JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Leveraging Inter-rater Agreement for Classification in the Presence of Noisy Labels

TUE-AM-327

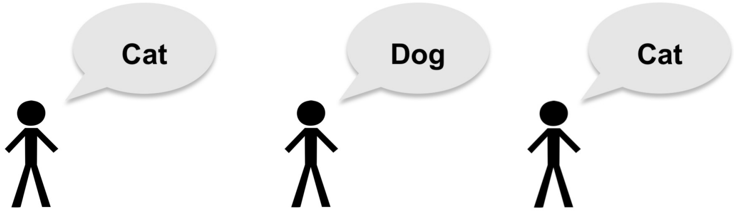
Maria Sofia Bucarelli[1], Lucas Cassano[2], Federico Siciliano[1], Amin Mantrach[2], Fabrizio Silvestri[1,3]





[1] Sapienza University of Rome, [2] Amazon, [3] ISTI-CNR, Pisa, Italy

What is this?



Background

- Classification datasets are obtained through a human labeling process.
- Annotators labels can be noisy
- Popular approaches for label combination are majority vote or soft labelling.
- No published results on leveraging IAA statistics for label noise estimation.
- Existing noise tolerant training methods rely on unknown quantities.

Contributions

- Methodology to estimate label noise distribution using IAA statistics.
- Leveraging the estimate to learn from noisy datasets.
- Providing generalization bounds based on IAA statistics.

Setting & Assumptions

Instance independent noise : $\mathbb{P}(y_a|y,x) = \mathbb{P}(y\_a|y)$

Noise Transition matrix :
$(T)^a_{ij} := \mathbb{P}(y_a = j|y = i)$

1) All annotators have the same noise transition matrix T.

2) T is symmetric and with diagonal elements larger than 0.5

Proposition

T is positive definite

3) Annotators are conditionally independent on the true label: $\mathbb{P}(y_a, y_b|y) = \mathbb{P}(y_a|y)\mathbb{P}(y_b|y)$

4) Classes distribution is known $v_i = \mathbb{P}(y = i),\ D := diag(v)$

Definition

The IAA matrix M_{ab} between annotators a and b is: $(M_{ab})_{ij} := \mathbb{P}(y_a = i, y_b = j)$

Proposition

M_{ab} can be written as a and b is: $M_{ab} = T_a^T\ D\ T_b$.

Estimation of the noise transition matrix

Lemma

If $D^{\frac{1}{2}}$ commutes with T we have that: $T = U\Lambda^{\frac{1}{2}}U^T$ where $U\Lambda U^T$ is an eigenvalue decomposition of $D^{-\frac{1}{2}}MD^{-\frac{1}{2}}$.

If the annotators have the same transition matrix, we can estimate M as follows:

$$(\widehat{M})_{ij} = \frac{1}{H(H-1)} \sum_{a=1}^{H} \sum_{\substack{b=1,\\ b\neq a}}^{H} \sum_{k=1}^{n} \frac{1_{[y_{a,k},=i, y_{b,k}=j]}}{n}$$

- Obtain the eignevalue decomposition of $D^{\frac{1}{2}}\widehat{M}D^{-\frac{1}{2}} = \widehat{U}\widehat{\Lambda}\,\widehat{U}^T$. $\longrightarrow$ $\widehat{T} = \widehat{U}\widehat{\Lambda}^{\frac{1}{2}}\,\widehat{U}^T$
- A more accurate estimate of T could be obtained as $\widehat{T} = \pi(\widehat{U}\widehat{\Lambda}^{\frac{1}{2}}\,\widehat{U}^T)$ where π is a projection operator to the set of doubly stochastic, positive definite and diagonally dominant matrices

Estimation of the noise transition matrix

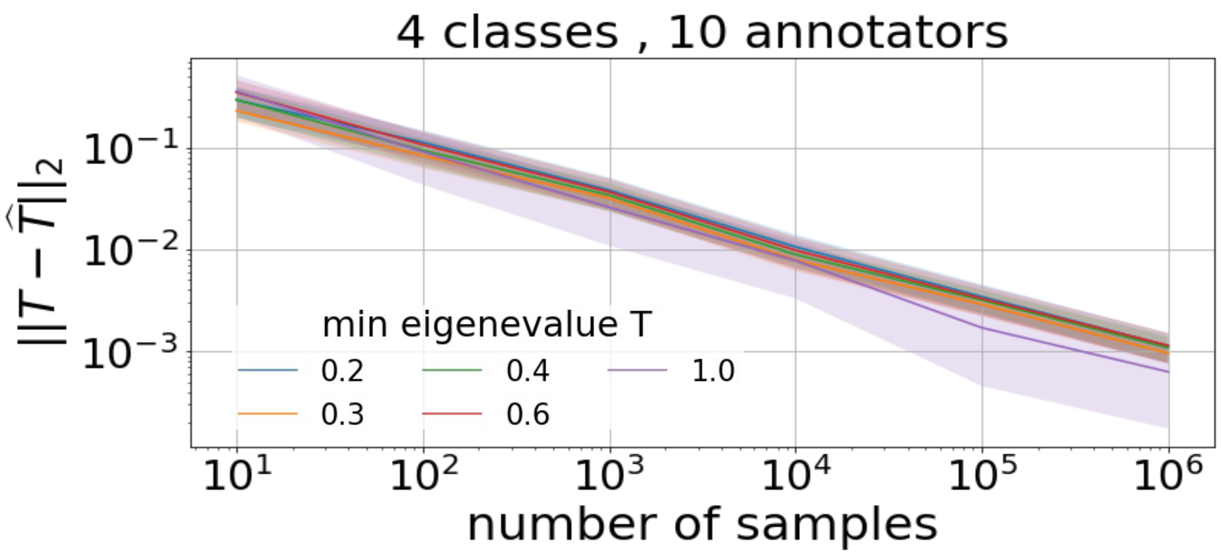
Theorem

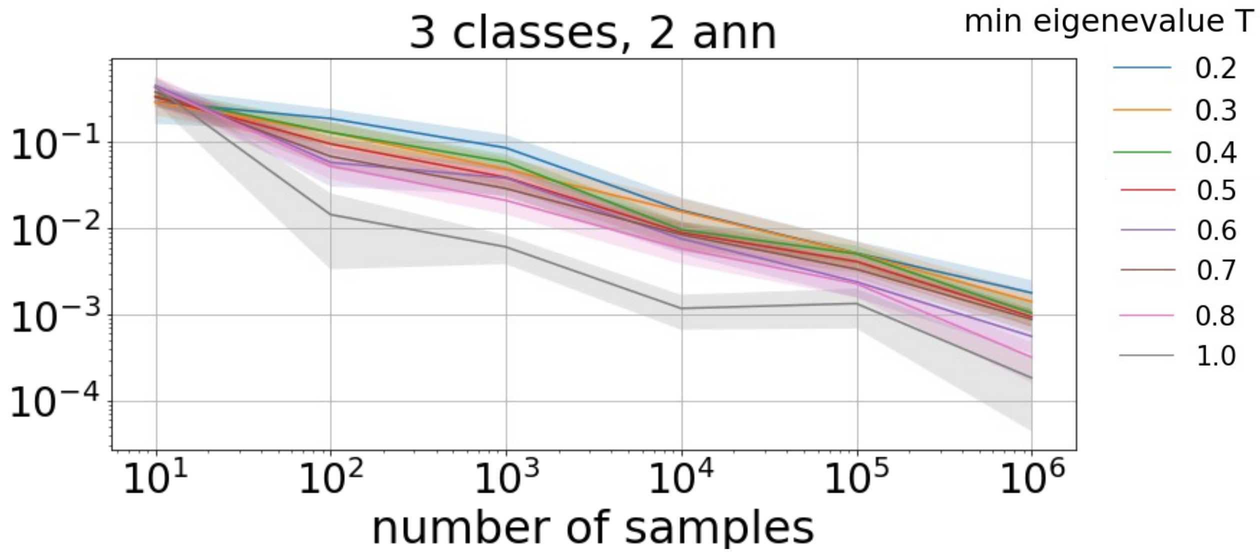
Let T be the noise transition matrix and $\hat{T}$ its estimate. With probability at least $1-\delta$:

$$|| T - \hat{T} ||_2 \leq \frac{C\left(\sqrt{C}+1\right)\lambda_{max}(D)}{\lambda_{min}(\hat{T})}\sqrt{\frac{1}{2n}\ln\frac{2C^2}{\delta}} \quad \text{and.} \quad || T^{-1} - \hat{T}^{-1} ||_2 \leq \frac{9C\left(\sqrt{C}+1\right)\lambda_{max}(D)}{\lambda_{min}(\hat{T})^2}\sqrt{\frac{1}{2n}\ln\frac{2C^2}{\delta}}$$

Experimental results

We performed experiments to validate the effectiveness of the method we propose for estimating T:





Leveraging T for learning

We can calculate the posterior distributions as follows:

$$\underbrace{\mathbb{P}(y_i = c \mid y_{1,i}, \dots, y_{H,i})}_{:= p_{c,i}} \propto \mathbb{P}(y_i = c) \prod_{h=1}^{H} \underbrace{\mathbb{P}(y_i = c \mid y_{1,i}, \dots, y_{H,i})}_{:= T_{c, y_{h,i}}}$$

Lemma

$$\lim_{H \to \infty} p_{c,i} = 1_{[y_i = c]} \quad a.s.$$

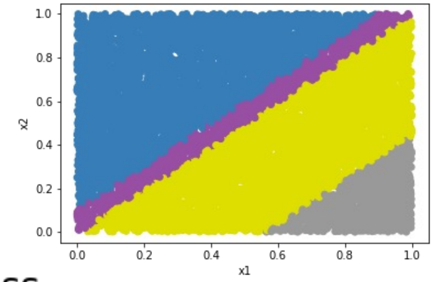
We can use the posterior distributions as soft-labels : $\ell\big(f(x_i), y_{1,i}, \dots, y_{H,i}\big) = \ell(f(x_i), \overline{p_i})$
where , $\overline{p_i} = \left[p_{1,i}, \dots, p_{c,i}\right]^T$.

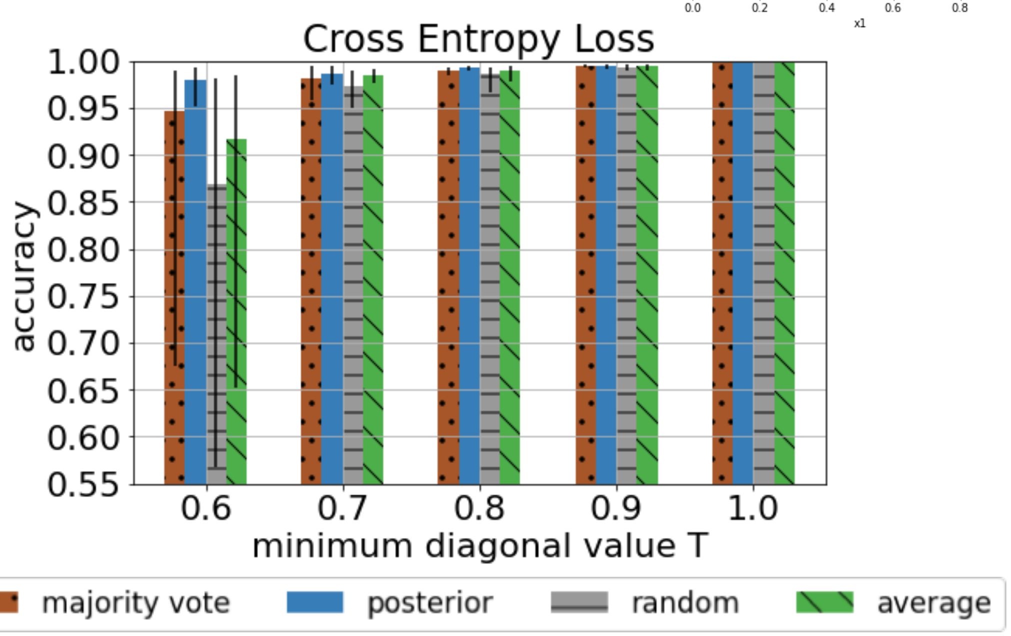
Or to weight the loss function : $\ell\big(f(x_i), y_{1,i}, \dots, y_{H,i}\big) = \sum_{c=1}^{C} p_{c,i}\, \ell(f(x_i), e_c)$
with e_c c-th vector of the standard basis of $\mathbb{R}^C$.

Leveraging T for learning

We performed experiments to show how the estimated T can be leveraged to train classifiers in the presence of noise labels.

Synthetic dataset with features generated uniformly in $[0,1]^2$





CIFAR10-N dataset. In this dataset there are no guarantees that the assumptions we made on T are satisfied, however, the method is still applicable with positive results.

Aggregation Method	Pretrained	Not-Pretrained
random	0.718 ± 0.035	0.579 ± 0.023
majority vote	0.740 ± 0.017	0.590 ± 0.006
average	0.762 ± 0.012	0.637 ± 0.016
posteriors (ours)	$\mathbf{0.794} \pm \mathbf{0.005}$	$\mathbf{0.652} \pm \mathbf{0.014}$

Table: Test Accuracy on CIFAR10-N with Resnet34

Generalization Gap Bounds

We can leverage the estimation of T in the *backward* and *forward* losses.
Let $\ell(t,y)$ be a generic loss.

$$\ell(t) \coloneqq [\ell(t,e_1),\dots,\ell(t,e_C)]$$
$$l_b\,(t,y) = (T^{-1}\,\ell(t))_y$$
$$l_b\,(t,y) = \ \ell(T^T t)_y$$

We derived generalization gap bounds for he backward loss computed using $\hat{T}$.

Theorem

Let ℓ_b be the backward loss for ℓ .

$$R_{\ell,\mathcal{D}}\,(\hat{f}) - \min_{f\in\mathcal{F}} R_{\ell,\mathcal{D}}\,(f) \leq \left[2\mathrm{L}\,\lambda_{min}\big(\hat{T}\big)^2 + \frac{\mu\lambda_{min}(D)}{\lambda_{min}\big(\hat{T}\big)^2}\sqrt{\frac{1}{n}\ln\left(\frac{4C}{\delta}\right)}\right]\mathfrak{R}(\mathcal{F})g(C),$$

with $1g(c) = 6C^2\sqrt{C} + 1$

Conclusions

We provided:

- A methodology to estimate label noise distribution using inter annotator agreement statistics
- A way to leverage estimated noise transition matrix to learn from noisy datasets
- Generalization bounds for backward loss based on IAA statistics. This bounds not dependent on true noise distribution (that is unkown), unlike previous works.

:

Thanks!