Active Learning from Weak and Strong Labelers

Chicheng Zhang CHZ038@ENG.UCSD.EDU
University of California, San Diego

Kamalika Chaudhuri KAMALIKA@ENG.UCSD.EDU
University of California, San Diego

Abstract

We study active learning with labels from multiple sources. Specifically, we consider the case where in addition to the usual labeling oracle, we are given a weak labeler. The weak labeler provides cheap labels which may be occassionally wrong, and our goal is to exploit it to reduce the number of queries made to the labeling oracle.

In this paper, we provide a learning theoretical formalization of this problem, and an active learning algorithm for our formalization. We provide an analysis of the number of high quality labels requested by our algorithm, and characterize when this algorithm can provide significant savings over using the high quality labels alone.

1. Introduction

Human interaction has the potential to make machine learning significantly easier by providing feedback to learning systems at all stages of learning. Feedback that can be provided is often complex in nature, and involves multiple users and different kinds of interaction from heterogeneous sources. While the theory of basic active learning, where labels are obtained interactively from a single source, has been well-developed, the effect of multiple annotators and different kinds of interaction is not as well understood, particularly in a formal theoretical setting.

In this paper, we take a step in this direction by considering a learning theoretic formalization of active learning when labels are obtained from heterogeneous sources. Specifically, we consider the case where in addition to the usual unlabeled data and a labeling oracle O, we have an extra weak labeler W. The labeling oracle O is an expert on the problem domain and provides high quality but expensive labels. The weak labeler W is cheap, but may provide incorrect labels on some inputs. In particular, querying the labeling oracle and the weak labeler at an x give labels from distributions $P_O(y|x)$ and $P_W(y|x)$ respectively. Our goal is to learn a classifier in a hypothesis class whose error with respect to the data labelled by the oracle O is low, while exploiting the weak labeler to reduce the number of queries made to O.

This setting models situations where high quality labels are expensive while low quality annotations may be readily obtained. For example, a physician's time may be valuable, where as lower quality labels may be obtained from medical residents. A key property of our formalization is that we allow the weak labeler to be correct in some regions of the input space and biased in others. This makes our formalization more realistic - medical residents do not diagnose every case incorrectly with some probability; rather, they can diagnose the easy cases or common cases, but make mistakes on rare cases or a certain type of cases.

We next provide an active learning algorithm that exploits the weak labeler. A natural approach is to learn a *difference classifier* to predict where the weak labeler differs from the labeling oracle, and then use a standard active learning algorithm which queries the weak labeler when this difference classifier predicts agreement. Our first observation is that this approach is statistically inconsistent, as false negative errors (that predict no difference when there is indeed a difference) lead to biased annotation. We address this problem by learning instead a *cost-sensitive difference classifier* that ensures that false negative errors rarely happen. Our second key observation is that as existing active learning algorithms usually query labels only in a localized region of space, it is sufficient to train the difference classifier restricted to this region and still maintain consistency. This leads to significant label savings as we can afford higher error (and thus require less labels) within this localized region. Combining these two ideas gives us an algorithm.

We analyze the label requirement of our algorithm, and characterize the conditions under which it provides label

Proceedings of the 31st International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

savings over simply using the labeling oracle. Our analysis shows that as expected we can achieve asymptotic label savings if the weak labeler agrees with the labeling oracle for a constant fraction of the examples close to the decision boundary. Moreover, when the target classification task is agnostic, the number of labels required to learn the difference classifier is of a lower order than the number of labels required for active learning; thus in realistic situations, learning the difference classifier adds a very small overhead to the total label requirement.

Related Work. There has been a fair amount of empirical work on active learning from multiple labelers (Donmez & Carbonell, 2008; Yan et al., 2011; 2012) to name a few; however theoretical formalization has been rare. (Urner et al., 2012) was the first to consider learning from weak teachers in a theoretical setting. In their model, the weak labeler is more likely to provide incorrect labels in heterogenous regions of space where similar examples have different labels. Their formalization is orthogonal to ours – while theirs is more natural in a non-parametric setting, ours is simpler and more natural for fitting classifiers in a hypothesis class. Our setting can also model situations where the weak labeler does not necessarily make mistakes close to the decision boundary – for example, when the data is clustered, and the weak labeler consistently labels one cluster incorrectly.

In a NIPS 2014 Workshop paper, (Malago et al., 2014) have also considered learning from strong and weak labelers; unlike ours, their work is in the online selective sampling setting, and applies only to linear classifiers and robust regression. In contrast, our strategy is completely general and applies to any classification problem.

Finally, there has been a large body of theoretical work on active learning (Balcan et al., 2009; Dasgupta, 2005; Dasgupta et al., 2007; Hanneke, 2007; Zhang & Chaudhuri, 2014; Balcan & Long, 2013; Beygelzimer et al., 2010). Our algorithm builds on disagreement-based active learning.

2. Preliminaries

The Model. We begin with our framework for actively learning from weak and strong labelers. In the standard active learning setting, we are given unlabelled data drawn from a distribution U over an input space $\mathscr{X}$, a label space $\mathscr{Y} = \{0,1\}$, a hypothesis class $\mathscr{H}$, and a labeling oracle O to which we can make interactive queries.

In our setting, we additionally have access to a labeling oracle W which we can query interactively. We call W the weak labeling oracle. Querying W is significantly cheaper than querying O; however, querying W generates a label drawn from a conditional distribution $P_W(y|x)$ which is not the same as the conditional distribution $P_O(y|x)$ of the oracle O. For simplicity we also assume that oracle O provides deterministic labels – for any x, $P_O(y|x)$ is either 0 or 1.

Let D be the data distribution over labelled examples such that: $\Pr_D(x,y) = \Pr_U(x)\Pr_O(y|x)$. Our goal is to learn a classifier h in the hypothesis class $\mathscr{H}$ such that with probability $\geq 1-\delta$ over the samples, we have:

$$\Pr_D(h(x) \neq y) \leq \min_{h^* \in \mathscr{H}} \Pr_D(h^*(x) \neq y) + \varepsilon$$

while making as few (interactive) queries to O as possible.

Some remarks on the model are in order. First, observe that W may disagree with the oracle O *anywhere* in the input space; this is unlike previous frameworks (Song et al., 2015) where labels assigned by the weak labeler are corrupted by random classification noise with a higher variance than the labeling oracle. Second, observe that to keep our model simple, we also do not assume that the mistakes made by the weak labeler are close to the decision boundary; however, we will see later that our algorithm will focus on mistakes made by W close to the decision boundary. Finally, we allow the oracle O to be *non-realizable* with respect to the target hypothesis class $\mathscr{H}$.

Background on Active Learning Algorithms. The standard active learning setting is very similar to ours, the only difference being that we have access to the weak oracle W.

There has been a long line of work on active learning (Balcan et al., 2009; Cohn et al., 1994; Dasgupta, 2005; Hanneke, 2007; Balcan & Long, 2013; Dasgupta et al., 2007; Beygelzimer et al., 2010; Zhang & Chaudhuri, 2014). The algorithms presented in this paper are based on a style of algorithms called *disagreement-based active learning (DBAL)*. The main idea behind DBAL is as follows. Based on the examples seen so far, the algorithm maintains acandidate set V_t of classifiers in $\mathscr{H}$ that is guaranteed with high probability to contain h^*, the classifier in $\mathscr{H}$ with the lowest error. Given a randomly drawn unlabeled example x_t, if all classifiers in V_t agree on its label, then this label is inferred. Otherwise, x_t is said to be in the *disagreement region* of V_t, and the algorithm queries O for its label. V_t is updated accordingly, and algorithm continues.

Later works (Dasgupta et al., 2007; Beygelzimer et al., 2010) have observed that it is possible to determine if an x_t is in the disagreement region of V_t without explicitly maintaining V_t. Instead, a labelled dataset S_t is maintained; the labels of the examples in S_t may be obtained by either querying the oracle or direct inference. To determine whether an x_t lies in the disagreement region, we perform two constrained ERM procedures; we constrain the classifier to output the label of x_t as 1 and -1 respectively, and then minimize the empirical risk over S_t. If the two classifiers obtained have similar training errors, then x_t lies in the disagreement region of V_t; otherwise its label can be safely inferred.

More Definitions and Notation. The error of a classifier h under a labelled data distribution Q is defined as: $\text{err}_Q(h) = \Pr_{(x,y)\sim Q}(h(x) \neq y)$; we use the notation $\widehat{\text{err}}(h,S)$ to denote its empirical error on a labelled data set S. We use the notation h^* to denote the classifier with the lowest error under D, where D is the target labelled data distribution.

Our active learning algorithm will implicitly maintain a $(1-\delta)$-*confidence set* for h^* throughout the algorithm. Given a set S of labelled examples, a set of classifiers $V(S) \subseteq \mathcal{H}$ is said to be a $(1-\delta)$-confidence set for h^* with respect to S if $h^* \in V$ with probability $\geq 1-\delta$ over the choice of S.

Given two classifiers h_1 and h_2 the disagreement between h_1 and h_2 under an unlabelled data distribution U, denoted by $\rho_U(h_1,h_2)$, is $\Pr_{x\sim U}(h_1(x) \neq h_2(x))$. Observe that the disagreements under U form a pseudometric over $\mathcal{H}$. We use $B_U(h,r)$ to denote a ball of radius r centered around h in this metric. The *disagreement region* of a set V of classifiers, denoted by $\text{DIS}(V)$, is the set of all examples $x \in \mathcal{X}$ such that there exist two classifiers h_1 and h_2 in V for which $h_1(x) \neq h_2(x)$.

3. Algorithm

Our algorithm is based on three key ideas, which we outline next.

A natural approach to our problem is to learn a *difference classifier* h^{df} in a hypothesis class $\mathcal{H}^{df}$ that predicts when W differs from O. This h^{df} is then used in conjunction with a standard active learning algorithm; on a label query, if h^{df} predicts a difference, then we query O, otherwise we query W. Our first key observation is that this procedure is statistically inconsistent if the region of difference between O and W is not realizable in $\mathcal{H}^{df}$. The reason is that false negative errors, that is, errors where the difference classifier incorrectly predicts agreement between W and O, are more pernicious than false positives, as they lead to biased annotation. We address this by instead learning a *cost-sensitive difference classifier*. Because we use cost-sensitive learning, we can impose a constraint that the false negative error of the difference classifier is very low, and then minimize the number of predicted positives (or disagreements between W and O) subject to this constraint. This ensures that the annotated data used by the active learning algorithm for the target classification task has diminishing bias, thus ensuring consistency.

The DBAL that builds the target classifier only makes label queries in the disagreement region $\text{DIS}(V)$ of V, the current $(1-\delta)$-confidence set for h^*. Our second key contribution is to exploit this fact to train the difference classifier restricted to $\text{DIS}(V)$. This procedure trivially maintains consistency. Additionally it provides label savings because we only need to train the difference classifier restricted to this region to an excess error of $O(\varepsilon/\phi)$, where ε is the target error and ϕ is the probability mass of $\text{DIS}(V)$. If we trained the difference classifier over the entire space, we would instead require an excess error of $O(\varepsilon)$, which would require more labeled examples.

A problem with this approach however is that as V is being constantly updated by Agnostic CAL, the disagreement region $\text{DIS}(V)$ is also constantly changing. Our third key contribution is to address this by deriving a novel epoch-based version of Agnostic CAL. We select the epochs such that after epoch k, the excess error of the target classifier is $\varepsilon_k \approx 1/2^k$. At the end of each epoch, V is updated, and a fresh difference classifier restricted to the disagreement region of the updated V is trained. An additional issue is how to determine the number of labeled examples to obtain in each epoch; if $\text{err}_D(h^*) = \nu$, then achieving an excess error of ε requires $\tilde{O}(d\nu/\varepsilon^2)$ labeled examples, where d is the VC dimension of the hypothesis class $\mathcal{H}$. However, as ν is unknown in advance, we cannot determine this number. We resolve this by using a doubling procedure that adaptively determines the number of labeled examples required to reach the target error at each epoch.

Main Algorithm. Our main algorithm combines these three key ideas together, and is described in Algorithm 1. Like certain versions of CAL, our algorithm implicitly maintains the $(1-\delta)$-confidence set by maintaining a labeled dataset S_k. In each epoch, the algorithm proceeds in three steps – (a) identify the current disagreement region and infer the labels that can be inferred (b) train a cost-sensitive difference classifier restricted to the disagreement region and (c) adaptively does active learning to update S_k, using the difference classifier to determine whether the weak or strong oracle should be queried. At the end of the last epoch, the algorithm returns a classifier in $\mathcal{H}$ for the target classification task.

Note that the procedure LEARN used by these algorithms is a constrainted empirical risk minimizer (ERM), of the form used by (Dasgupta et al., 2007; Beygelzimer et al., 2010). Given a hypothesis class H, a labelled dataset S and a set of constraining labelled examples C, $\text{LEARN}_H(C,S)$ returns a classifier in H that minimizes the empirical error on S subject to the constraint that $h(x_i) = y_i$ for each $(x_i, y_i) \in C$.

Algorithm 1 Active Learning Algorithm from Weak and Strong Labelers

```
Input: Unlabelled Distribution $U$, Target Error $\epsilon$, Labeling oracle $O$, Weak oracle $W$, hypothesis class $\mathcal{H}$, hypothesis class for difference classifier $\mathcal{H}^{df}$.
Output: Classifier $\hat{h}$ in $\mathcal{H}$.
Initialize: Initial error $\epsilon_0 = 1/2$. Total number of epochs $k_0 = \lceil \log \frac{1}{\epsilon} \rceil$.
Draw $n_0 = \tilde{O}(d/\epsilon_0^2)$ examples and query $O$ for their labels, forming $S_0$.
for $k = 1, 2, \ldots, k_0$ do
    Set Target error $\epsilon_k = \epsilon_{k-1}/2$.
    Set $n_k = \tilde{O}(d/\epsilon_k^2)$. Draw $n_k$ unlabelled examples to form $T_k$.
    # Identify Disagreement Region and Infer Labels
    $A_k \leftarrow$ subset of $T_k$ that lies in the disagreement region. $C_k \leftarrow$ rest of $T_k$, along with their inferred labels.
    # Train Difference Classifier
    $\hat{h}_k^{df} \leftarrow$ Train difference classifier on input data $A_k$, oracles $W$ and $O$, target false negative error $\epsilon_k / 2\mathbb{P}_{T_k}(A_k)$.
    # Active Learning using Difference Classifier
    Adaptively draw $m_k \approx \tilde{O}(\frac{\mathbb{P}_{T_k}(A_k) d\nu}{\epsilon_k^2})$ examples randomly from $A_k$. For each example $x$, if $\hat{h}_k^{df}(x) = 1$, then query $O$ for its label; else query $W$. Add these labeled examples to $S_k$.
end for
return $\hat{h} \leftarrow \text{LEARN}_{\mathcal{H}}(\emptyset, S_{k_0})$.
```

4. Performance Guarantees

We now provide analytical guarantees on the performance of Algorithm 1. Our guarantees are based on the following simple assumption on the difference classifier.

Assumption 1. *For any $r, \eta > 0$, there exists an $f_{\eta,r} \in \mathcal{H}^{df}$ with the following properties:*

$$\Pr(f_{\eta,r}(x) = -1 \wedge x \in DIS(B_U(h^*, r)) \wedge y_O \neq y_W) \leq \eta \tag{1}$$

$$\Pr(f_{\eta,r}(x) = 1 \wedge x \in DIS(B_U(h^*, r))) \leq \alpha(r, \eta) \tag{2}$$

(1) states that there exists an $f_{\eta,r}$ in $\mathcal{H}^{df}$ with low false negative error in the disagreement region of $B_U(h^*, r)$; this assumption is trivially satisfied if O is a deterministic labeler and if $\mathcal{H}^{df}$ includes the constant classifier that always predicts 1. (2) in addition states that the number of positives predicted by this classifier is not very high. We note $\alpha(r, \eta) \leq \Pr(\text{DIS}(B_U(h^*, r)))$ always; we will obtain a performance gain when $\alpha(r, \eta)$ is significantly less.

We next show that Algorithm 1 is statistically consistent – namely, that it achieves its target excess error with high probability.

Theorem 1 (Consistency)**.** *Let h^* be the classifier that minimizes the error with respect to D. If Assumption 1 holds, then with probability $\geq 1 - \delta$, the classifier $\hat{h}$ output by Algorithm 1 satisfies: $err_D(\hat{h}) \leq err_D(h^*) + \epsilon$.*

The label complexity of standard DBAL algorithms are measured in terms of the disagreement coefficient. The disagreement coefficient $\theta(r)$ at scale r is defined as: $\theta(r) = \sup_{r' \geq r} \frac{\Pr_U(\text{DIS}(B_U(h^*, r')))}{r'}$; intuitively, this measures the rate of shrinkage of the disagreement region with the radius of the ball $B_U(h^*, r)$.

Theorem 2 (Label Complexity)**.** *Let d be the VC dimension of $\mathcal{H}$ and let d' be the VC dimension of $\mathcal{H}^{df}$. If Assumption 1 holds, and if the error of the best classifier in $\mathcal{H}$ on D is at most ν, then with probability $\geq 1 - \delta$, the following hold:*

1. *The number of label queries made by Algorithm 1 to the oracle O in epoch k at most:*

$$m_k = \tilde{O}\Big(\frac{d(2\nu + \epsilon_{k-1})\alpha(2\nu + \epsilon_{k-1}, \epsilon_{k-1}/64)}{\epsilon_k^2} + \frac{d' \Pr(DIS(B_U(h^*, 2\nu + \epsilon_{k-1})))}{\epsilon_k} \Big) \tag{3}$$

2. *The total number of label queries made by Algorithm 1 to the oracle O is at most:*

$$\tilde{O}\Big(\sup_{r \geq \epsilon} \frac{\alpha(2\nu + r, r/64)}{2\nu + r} \cdot d \left(\frac{\nu^2}{\epsilon^2} + 1 \right) + \theta(2\nu + \epsilon) d' \left(\frac{\nu}{\epsilon} + 1 \right) \Big) \tag{4}$$

Some remarks on Theorem 2 are in order. The first terms in (3) and (4) represent the number of labels required for learning the target classifier, and second terms represent the overhead incurred to learn the difference classifier. We observe that provided $d \approx d'$, the second term is a *lower order term* in the more realistic agnostic case (when $\nu > 0$) and is of the same order when the target classifier is realizable; thus fitting the difference classifier does not incur a high overhead. We believe that this is because the cost-sensitive learning problem is realizable and we train the difference classifier in increasingly smaller regions of the instance space.

Second, since $\sup_{r \geq \epsilon} \frac{\alpha(2\nu + r, r/64)}{2\nu + r} \leq \theta(2\nu + \epsilon)$, the worst case asymptotic label complexity is the same as that of standard disagreement-based active learning. This label complexity may be considerably better however if $\sup_{r \geq \epsilon} \frac{\alpha(2\nu + r, r/64)}{2\nu + r}$ is significantly less than the disagreement coefficient. As we expect, this will happen when the region of difference between W and O restricted to the disagreement regions is small, and this region is well-modeled by the difference hypothesis class $\mathcal{H}^{df}$.

References

Balcan, M.-F. and Long, P. M. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, 2009.

Beygelzimer, A., Hsu, D., Langford, J., and Zhang, T. Agnostic active learning without constraints. In *NIPS*, 2010.

Cohn, D. A., Atlas, L. E., and Ladner, R. E. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.

Dasgupta, S. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

Dasgupta, S., Hsu, D., and Monteleoni, C. A general agnostic active learning algorithm. In *NIPS*, 2007.

Donmez, P. and Carbonell, J. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *CIKM*, 2008.

Hanneke, S. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

Malago, J., Cesa-Bianchi, N., and Renders, J. Online active learning with strong and weak annotators. In *NIPS Workshop on Learning from the Wisdom of Crowds*, 2014.

Song, S., Chaudhuri, K., and Sarwate, A. D. Learning from data with hetergeneous noise using sgd. In *Proc. of AISTATS*, 2015.

Urner, Ruth, Ben-David, Shai, and Shamir, Ohad. Learning from weak teachers. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, pp. 1252–1260, 2012.

Yan, Yan, Rosales, Rómer, Fung, Glenn, and Dy, Jennifer G. Active learning from crowds. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1161–1168, 2011.

Yan, Yan, Rosales, Rómer, Fung, Glenn, Farooq, Faisal, Rao, Bharat, and Dy, Jennifer G. Active learning from multiple knowledge sources. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, pp. 1350–1357, 2012.

Zhang, C. and Chaudhuri, K. Beyond disagreement-based agnostic active learning. In *Proc. of NIPS*, 2014.