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

$y : \mathcal{X} \mapsto \{0, 1\}$

Example: classify ambulance.

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

$$\min_{h \in \mathcal{H}} \epsilon(h) := \min_{h \in \mathcal{H}} \mathbb{P}_D(y(x) \neq h(x))$$

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 tasks of interest.

Weak Supervision Framework [1]

![Diagram of weak supervision framework](image)

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 $\epsilon_i = \epsilon(\ell_i)$ easy to estimate with few labeled data

• Let $S(\tilde{e})$ be the set of all set of labelers that have error rates equal to $\tilde{e} = (\epsilon_1, \ldots, \epsilon_n)$

• Given a vector $\tilde{a} \in \{0, 1\}^n$, let $\lambda(\tilde{a})$ be its majority vote.

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

$$\max_{S(\tilde{e})} \epsilon(\lambda \circ \tilde{e}) \geq \text{median}\{\epsilon_1, \ldots, \epsilon_n\}$$

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.

$$D_{ij} = \mathbb{P}_D(\ell_i(x) \neq \ell_j(x))$$

only need unlabeled data to estimate it

$$\text{Three labels with same error rate } \epsilon \text{ and same pairwise difference } 1$$

(!) 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 $\epsilon'$ and pairwise difference $D'$, we have that:

$$\max_{S(\tilde{e})} \epsilon(\lambda \circ \tilde{e}) = \max_{a \in \{0, 1\}^n} \sum_{i=1}^n p_{\tilde{a}_i}$$

(a) $\sum_{i=1}^n p_{\tilde{a}_i} = \epsilon'_i$ for $i = 1, \ldots, n$

(b) $\sum_{i=1}^n p_{\tilde{a}_i} = D_i$ for $i \neq j$

(c) $\sum_{i=1}^n p_{\tilde{a}_i} = 1$

(d) $p_{\tilde{a}_i} \geq 0 \forall \tilde{a}_i$

Linear program with $O(2^n)$ variables and $O(n^2)$ constraints

Experiments

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

Baseline

State-of-the-art [3]

Our methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority Vote</th>
<th>Dense-Slime</th>
<th>ALL</th>
<th>PGMV</th>
<th>PGMV-P</th>
<th>PGMV-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AwA2 (1)</td>
<td>79.1 ± 0.1</td>
<td>98.0 ± 0.1</td>
<td>98.0 ± 0.1</td>
<td>98.0 ± 0.1</td>
<td>98.0 ± 0.1</td>
<td>98.0 ± 0.1</td>
</tr>
<tr>
<td>AwA2 (2)</td>
<td>94.0 ± 0.7</td>
<td>94.0 ± 0.7</td>
<td>94.0 ± 0.7</td>
<td>94.0 ± 0.7</td>
<td>94.0 ± 0.7</td>
<td>94.0 ± 0.7</td>
</tr>
<tr>
<td>AwA2 (3)</td>
<td>95.5 ± 0.5</td>
<td>95.5 ± 0.5</td>
<td>95.5 ± 0.5</td>
<td>95.5 ± 0.5</td>
<td>95.5 ± 0.5</td>
<td>95.5 ± 0.5</td>
</tr>
<tr>
<td>AwA2 (4)</td>
<td>96.2 ± 0.3</td>
<td>96.2 ± 0.3</td>
<td>96.2 ± 0.3</td>
<td>96.2 ± 0.3</td>
<td>96.2 ± 0.3</td>
<td>96.2 ± 0.3</td>
</tr>
<tr>
<td>AwA2 (5)</td>
<td>97.0 ± 0.3</td>
<td>97.0 ± 0.3</td>
<td>97.0 ± 0.3</td>
<td>97.0 ± 0.3</td>
<td>97.0 ± 0.3</td>
<td>97.0 ± 0.3</td>
</tr>
</tbody>
</table>

References:


Images from flaticon.com