Black-box Teaching: Improving Generalization and Convergence of Mentored Learner

Xiaofeng Cao

March 3, 2025

Organization

Active Learning

Definition

There are situations in which unlabeled data is abundant but manual labeling is expensive. In such a scenario, learning algorithms can actively query those unlabelled examples. This type of iterative supervised learning is called Active Learning.

Learning Scenarios

Active learning interactively prunes a pre-specified hypothesis class $\mathcal{H}$ to find one desired output, which improves the convergence of any learning algorithm using as few labels as possible. The setting is that the learner has access to a pool of unlabeled data and can query labels from human annotators, where the hypotheses are generated from a functional assumption, e.g., MLP, CNN, etc.

Example

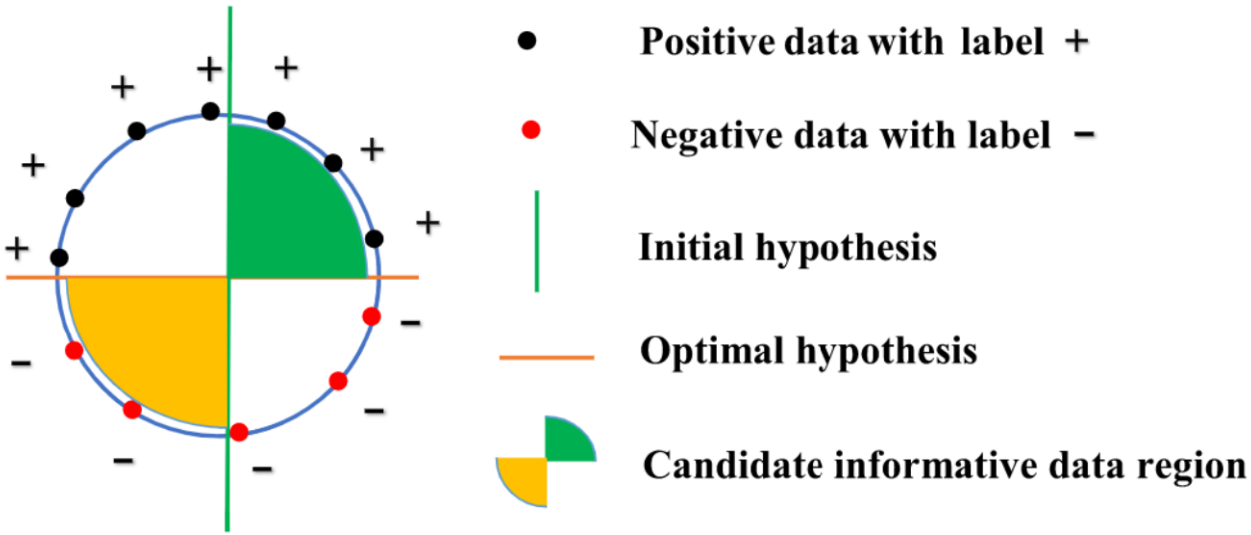

Figure: Binary classification issue over a unit sphere with a radius of R, where $+$ and $-$ denote the class labels. Hypothesis pruning strategy prunes the hypothesis set (reduce the number of candidate hypotheses, i.e., the diameters across the colored regions) via querying data distributed in the colored pool.

Research Problem

Assumption

The infinite hypothesis class exists the optimal hypothesis that may be incrementally updated from one passive initialization.

Problem

The previous theoretical results usually are infeasible for the convergence of the incremental updates in hypothesis class, that is h^* can not be easily obtained from these updates.

Proposal

We introduce a black-box teacher who can provide guidance for the learner but does not disclose any its cue.

Black-box teaching maintains a fair teaching scenario compared to those non-educated learners!

Solution:Black-box Teaching

Solution

The teacher gives a black-box hypothesis $h^{\mathcal{T}}$, where $h^{\mathcal{T}}$ is an acceptable approximation for the optimal hypothesis h^*. With $h^{\mathcal{T}}$, an active learner can easily replace the infeasible h^* and select those unlabeled data which maximize the disagreement of the feedback of teacher and learner. It is also called "Mentored Learning", which consists of 'How to teach' and 'How to learn'.

Implementation

Teaching Improves Hypothesis Pruning: We construct a new hypothesis pruning strategy by using teacher and student disagreement.
Self-improvement of Teaching: The self-improvement of teaching is firstly proposed to improve the initial teacher.

Example

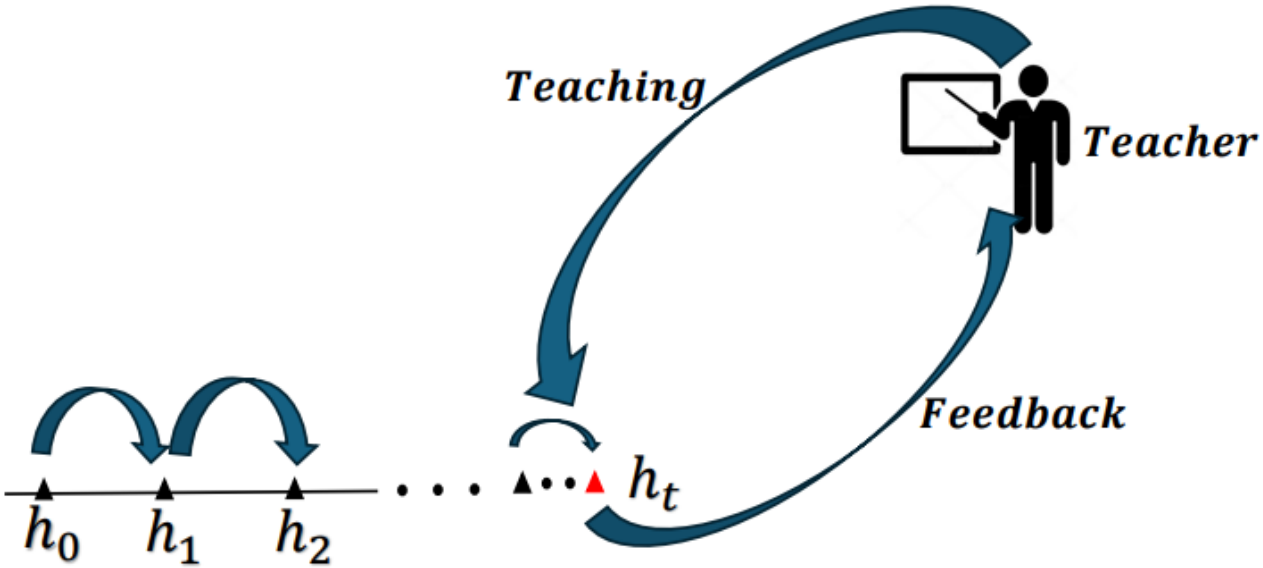

Figure: Hypothesis pruning of Learning vs. Teaching. Learning focuses on 'how to learn', and teaching focuses on 'how to teach'. With teaching, the general learning shifts into "Mentored Learning", which consists of the dual questions.

Importance Weighted Active Learning (IWAL)

Give a hypothesis set $H_1 = \mathcal{H}$, IWAL receives $x_t \in \mathcal{X}$ drawn i.i.d. according to $\mathcal{D}_{\mathcal{X}}$. For each round $t \in [T] = \{1, ..., T\}$, the algorithm decides whether to query the label of x_t and prunes H_t to H_{t+1}.

- **Query** At t-time, IWAL does a Bernoulli experiment Q_t with success probability: $p_t = \max_{h,h' \in H_t} \max_y |\ell(h(x_t), y) - \ell(h'(x_t), y)|$;
- **Hypothesis pruning** IWAL prunes H_t to H_{t+1} through $L_t(\widehat{h}_t)$ and an allowed slack $2\Delta_t$: $H_{t+1} = \left\{ h \in H_t : L_t(h) \leq L_t(\widehat{h}_t) + 2\Delta_t \right\}$, where $L_t(h) = \sum_{s=1}^{t} \frac{Q_s}{p_s} \ell\left(h\left(x_s\right), y_s\right), \widehat{h}_t = \underset{h \in H_t}{\operatorname{argmin}}\ L_t(h)$.

Learning Guarantees

Remark 1

Whether the optimal hypothesis h^* can usually be maintained in the candidate hypothesis set H_t is a necessary condition for the success of an active learning algorithm.

Remark 2

Two factors measure the quality of an active learning algorithm: 1) tighter bound on generalization error $R(\widehat{h}_T)$, where $\widehat{h}_T$ is the hypothesis returned by the algorithm after T rounds, and 2) tighter bound on label complexity τ_T, where τ_T is the total number of query labels within T rounds.

The active learning algorithm needs to satisfy the three factors, including 1) maintaining the optimal hypothesis, 2) tighter bound on generalization error, and 3) tighter bound on label complexity.

Organization

Assumption and Definition

Teaching Assumption

For any hypothesis class $\mathcal{H}$, assume that there exists a teacher hypothesis $h^{\mathcal{T}}$ which tolerates an error bias ϵ,

$$\mathcal{L}\left(h^*, h^{\mathcal{T}}\right) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \left[\max_y \left| \ell(h^*(x), y) - \ell(h^{\mathcal{T}}(x), y) \right| \right] < \epsilon.$$

Definition

For any hypothesis class $\mathcal{H}$, $h^{\mathcal{T}}$ is a teacher hypothesis that satisfies Teaching Assumption. If there exists a finite hypothesis class $\mathcal{H}^{\mathcal{T}}$ s.t. $h^{\mathcal{T}} = \mathop{\text{argmin}}_{h \in \mathcal{H}^{\mathcal{T}}} R(h)$ is the best-in-class hypothesis in $\mathcal{H}^{\mathcal{T}}$, then $\mathcal{H}^{\mathcal{T}}$ is called the **teacher-hypothesis-class** of $\mathcal{H}$.

Teaching Model

We use $\mathcal{F}^{\mathcal{T}}(\cdot) = \mathcal{L}(h^{\mathcal{T}}, \cdot)$ to denote a disagreement feedback function with operation $\mathcal{H} \to [0, 1]$.
Teacher: the teacher has a teaching hypothesis $h^{\mathcal{T}}$, which only can provide the disagreement feedback $\mathcal{F}^{\mathcal{T}}(\cdot)$ to the Learner.
Learner: the learner has a teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, which prunes $\mathcal{H}^{\mathcal{T}}$ by identifying the disagreement feedback $\mathcal{F}^{\mathcal{T}}(\cdot)$ with the Teacher.

Teaching-based Hypothesis Pruning

We still follow the pruning manner of IWAL to supervise the updates of the candidate hypothesis set, where the main difference is that we introduce a teacher $h^{\mathcal{T}}$ to control. Specifically, the slack constraint $2\Delta_t$ is tightened as $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$ s.t. $\mathcal{F}^{\mathcal{T}}(\widehat{h}_t) \leq 1$ by invoking the guidance of a teacher:

$$H_{t+1}^{\mathcal{T}} = \left\{ h \in H_t^{\mathcal{T}} : L_t(h) \leq L_t(\widehat{h}_t) + \left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t \right\}.$$

Pruning Speed

With a fast hypothesis pruning speed, the candidate hypothesis set $H_t^{\mathcal{T}}$ is shrunk rapidly, which reduces the learning difficulty, easily converting into $h^{\mathcal{T}}$. The primary determinant of pruning speed is the pruning slack term, i.e., $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$. There exists $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t \leq 2\Delta_t$, which means the teaching-based hypothesis pruning employs a tighter slack term to shrink $H_t^{\mathcal{T}}$ than IWAL.

Retain the Teacher

Theorem

For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, teaching an active learner runs on $\mathcal{H}^{\mathcal{T}}$. Given any $\delta > 0$, with a probability at least $1 - \delta$, for any $t \in \mathbb{N}^{+}$, the following inequality holds:

$$L_t(h^{\mathcal{T}}) - L_t(\widehat{h}_t) \leq \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right) \Delta_t.$$

Theorem shows that the teacher hypothesis satisfies the pruning rule with a high probability at any t-time. Thus teaching-based hypothesis pruning maintains the optimal hypothesis in the candidate hypothesis set with a high probability.

Learning Guarantees

Theorem

For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, teaching an active learner runs on $\mathcal{H}^{\mathcal{T}}$. Given any $\delta > 0$, with a probability at least $1-\delta$, for any $T \in \mathbb{N}^+$, the following holds:

1) the generalization error holds

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon;$$

2) if the learning problem has disagreement coefficient θ, the label complexity is at most

$$\tau_T \leq 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1})\right)\right) O(\sqrt{T}) + 2T\epsilon\Big).$$

Teaching an active learner has a tighter upper bound on the generalization error and label complexity than that of the error disagreement-based active learning!

The Problem of Teaching

For **Teaching Assumption**, if the teacher hypothesis is loosely approximated to the optimal hypothesis, i.e. ϵ is large, how do we guarantee the convergence of black-box teaching? We thus design a self-improvement of teaching strategy, which generates new hypotheses after each hypothesis pruning and determines whether to update the teacher.

New Hypotheses

Since hypothesis pruning is a process of shrinking the candidate hypothesis set, generating new hypotheses should not interrupt this process. Therefore, after pruning the hypothesis-set from $H_t^{\mathcal{T}}$ to H'_t at t-time, we generate new hypotheses $\tilde{h}$ from the convex hull of H'_t:

$$\tilde{h} = \sum_{j}^{m} \lambda_j h_j, h_j \in H'_t,$$

subjected to $\sum_j^m \lambda_j = 1, \lambda_j \in [0, 1], j = [m]$, where m denotes the size of H'_t. We generate n hypotheses to obtain $\widetilde{H}'_t = \{\tilde{h}_i; i \in [n]\}$ and combine it with H'_t as the candidate hypothesis set next time: $H_{t+1}^{\mathcal{T}} = H'_t \cup \widetilde{H}'_t$.

Self-improvement

The new hypotheses may perform better than the teacher, that is, $R(\tilde{h}) \leq R(h^{\mathcal{T}})$. The following theorem presents the determine condition of self-improvement:

Theorem

For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, teaching an active learner runs on $\mathcal{H}^{\mathcal{T}}$, where the sequence of candidate hypothesis sets satisfies $\mathrm{Conv}(H_{t+1}^{\mathcal{T}}) \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$ with $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. For any $t \in \mathbb{N}^+$, given any $\delta > 0$, with a probability at least $1 - \delta$, for any $\tilde{h} \in \widetilde{H}_t'$, the following inequality holds:

$$R(h^{\mathcal{T}}) - R(\tilde{h}) \geq L_t(h^{\mathcal{T}}) - L_t(\tilde{h}) - \left(1 + \mathcal{F}^{\mathcal{T}}(\tilde{h})\right) \Delta_t. \tag{1}$$

If $\beta_i^{(t)} = L_t(h^{\mathcal{T}}) - L_t(\tilde{h}_i) - \left(1 + \mathcal{F}^{\mathcal{T}}(\tilde{h}_i)\right) \Delta_t > 0$, i.e., $R(\tilde{h}_i) < R(h^{\mathcal{T}})$, i.e., $R(\tilde{h}_i) < R(h^{\mathcal{T}})$, then the teacher hypothesis is updated to $h^{\mathcal{T}} = \tilde{h}_i$.

Improvement of Teaching Performance

Self-improvement of teaching strategy obtain a teacher hypothesis sequence $\{h_1^{\mathcal{T}}, ..., h_T^{\mathcal{T}}\}$, where $h_t^{\mathcal{T}}$ denote the best-in-class hypothesis in $\bigcup_{k=1}^{t} H_t^{\mathcal{T}}$. The following Corollary gives the improvement of teaching performance.

Corollary

For any teacher-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, teaching an active learner runs on $\mathcal{H}^{\mathcal{T}}$. If the self-improvement of teaching is applied, let $\alpha_t = \max_i \beta_i^{(t)}$. Give any $\delta > 0$, with a probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$, holds $R(h_T^{\mathcal{T}}) \leq R(h_1^{\mathcal{T}}) - \sum_{t=1}^{T-1} \alpha_t$.

Corollary guarantees that under high probability, self-improvement of teaching can reduce the generalization error of the initial teacher by at least $\sum_{t=1}^{T-1} \alpha_t$.

Learning Guarantees

Theorem

For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, teaching an active learner runs on $\mathcal{H}^{\mathcal{T}}$. If the self-improvement of teaching is applied, given any $\delta > 0$, with a probability at least $1-\delta$, for any $T \in \mathbb{N}^+$, the following holds: 1) for any $t \in [T]$, holds $h_t^{\mathcal{T}} \in H_t^{\mathcal{T}}$; 2) the generalization error holds

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon_{T-1};$$

3) if the learning problem has disagreement coefficient θ, the label complexity is at most

$$\tau_T \leq 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1})\right) O(\sqrt{T}) + 2T\epsilon_{T-1}\right).$$

By generating new hypotheses, self-improvement of teaching tightens the approximation of the teacher hypotheses to the optimal hypothesis, which provides more favorable learning guarantees.

Organization

Teaching a White-box Learner

- Query:

$$p_t = \max_{h,h' \in H_t^{\mathcal{T}}} \max_{y} |\ell(h(x_t), y) - \ell(h'(x_t), y)|.$$

- Hypothesis pruning (shrinking the candidate hypothesis class):

$$H_{t+1}^{\mathcal{T}} = \left\{ h \in H_t^{\mathcal{T}} : L_t(h) \leq L_t(\widehat{h}_t) + \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right) \Delta_t \right\}.$$

- Self-improvement: generates new hypotheses according to

$$\tilde{h}_i = \lambda_i h_t^{\mathcal{T}} + (1 - \lambda_i)\widehat{h}_t,$$

updates the teacher according to

$$R(h^{\mathcal{T}}) - R(\tilde{h}) \geq L_t(h^{\mathcal{T}}) - L_t(\tilde{h}) - \left(1 + \mathcal{F}^{\mathcal{T}}(\tilde{h})\right) \Delta_t.$$

Teaching a Black-box Learner

- Query: $p_t = \max_y |\ell(h_t^{\mathcal{T}}(x), y) - \ell(\widehat{h}_{t-1}(x), y)|$.
  Before the update on $\widehat{h}_t$ at t-time, $\widehat{h}_{t-1}$ is used to approximate $\widehat{h}_t$.
- Hypothesis pruning (search with limited hypothesis radius):

$$L_{t-1}(\widehat{h}_t) \leq L_{t-1}(\widehat{h}_{t-1}) + \left(1 + \mathcal{F}_{t-1}^{\mathcal{T}}(\widehat{h}_{t-1})\right) \Delta_{t-1}$$

  If the above equation does not satisfy, then we backtrack the learner $\widehat{h}_t = \widehat{h}_{t-1}$.
- Self-improvement:

$$\tilde{h} = \lambda h_t^{\mathcal{T}} + (1 - \lambda)\widehat{h}_t,$$

  and update the teacher.

Organization

Experimental Setup

Empirical Studies:

- Whether the teaching-based hypothesis pruning of BTAL can prune the candidate hypothesis set faster than hypothesis pruning of IWAL.
- Whether self-improvement of teaching strategy of BTAL can reduce the generalization error of teacher hypothesis.

Real-world Studies:

- White-box learner
  - The error rate of IWAL, IWAL-D, and BTAL.
  - The number of query labels of IWAL, IWAL-D, and BTAL.
- Black-box learner
  - The accuracy of Random, MVR, ME, and BTAL$^+$.

Empirical Studies

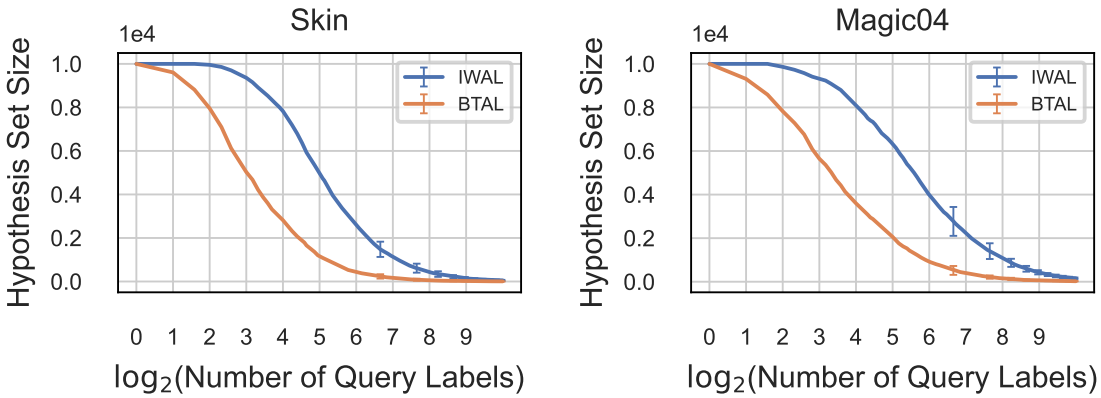

Figure: The size of the candidate hypothesis set of IWAL and BTAL vs. the number of query labels (log 2 scale).

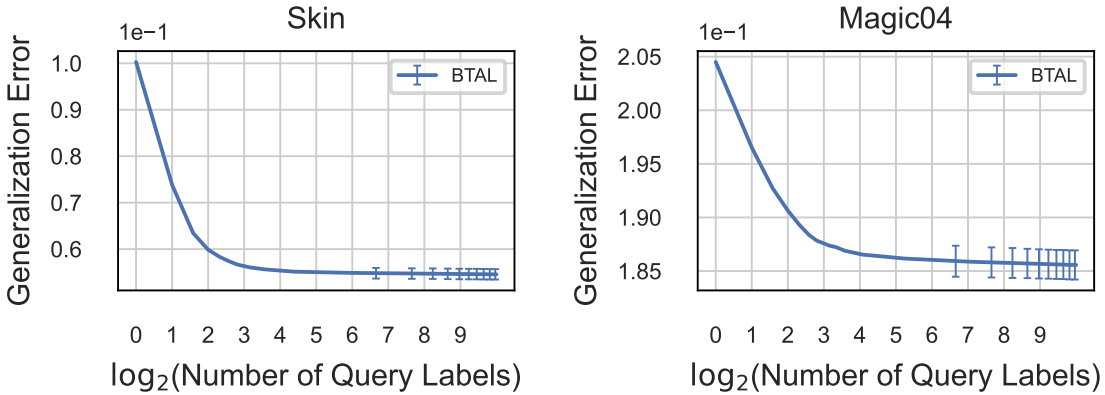

Figure: The generalization error of self-improving teacher of BTAL vs. the number of query labels (log 2 scale).

White-box learner

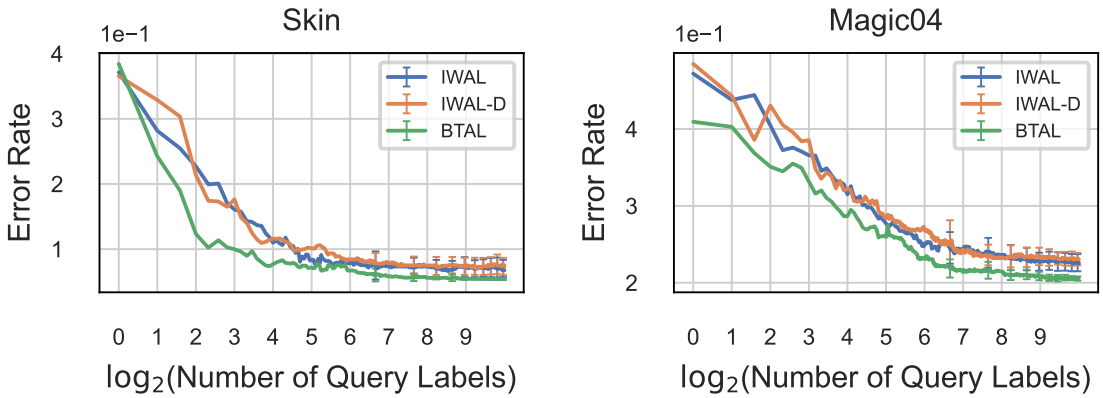

Figure: The error rate of IWAL, IWAL-D, and BTAL on the test dataset vs. the number of query labels (log 2 scale).

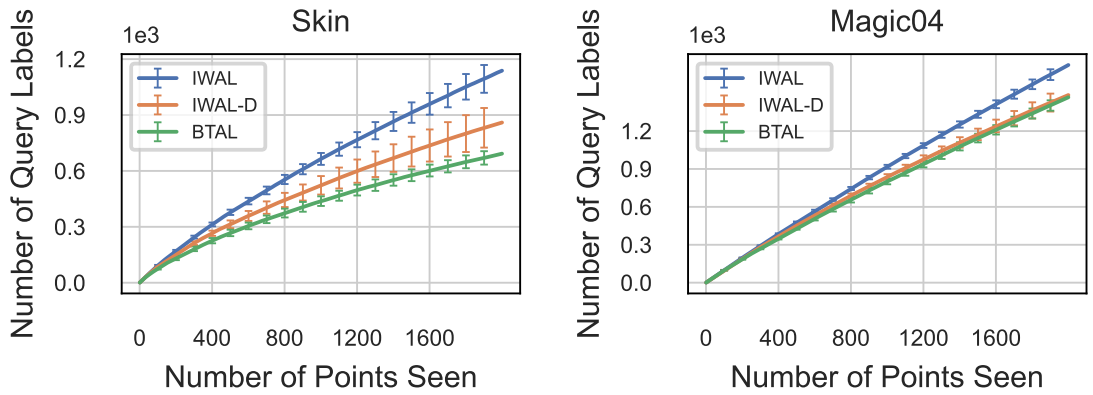

Figure: The number of query labels of IWAL, IWAL-D, and BTAL vs. the number of points seen.

Black-box learner

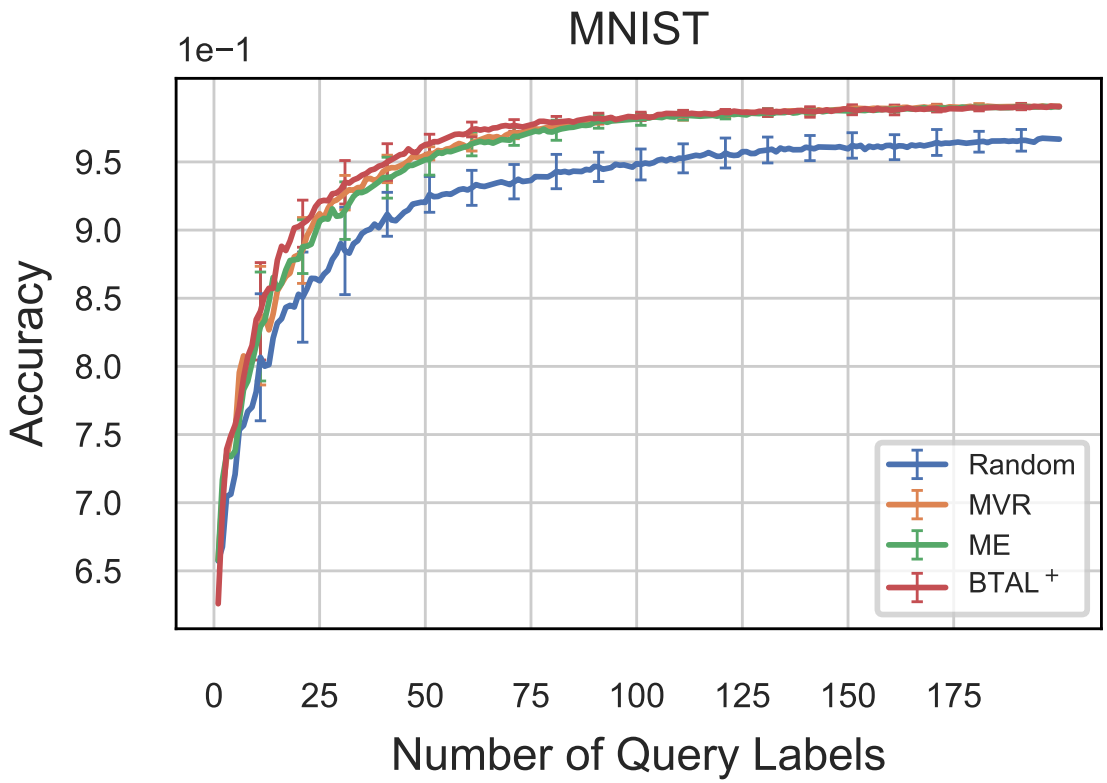

Figure: The accuracy of Random, MVR, ME, and BTAL$^+$ on the test dataset vs. the number of query labels.

Organization

Conclusion

- We introduce a new idea to the traditional active learning community: black-box teaching an active learner.
- We introduce a teacher hypothesis to improve the hypothesis pruning, which results in tighter bounds on the generalization error and label complexity.
- We present the self-improvement of teaching to improve the teaching performance.
- We present a black-box teaching-based active learning (BTAL) algorithm, which spends fewer annotations to converge, yielding more effective performance than those typical active learning baselines.

The End