Semi-Supervised Aggregation of Dependent Weak Supervision Sources With Performance Guarantees

Alessio Mazzetto*, Dylan Sam, Andrew Park, Eli Upfal, Stephen H. Bach

*alessio_mazzetto@brown.edu

Introduction

Binary classification

$$y : \boxed{\mathcal{X}} \longrightarrow \{0,1\}$$

Classification domain with distribution $\mathcal{D}$

Example: classify ambulance.

$y(x)=1$ $y(x)=0$

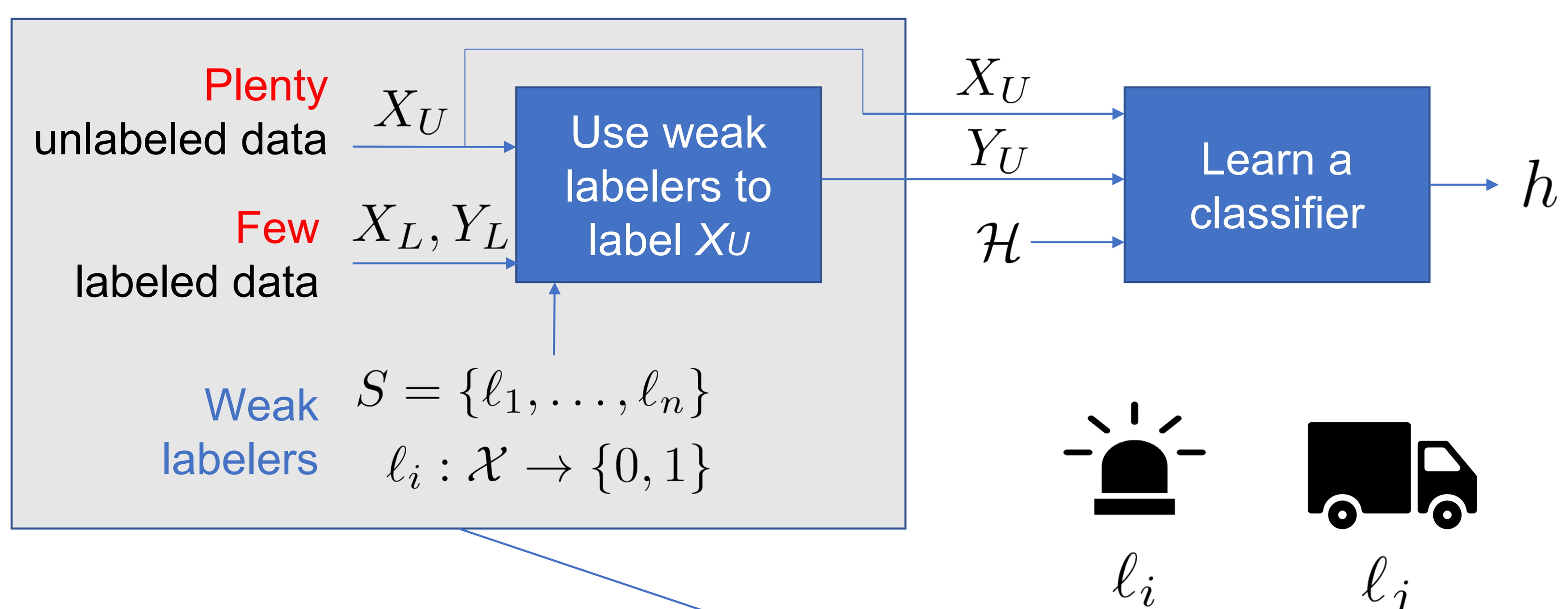
Given hypothesis class $\mathcal{H}$, we want to find the hypothesis h s.t

$$\min_{h\in\mathcal{H}} \varepsilon(h) := \min_{h\in\mathcal{H}} \underbrace{\Pr_{x\sim\mathcal{D}}(y(x)\neq h(x))}_{\text{error of } h}$$

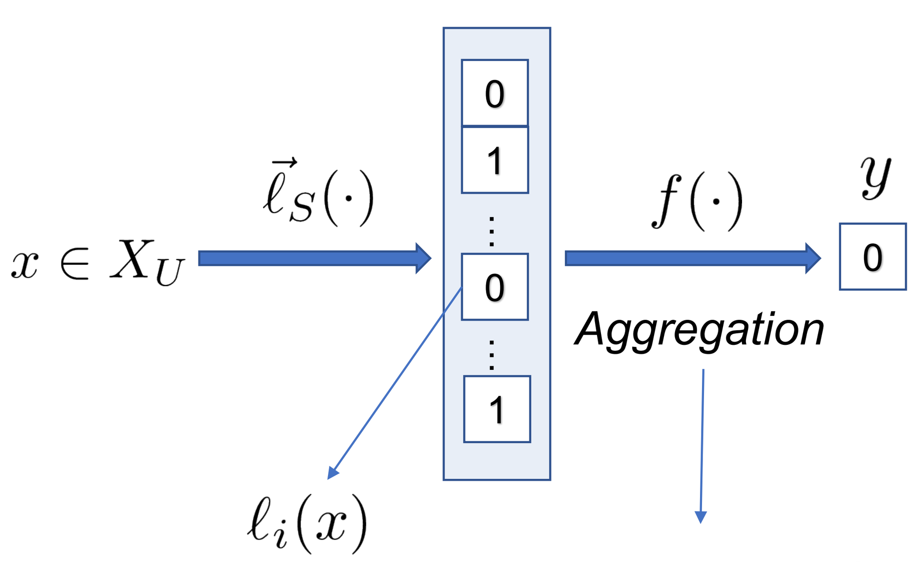
Supervised learning of a binary classification task requires a lot of labeled data for high-dimensional hypothesis classes (e.g., DNN).

Labeled data is costly and scarce for a lot of binary classification task of interest.

Weak Supervision Framework [1]



Our focus



We use weak labelers in order to provide *noisy labels* to a unlabeled dataset.

Contribution

Previous work unrealistically usually assumes *independence* or a *distribution family* between weak labelers' errors to do aggregation.

Our contribution:

- First theoretical bound to the worst-case error of the majority vote of a set of weak labelers without those assumptions.
- Novel algorithm that uses the bound above to provide the first theoretical guarantees in learning an aggregation of an arbitrary set of weak labelers.

Intuition

Preliminary definitions:

- Let error rate of *i*-th labeler be $\epsilon_i = \varepsilon(\ell_i) \longrightarrow$ easy to estimate with few labeled data
- Let $S(\vec{\epsilon})$ be the set of all set of labelers that have error rates equal to
$$\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$$
- Given a vector $\vec{a} \in \{0,1\}^n$, let $\lambda(\vec{a})$ be its majority vote.

Our result (knowledge of error rates is not enough):

$$\boxed{\max_{S\in S(\vec{\epsilon})} \varepsilon(\lambda \circ \vec{\ell}_S) \geq \text{median}\{\epsilon_1, \ldots, \epsilon_n\}}$$

Worst-case, we cannot improve upon the best labeler, if we only know their error rates

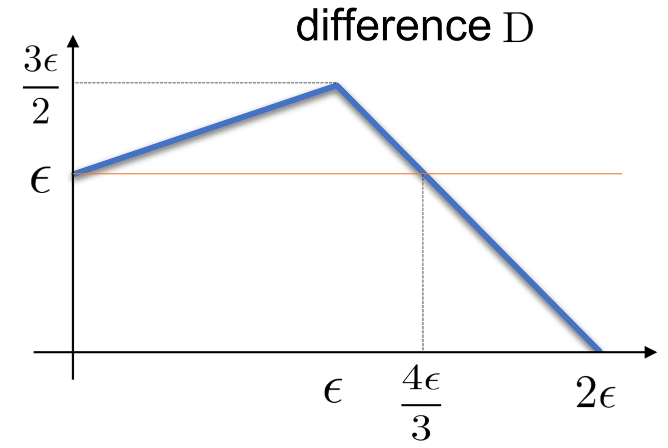
(!) With independence assumption, error of majority vote would go to zero.

We need auxiliary information: pairwise difference between labelers.

$$\mathbf{D}_{ij} = \Pr_{x\sim\mathcal{D}}(\ell_i(x) \neq \ell_j(x)) \longrightarrow \text{only need unlabeled data to estimate it}$$

Three labelers with same error rate ϵ and same pairwise difference D

Worst-case error of majority vote $\max_{S\in S(\vec{\epsilon},\mathbf{D})} \varepsilon(\lambda \circ \vec{\ell}_S)$



Pairwise difference D

Take into account pairwise difference

(!) For large pairwise difference, we can improve upon best labeler.

Method

Goal: find subset of weak labelers with lowest worst-case error on their majority vote.

- Closed formula for set of three weak labelers.
- Heuristic: iteratively add the two labelers that yield the lowest worst-case error on their majority vote.

For a set of labelers $S = \{\ell_1, \ldots, \ell_n\}$ with error rate $\vec{\epsilon}$ and pairwise difference $\mathbf{D}$, we have that:

$$\max_{S\in\mathcal{S}(\vec{\epsilon},\mathbf{D})} \varepsilon(\lambda \circ \vec{\ell}_S) = \max \sum_{\vec{a}\in\{0,1\}^n : |\vec{a}|_1 < n/2} p_{\vec{a}}$$

$$(a) \quad \sum_{\vec{a}\in\{0,1\}^n : a_i = 0} p_{\vec{a}} = \epsilon_i \quad \text{for } i = 1, \ldots, n$$

$$(b) \quad \sum_{\vec{a}\in\{0,1\}^n : a_i \neq a_j} p_{\vec{a}} = \mathbf{D}_{ij} \quad \text{for } i \neq j$$

$$(c) \quad \sum_{\vec{a}} p_{\vec{a}} = 1$$

Linear program with O(2ⁿ) variables and O(n²) constraints

$$(d) \quad p_{\vec{a}} \geq 0 \ \forall \vec{a}$$

Experiments

Animals With Attribute (AwA2 [2]) dataset. Each class has 85 attributes, used to create weak classifiers.

	Baselines		State-of-the-art [3]	Our methods		
Dataset	Majority Vote	Dawid-Skene	ALL	**PGMV**	**PGMV-P**	**PGMV-D**
AwA2 (1)	79.1 ± 1.1	80.0 ± 1.8	84.2 ± 0.9	82.0 ± 1.1	85.5 ± 0.9	84.3 ± 1.3
AwA2 (2)	90.0 ± 0.7	94.7 ± 0.4	93.5 ± 0.5	93.7 ± 0.4	93.7 ± 0.5	94.1 ± 0.4
AwA2 (3)	92.3 ± 1.0	96.7 ± 0.3	95.5 ± 0.5	95.4 ± 0.3	95.9 ± 0.3	96.3 ± 0.2
AwA2 (4)	94.2 ± 0.6	96.8 ± 0.2	93.8 ± 0.8	96.8 ± 0.2	97.0 ± 0.3	96.8 ± 0.2
AwA2 (5)	97.6 ± 0.6	99.0 ± 0.2	96.3 ± 0.7	97.5 ± 0.3	98.3 ± 0.3	98.8 ± 0.2

Accuracy over different tasks, grouped by quality of the weak labelers, using ~800 unlabeled and labeled data.

References:

[1]: Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and R´e, C. (2017). Snorkel: Rapid training data creation with weak supervision. PVLDB.

[2]:Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. PAMI

[3]: Arachie, C. and Huang, B. (2019). Adversarial label learning. AAAI.