Search Improves Label for Active Learning

Alina Beygelzimer BEYGEL@YAHOO-INC.COM
Yahoo Labs New York, NY

Daniel Hsu DJHSU@COLUMBIA.EDU
Department of Computer Science, Columbia University New York, NY

John Langford JCL@MICROSOFT.COM
Microsoft Research New York, NY

Chicheng Zhang CHICHENGZHANG@UCSD.EDU
Department of Computer Science and Engineering, University of California, San Diego La Jolla, CA

Abstract

We investigate active learning with access to two distinct oracles: LABEL (which is standard) and SEARCH (which is not). The SEARCH oracle models the situation where a human searches a database to seed or counterexample an existing solution. SEARCH is stronger than LABEL while being natural to implement in many situations. We show that an algorithm using both oracles can provide exponentially large problem-dependent improvements over LABEL alone.

Introduction Traditional active learning uses selective sampling with a LABEL oracle: the learning algorithm provides an unlabeled example to the oracle, and the oracle responds with a (possibly noisy) label. Using LABEL in an active learning algorithm is known to give (possibly exponentially large) problem-dependent improvements in label complexity, even in agnostic settings when no assumption is made about the labeling mechanism (e.g., Balcan et al., 2006; Hanneke, 2007; 2014).

A well-known deficiency of LABEL arises in the presence of rare classes in classification problems, frequently the case in practice (Attenberg and Provost, 2010). Class imbalance may be so extreme that simply *finding* an example from the rare class can exhaust the labeling budget. A good illustration of this is the problem of learning interval functions in $[0, 1]$. Any LABEL-only active learner needs at least $\Omega(1/\epsilon)$ LABEL queries to learn an arbitrary target interval with error at most ϵ (Dasgupta, 2005). As soon as any positive example from the interval is found, the sample complexity of learning intervals collapses to $O(\log(1/\epsilon))$— we can simply do a binary search for each of the end points. How can this observation be generalized and used effectively?

Searching for examples of the rare class to seed active learning is the way this hurdle is successfully dealt with in practice (Attenberg and Provost, 2010). Domain experts are often adept at finding examples of a class by various, often clever means. When building a hate speech filter, a simple web search can readily produce several positive examples. Sending a random batch of unlabeled examples to LABEL is unlikely to produce any positive examples at all.

In practice, it is also common to have counterexamples to a learned predictor. When monitoring the content stream filtered by the current hate speech filter, a human editor may spot an example of hate speech that seeped through the filter. The editors, using all search tools available to them, can be tasked with finding such counterexamples, interactively correcting the learning process.

We define a new oracle, SEARCH, that provides *counterexamples* to *version spaces*. Given a set of possible classifiers H mapping unlabeled points to labels, a *version space* $V \subseteq H$ is the subset of classifiers that are plausibly optimal. A *counterexample* to a version space is a labeled example which every hypothesis in the version space classifies incorrectly. When there is no counterexample to the version space, SEARCH returns $\perp$.

Why not counterexample a single classifier? Consider a learned interval classifier on the real line. A valid counterexample to this classifier may be arbitrarily close to an interval endpoint, yielding no useful information. SEARCH formalizes "counterexample away from decision boundary," avoiding this. Thus the learning algorithm must guide the search effort to parts of the space where it would be most effective.

How can a counterexample to the version space be used? We consider a nested sequence of hypothesis classes of increasing complexity, akin to Structural Risk Minimization (SRM) in passive learning (see, e.g., Vapnik, 1982; Devroye et al., 1996). When SEARCH produces a counterexample to the version space, it gives a proof that the current hypothesis class is too simplistic to solve the problem effectively. We show that this guided increase in hypothesis complexity results in radically lower LABEL complexity than directly learning on the complex space.

SEARCH can easily model the practice of seeding, discussed earlier. If the first hypothesis class in the sequence has just the constant -1 function, a seed example with label $+1$ is a counterexample to the version space.

We require that SEARCH always returns the label of the best predictor in the nested sequence. For many natural hypothesis sequences, the Bayes optimal classifier is eventually in the sequence. Unlike with LABEL queries where the labeler has no choice of what to label, here the labeler *chooses* a counterexample. If a human editor spots a piece of content that seeped through the filter and says that it is unquestionably hate speech, it likely is. These counterexamples should be consistent with the Bayes optimal predictor for any sensible feature representation.

Balcan and Hanneke (Balcan and Hanneke, 2012) define the Class Conditional Query (CCQ) oracle. Here, a query specifies a subset of unlabeled examples and a label, with the oracle returning one of the examples in the subset with the specified label, if one exists. While the definition of the CCQ oracle doesn't require the subset to be explicitly enumerated and finite, the motivation and the algorithms proposed in the paper do. In contrast, SEARCH has an implicit domain of all examples satisfying some filter, so search can more plausibly discover relevant counterexamples. The use of SEARCH in this paper is substantially different from the use of CCQ in (Balcan and Hanneke, 2012). Our motivation is to use SEARCH to assist LABEL, as opposed to using SEARCH alone. This is especially useful in the setting where the cost of SEARCH is significantly higher than the cost of LABEL (and class skew is only moderate)—we hope to avoid using SEARCH queries whenever it is possible to make progress using LABEL queries.

The Relative Power of Oracles As given by the intervals example, SEARCH can be exponentially more powerful than LABEL. Does it dominate LABEL?

Although SEARCH cannot always implement LABEL, we show that it is at least as effective in reducing the region of disagreement of the current version space. The clearest example is learning threshold classifiers $H := \{h_w : w \in [0,1]\}$ in the realizable case, where $h_w(x) = +1$ if $w \leq x \leq 1$, and -1 if $0 \leq x < w$. A simple binary search with LABEL achieves an exponential improvement in query complexity over passive learning. The agreement region of any set of threshold classifiers with thresholds in $[w_{\min}, w_{\max}]$ is $[0, w_{\min}) \cup [w_{\max}, 1]$. Since SEARCH is allowed to return any counterexample in the agreement region, there is no mechanism for forcing SEARCH to return the label of a particular point we want. However, this is not needed to achieve logarithmic query complexity with SEARCH: If binary search starts with querying the label of $x \in [0,1]$, we can query SEARCH(V_x), where $V_x := \{h_w \in H : w < x\}$ instead.

If SEARCH returns $\perp$, we know that the target $w^* \leq x$ and can safely reduce the region of disagreement to $[0, x)$. If SEARCH returns a counterexample $(x_0, -1)$ with $x_0 \geq x$, we know that $w^* > x_0$ and can reduce the region of disagreement to $(x_0, 1]$. This observation holds more generally: For any call to LABEL, we can always construct a call to SEARCH that achieves a no lesser reduction in the region of disagreement.

In the realizable setting where a zero-error classifier exists in the nested sequence, any call to SEARCH can be simulated with at most two calls to CCQ. Thus CCQ is at least as powerful and at least as difficult to implement in the realizable setting.

Our Results We propose and analyze a general purpose agnostic algorithm, LARCH, that uses SEARCH and LABEL (see (Beygelzimer et al., 2016) for details). As an implication of our general theorem in the case when the target hypothesis is a union of k^* non-trivial intervals in $[0,1]$, LARCH makes at most $k^* + \log(1/\epsilon)$ queries to SEARCH and at most $\tilde{O}((k^*)^3 \log(1/\epsilon) + (k^*)^2 \log^3(1/\epsilon))$ queries to LABEL, with high probability—an exponential improvement over any LABEL-based active learner.

In practical applications, it is critical to consider the relative cost of implementing the two oracles. We show that an amortized approach to explicitly trading off using LABEL and SEARCH yields an algorithm with a good guarantee on the total cost (Beygelzimer et al., 2016).

Discussion Our results demonstrate that SEARCH can significantly benefit LABEL-based active learning algorithms. Are there less powerful oracles that are as benefitial and still plausible to implement?

Another key question is computational efficiency. Can the benefits of SEARCH be provided in a computationally efficient general purpose manner? Attenberg and Provost showed that simply finding a set of examples of the rare class to seed supervised learning or LABEL-based active learning is already very powerful empirically (Attenberg and Provost, 2010). Can we do better with a truly interactive yet efficient algorithm?

References

Josh Attenberg and Foster J. Provost. Why label when you can search? alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 423–432, 2010.

Maria-Florina Balcan and Steve Hanneke. Robust interactive learning. In *COLT*, 2012.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *ICML*, 2006.

Alina Beygelzimer, Daniel J. Hsu, John Langford, and Chicheng Zhang. Search improves label for active learning. *CoRR*, abs/1602.07265, 2016. URL http://arxiv.org/abs/1602.07265.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005.

Luc Devroye, László Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, pages 249–278, 2007.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014. ISSN 1935-8237. doi: 10.1561/2200000037.

V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1982.