need T?

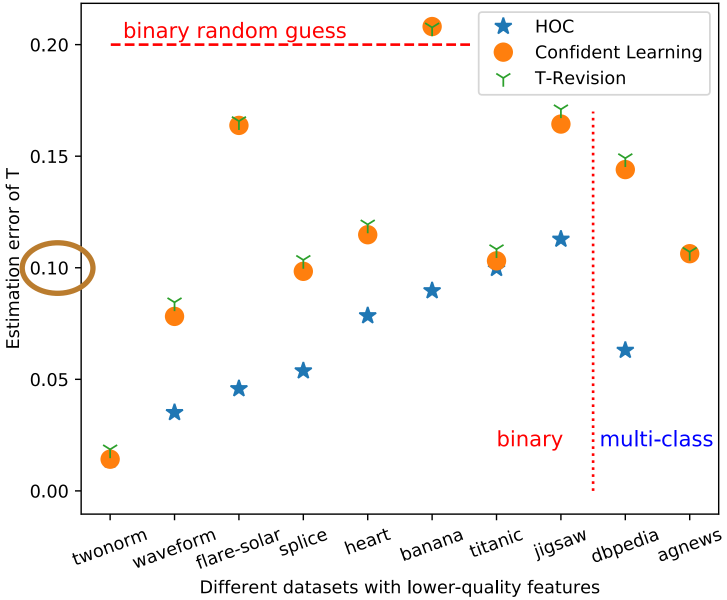
- **T** is important in many applications, e.g.
  - Learning noise-tolerant classifiers (Natarajan et al., 2013)
  - Corrupted label detection (Northcutt et al., 2021; Zhu et al., 2022)
  - Machine learning fairness (Lamy et al., 2019; Wang et al., 2021)
  - Crowdsourcing (Liu & Liu, 2015)
  - Medical applications (McCormick et al., 2016)

[1] Natarajan, N. et al. Learning with noisy labels. *NeurIPS 2013.*
[2] Northcutt, C. G. et al. Pervasive label errors in test sets destabilize machine learning benchmarks. *NeurIPS 2021.*
[3] Zhu, Z. et al. Detecting corrupted labels without training a model to predict. *ICML 2022.*
[4] Lamy, A. et al. Noise-tolerant fair classification. *NeurIPS 2019.*
[5] Wang, J. et al. Fair classification with group-dependent label noise. *FAccT 2021.*
[6] Liu, Y. and Liu, M. An online learning approach to improving the quality of crowdsourcing. *SIGMETRICS 2015.*
[7] McCormick, T. H. et al. Probabilistic cause-of-death assignment using verbal autopsies. *JASA 2016.*

BUT existing methods focus on image classifications

- Failures in tasks beyond images

Large error for binary classifications



Main reasons:

- Complicated task
- Lower-quality features

HOC: Z. Zhu, Y. Song, Y. Liu. Clusterability as an alternative to anchor points when learning with noisy labels. *ICML 2021.*
CL: Northcutt, C., Jiang, L., and Chuang, I. Confident learning: Estimating uncertainty in dataset labels. *JAIR 2021.*
T-Revision: Xia, X. et al. Are anchor points really indispensable in label-noise learning? *NeurIPS 2019.*

Our Contributions

f-MI: f-mutual information

We estimate **T** for tasks
with **lower-quality** features!

Highlights:

- ✓ Reweighting mechanism (Sec. 3)
- ✓ Get proper weights with only noisy labels: Noisy f-MI → Clean f-MI (Sec. 4)

We build on HOC

comparable representations $\xrightarrow{\text{(\textbf{without} prob. of each class)}}$ **T**

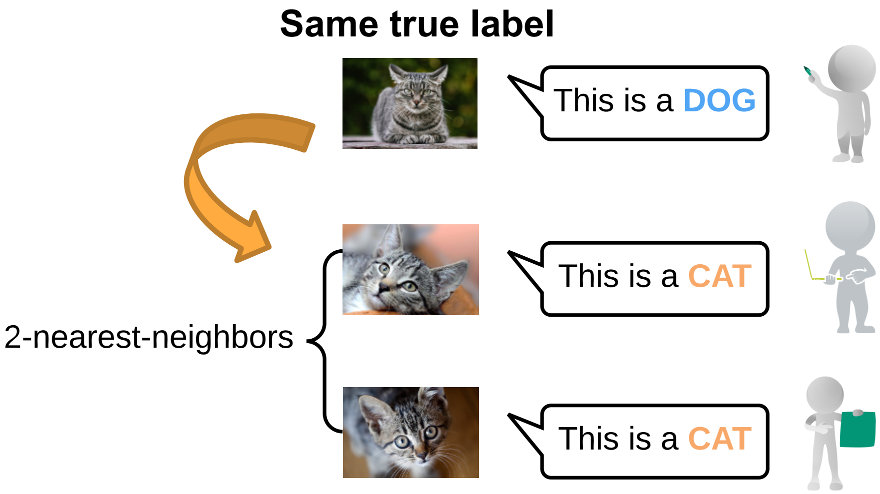
Benefits:

1. Avoid complicated hyperparameters tuning
2. Build light tool for more general applications

A Data-Centric Method

What is HOC?

- Features that satisfy 2-NN label clusterability



- Check *label consensuses* of 2-NN features

Pattern	#
(CAT, CAT, CAT)	2
(DOG, CAT, CAT)	5
…	…

Questions:

➥ **Features with clusterability (our focus)**

➥ # Consensus patterns → Decode **T (HOC)**

HOC: Z. Zhu, Y. Song, Y. Liu. Clusterability as an alternative to anchor points when learning with noisy labels. *ICML 2021*.

Improve Clusterability

- **Challenge:**
  - Lower-quality features → poor clusterability
  - Find a similarity measure that improves clusterability

- **Intuition:** Downweight less informative parts by matrix ***W***
- **Solution:** Soft cosine similarity (Definition 3.1)

$$\text{Sim}_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{(\sqrt{\boldsymbol{W}}\boldsymbol{x})^\top (\sqrt{\boldsymbol{W}}\boldsymbol{x}')}{\|\sqrt{\boldsymbol{W}}\boldsymbol{x}\|_2 \|\sqrt{\boldsymbol{W}}\boldsymbol{x}'\|_2}$$

UC SANTA CRUZ

Property of *W*

$$\text{Sim}_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{(\sqrt{\boldsymbol{W}}\boldsymbol{x})^\top(\sqrt{\boldsymbol{W}}\boldsymbol{x}')}{\|\sqrt{\boldsymbol{W}}\boldsymbol{x}\|_2\|\sqrt{\boldsymbol{W}}\boldsymbol{x}'\|_2}$$

- **Symmetric:** $\text{Sim}_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{x}') = \text{Sim}_{\boldsymbol{W}}(\boldsymbol{x}', \boldsymbol{x})$

- **Correlation** monotone (between different elements of X)

  (off-diagonal) Less correlated parts have lower weights → Remove correlation

- **Information** monotone (between X, Y) → Noisy f-MI → Clean f-MI

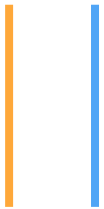
  (diagonal) Less informative parts have lower weights

Remove Correlation

$$\text{Sim}_{\boldsymbol{W}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{(\sqrt{\boldsymbol{W}}\boldsymbol{x})^\top (\sqrt{\boldsymbol{W}}\boldsymbol{x}')}{\|\sqrt{\boldsymbol{W}}\boldsymbol{x}\|_2 \|\sqrt{\boldsymbol{W}}\boldsymbol{x}'\|_2}$$

Correlated features $\boldsymbol{X}$

Project onto its **eigen** basis



Discorrelated features $\boldsymbol{Z}$

***W* could be simplified!**

$$\begin{bmatrix} I_f(Z_1, Y), & 0, & \cdots, & 0 \\ 0, & I_f(Z_2, Y), & \cdots, & 0 \\ \vdots & & \ddots & \vdots \\ 0, & 0, & \cdots, & I_f(Z_2, Y) \end{bmatrix}$$

Estimate clean f-MI with only noisy labels

Apply $\text{Sim}_{\boldsymbol{W}}(\boldsymbol{z}, \boldsymbol{z}')$ in HOC

T

f-MI: f-mutual information

How does noisy *f*-MI affect?
We care about the information monotone!

Definition 4.1 (ϵ-Order-Preserving Under Label Noise)
(Sketch) **Noisy** f-MI preserves the same order as **clean** f-MI with tolerance ϵ.

Binary classifications:

Theorem 4.4 (Total-Variation)

Total-Variation is 0-Order-Preserving.

Theorem 4.5 (KL-Divergence)

KL Divergence is ϵ-Order-Preserving, where

$$\epsilon = e_1[\delta \log \delta - (1+\delta)\log(1+\delta)] + H(e_1), \quad e_1 = \delta e_2.$$

Experiments

Estimation Error ↓

Text Datasets $(d, [N_1, \cdots, N_K])$	Noise Rate	T-Rev	CL	HOC	Ours-$\boldsymbol{X}$-KL	Ours-$\boldsymbol{X}$-TV	Ours-$\boldsymbol{A}$-KL	Ours-$\boldsymbol{A}$-TV
AG's news (BERT) (768, [30k × 4])	Low	10.38	11.41	13.32	12.65	12.75	**8.36**	**8.35**
	Medium	10.71	10.63	10.62	10.13	10.45	**6.44**	**6.52**
	High	13.97	13.82	6.80	6.83	6.69	**4.54**	**4.19**
DBpedia (BERT) (768, [40k × 14])	Low	6.80	5.31	7.57	6.76	6.94	**2.52**	**2.52**
	Medium	14.91	14.40	6.30	5.66	5.78	**2.33**	**2.28**
	High	24.23	23.28	6.00	5.18	5.22	**2.42**	**2.43**
Yelp-5 (BERT) (768, [130k × 5])	Low	38.49	38.75	40.87	40.71	40.58	**37.37**	**37.19**
	Medium	35.46	36.05	33.63	34.23	33.88	**31.79**	**31.94**
	High	21.20	20.88	19.09	18.56	20.13	**18.11**	**18.06**
Jigsaw (BERT) (768, [144,277, 15,294])	Low	20.92	20.17	14.25	14.07	14.24	**9.76**	**9.97**
	Medium	17.10	16.44	11.28	11.80	12.23	**7.45**	**7.66**
	High	7.19	6.81	4.84	4.85	3.43	**0.78**	**1.02**

Consistent improvement under BERT embeddings.

Experiments

- Downstream applications: Learn a noise-consistent classifier

Method	AG's news		DBpedia	
	Last	Best	Last	Best
HOC (Zhu et al., 2021c)	82.17	83.08	91.06	91.06
Ours-A-TV	**85.01**	**85.17**	**97.71**	**97.77**

Test Accuracy ↑

UC SANTA CRUZ

Thank you !

Hall E
#517

Want to know more?

1. Scan the QR code to reproduce
2. Join our **onsite** poster session

Code