Active Learning with Disagreement Graphs

Corinna Cortes[1], Giulia DeSalvo[1], Claudio Gentile[1],
Mehryar Mohri[1,2], **Ningshan Zhang**[3]

[1] Google Research [2] Courant Institute, NYU [3] NYU

ICML, June 12, 2019

On-line Active Learning Setup

- At each round $t \in [T]$, receives unlabeled $x_t \sim \mathcal{D}_{\mathcal{X}}$ i.i.d.
- Decides whether to request label:
  - If label requested, receives y_t.
- After T rounds, returns a hypothesis $h_T \in \mathcal{H}$.

Objective:

- Generalizations error:
  - Accurate predictor h_T: small expected loss $R(h_T) = \mathbb{E}_{x,y}\left[\ell(h_T(x), y)\right]$.
  - Close to best-in-class $h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$.
- Label complexity: few label requests.

Disagreement-based Active Learning

Key idea: Request label when there is some **disagreement** among hypotheses.
Examples:

- Separable case: CAL (Cohn et al., 1994).
- Non-separable case: A^2 (Balcan et al., 2006), DHM (Dasgupta et al., 2008).
- IWAL (Beygelzimer et al., 2009).

Can we improve upon existing disagreement-based algorithms, such as IWAL?

- Better guarantees?
- Leverage average disagreements?

This talk

- IWAL-D algorithm: enhanced IWAL with disagreement graph.
- IZOOM algorithm: enhanced IWAL-D with zooming-in.
- Better generalization and label complexity guarantees.
- Experimental results.

Disagreement Graph (D-Graph)

- Vertices: hypotheses in $\mathcal{H}$ (a finite hypothesis set)
- Edges: fully connected. The edge between $h, h' \in \mathcal{H}$ is weighted by their expected disagreement:

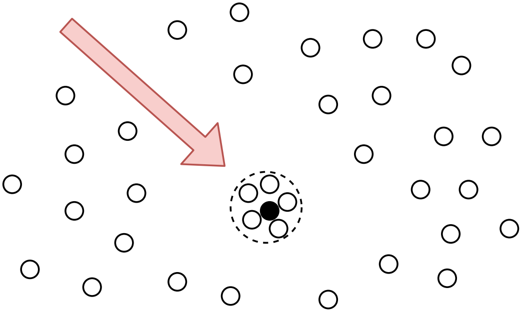
$$\mathcal{L}(h, h') = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\max_{y \in \mathcal{Y}} \left| \ell(h(x), y) - \ell(h'(x), y) \right| \right].$$

  $\mathcal{L}$ symmetric, $\ell \leq 1 \Rightarrow \mathcal{L} \leq 1$.
- D-Graph can be accurately estimated using unlabeled data.

Disagreement Graph (D-Graph)

One favorable scenario:

- Best-in-class h^* (●) is within an isolated cluster;
- $\mathcal{L}(h, h^*)$ is small within the cluster.



IWAL-D Algorithm: IWAL with D-Graph

- At round $t \in [T]$, receive x_t.
  1. Flip a coin $Q_t \sim \text{Ber}(p_t)$, with disagreement-based bias:
  $$p_t = \max_{h,h' \in \mathcal{H}_t} \max_{y \in \mathcal{Y}} \left| \ell(h(x_t), y) - \ell(h'(x_t), y) \right|.$$
  2. If $Q_t = 1$, request the label y_t.
  3. Trim the version space:
  $$\mathcal{H}_{t+1} = \left\{ h \in \mathcal{H}_t \colon L_t(h) \leq L_t(\widehat{h}_t) + \left(1 + \mathcal{L}(h, \widehat{h}_t)\right) \Delta_t \right\},$$
  which uses importance weighted empirical risk
  $$L_t(h) = \frac{1}{t} \sum_{s=1}^{t} \frac{Q_s}{p_s} \ell(h(x_s), y_s), \quad \widehat{h}_t = \operatorname*{argmin}_{h \in \mathcal{H}_t} L_t(h), \quad \Delta_t = \widetilde{O}\left(\sqrt{\frac{\log(T|\mathcal{H}|)}{t}}\right).$$
- After T rounds, return $\widehat{h}_T$.

IWAL-D vs. IWAL: Quantify the Improvement

Theorem (IWAL-D) With high probability,

$$R(\widehat{h}_T) \leq R^* + \left(1 + \mathcal{L}(\widehat{h}_T, h^*)\right)\Delta_T,$$

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[p_t|\mathcal{F}_{t-1}\right] \leq 2\theta\left[2R^* + \max_{h\in\mathcal{H}_t}\left(2 + \mathcal{L}(h, \widehat{h}_{t-1}) + \mathcal{L}(h, h^*)\right)\Delta_{t-1}\right].$$

- θ: disagreement coefficient (Hanneke, 2007).
- More aggressive trimming of the version space.
- Slightly better generalization guarantee and label complexity.

IWAL and IWAL-D

Problem:

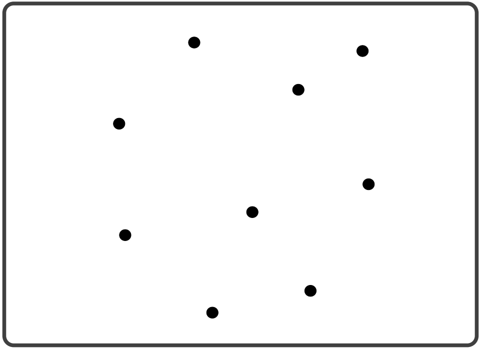
- Theoretical guarantees only hold for finite hypothesis sets.
- Need ϵ-cover to extend to infinite hypothesis sets.
- Expensive to construct ϵ-cover in practice.

Can we adaptively enrich the hypothesis set, with theoretical guarantees?

IZOOM: IWAL-D with Zooming-in

At round t,

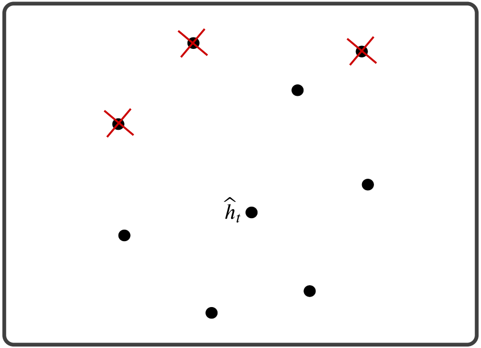
- Request label based on dis. of $(\mathcal{H}_t)$



IZOOM: IWAL-D with Zooming-in

At round t,

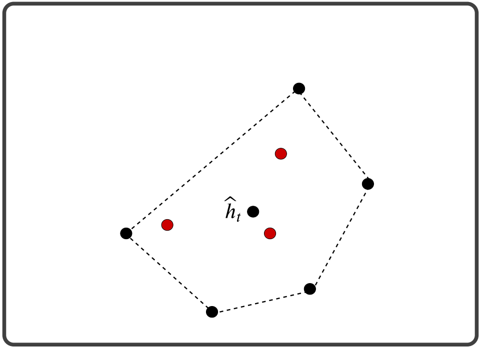
- Request label based on dis. of $(\mathcal{H}_t)$
- $\mathcal{H}'_{t+1} \leftarrow \text{Trim}(\mathcal{H}_t)$



IZOOM: IWAL-D with Zooming-in

At round t,

- Request label based on dis. of $(\mathcal{H}_t)$
- $\mathcal{H}'_{t+1} \leftarrow \text{Trim}(\mathcal{H}_t)$
- $\mathcal{H}''_{t+1} \leftarrow \textbf{Resample}(\mathcal{H}'_{t+1})$



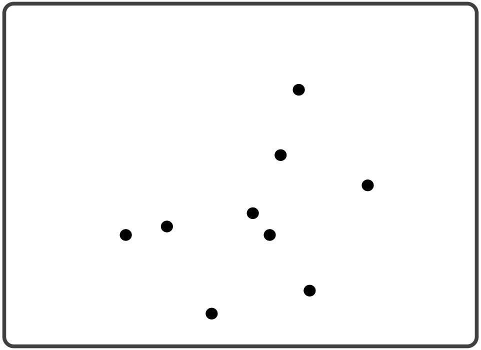
Resample$(\mathcal{H}'_{t+1})$: sample new $h \in \text{ConvexHull}(\mathcal{H}'_{t+1})$.

- E.g., random convex combination of $\widehat{h}_t$ and $h \in \mathcal{H}'_{t+1}$.

IZOOM: IWAL-D with Zooming-in

At round t,

- Request label based on dis. of $(\mathcal{H}_t)$
- $\mathcal{H}'_{t+1} \leftarrow \text{Trim}(\mathcal{H}_t)$
- $\mathcal{H}''_{t+1} \leftarrow$ **Resample**$(\mathcal{H}'_{t+1})$
- $\mathcal{H}_{t+1} \leftarrow \mathcal{H}'_{t+1} \cup \mathcal{H}''_{t+1}$



IZOOM vs. IWAL-D

Let $\mathbb{H}_t = \cup_{s=1}^{t} \mathcal{H}_t$, i.e. all the hypotheses ever considered up to time t. Let $h_t^* = \operatorname{argmin}_{h \in \mathbb{H}_t} R(h)$.

Theorem (IZOOM) With high probability,

$$R(\widehat{h}_T) \leq R_T^* + \big(1 + \mathcal{L}(\widehat{h}_T, h_T^*)\big)\Delta_T + O(\tfrac{1}{T}),$$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}\big[p_{t+1} | \mathcal{F}_t\big] \leq 2\theta_t \big[2R_t^* + \max_{h \in \mathcal{H}_{t+1}} \big(2 + \mathcal{L}(h, \widehat{h}_t) + \mathcal{L}(h, h_t^*)\big)\Delta_t\big] + O(\tfrac{1}{T}).$$

- $R_t^* = \min_{h \in \mathbb{H}_t} R(h)$ is smaller than $R^* = \min_{h \in \mathcal{H}_0} R(h)$.
- More accurate $\widehat{h}_T$, with fewer label requests.

Experiments

Tasks: 8 Binary classification datasets from UCI repository.

- ℓ: logistic loss rescaled to $[0, 1]$.

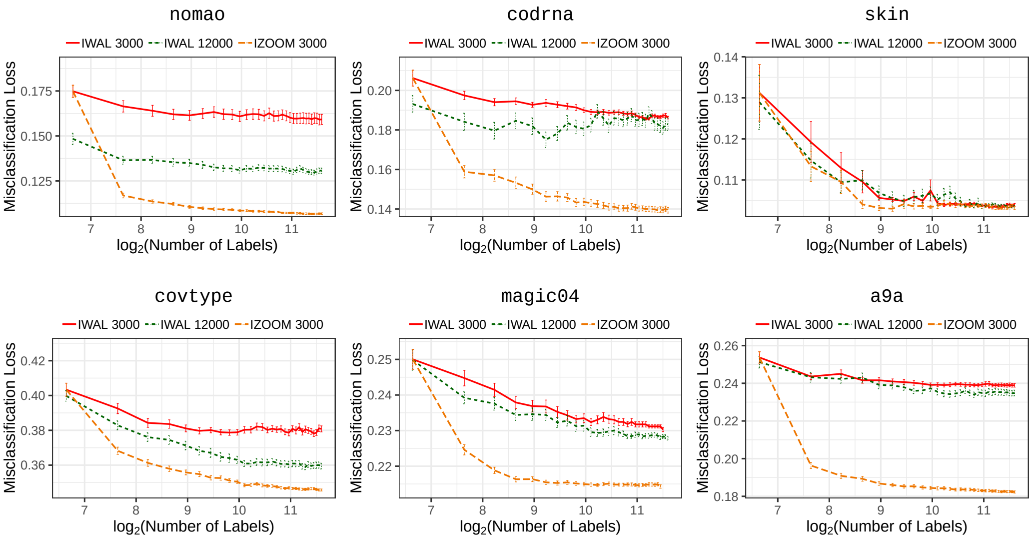
Baselines:

- IWAL with 3,000 hypotheses.
- IWAL with 12,000 hypotheses.
- IZOOM with 3,000 hypotheses.

Performance measure:

- 0-1 loss on test data vs. number of label requests.

Experiments



Conclusion

- Key introduction and role of disagreement graph.
- More favorable generalization and label complexity guarantees.
- Substantial performance improvements.
- Effective solutions for active learning.

Poster: Pacific Ballroom #265

KDD workshop (Alaska, August 2019) on Active Learning:
Data Collection, Curation, and Labeling for Mining and Learning.