Active Learning with Partially Labeled Data via Bias Reduction

Minoo Aminian
Computer Science Dept.
SUNY Albany, Albany
NY, USA, 12222
minoo@cs.albany.edu

Ian Davidson
Computer Science Dept.
SUNY Albany, Albany
NY, USA, 12222
davidson@cs.albany.edu

Abstract
With active learning the learner participates in the process of selecting instances so as to speed-up convergence to the “best” model. This paper presents a principled method of instance selection based on the recent bias variance decomposition work for a 0-1 loss function. We focus on bias reduction to reduce 0-1 loss by using an approximation to the optimal Bayes classifier to calculate the bias for an instance. We have applied the proposed method to naive Bayes learning on a number of benchmark data sets showing that using this active learning approach decreases the generalization error at a faster rate than randomly adding instances and converges to the optimal Bayes classifier error obtained from the original data set.

1 Introduction and Motivation
An active learner seeks instance(s) to maximize its performance or speed up the process of learning; a passive learner, however, receives instances from the provider or randomly draws out of a distribution. Naturally, active learning can be invaluable where we have limited amount of labeled data, and labeling instances are expensive or difficult. To minimize the generalization error associated with the learner we can decompose the error into bias and variance. Recently several authors have proposed corresponding decompositions for zero-one loss and here, we use the bias-variance decomposition proposed by (Domingos, 2000) and use the bias associated with each example to guide active learning.

The main contribution of this paper is using an approximation to Bayesian optimal classifier (BOC) as a guide to label the unlabeled instances for active learning. We use bootstrap model averaging technique to approximate the BOC (Davidson, 2004) and use active learning to select the instances with the maximum bias with respect to this optimal classifier. We begin the paper by bias variance decomposition for classification loss; next we explain how we have used this interpretation of bias as a guide for active learning. We carry out a number of experiments to verify the idea, and finally discuss the associated issues and future work.

2 Bias Variance Decomposition for Classification Loss
The goal of learning can be stated as producing a model with the smallest possible loss. Suppose we have a training set of pairs \{(x_i, t_i), i = 1, ..., n\}, and a model which produces an estimate \(y_i\) of the true value \(t_i\) for \(x_i\). The zero-one loss is zero if \(y_i = t_i\), and is one otherwise. Based on the definitions in (Domingos, 2000), the optimal prediction for a specific example \(x_i\) is the lowest loss prediction irrespective of our model or formally:

\[
y_\ast = \arg \min_{y} E_{t_i}[L(t_i, y)]
\]

And the main prediction, \(y_m\), for the specific value \(x\), a specific loss function \(L\) and a set of training sets \(D\), is defined to be the value that differs least from all other predictions \(y\) according to \(L\).

\[
y_m = \arg \min_{y} E_{D}[L(y, y')] \tag{2}
\]

Then bias of a learner on a specific example is defined as:

\[
B(x_i) = L(y_\ast, y_m) \tag{3}
\]

And the variance of the learner on an example as:

\[
V(x_i) = E_{D}[L(y, y_m)] \tag{4}
\]

Noise is defined as:

\[
N(x_i) = E_{D}[L(t, y_m)] \tag{5}
\]

And based on all these the following decomposition holds:

\[
E_{D,t_i}[L(t(x_i)) = c_1 E_{D}[L(t, y_\ast)] + L(y_\ast, y_m) + c_2 E_{D}[L(y_m, y_\ast)] = c_1 N(x_i) + B(x_i) + c_2 V(x_i)
\]

In which \(c_1\) and \(c_2\) are multiplicative factors which will take different values for different loss functions.

3 Active Learning for Bias Reduction
Based on the definition of bias that we mentioned before (3), we can measure bias of an instance relative to the optimal classifier and, select new instance to be added to the training set such that it will minimize the expected
value of loss over the entire domain. We can state the procedure formally in the following algorithm in table 1.

Given:
- a set $D_L$ of labeled training instances
- a set $U$ of unlabeled instances
- an optimal classifier $O$ trained on $D_L$

Loop for $k$ iterations:
- Pick a random pool $R$ from the set $U$.
- Use $D_L$ to train the classifier $T_m$.
- Allow $T_m$ to label instances in $R$.
- Allow $O$ to label instances in $R$.

Let $S_1$ be the set of instances from $R$ on which $T_m$ makes the predictions such that

$$\forall x \in S_1: T_m(x) \neq O(x)$$

- $\forall x \in S_1$ label the instances with $O$ and add them to the training data set $D_L$.

| Table 1 : Active learning to reduce bias |

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Loss Reduction (Active)</th>
<th>Loss Reduction (Passive)</th>
<th>Accuracy BOC (Initial sample size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cpu</td>
<td>2.12</td>
<td>1.50</td>
<td>82.02 (20)</td>
</tr>
<tr>
<td>Wine</td>
<td>2.34</td>
<td>1.56</td>
<td>84.33 (18)</td>
</tr>
<tr>
<td>Labor</td>
<td>7.66</td>
<td>4.3</td>
<td>72.16 (6)</td>
</tr>
<tr>
<td>Iris</td>
<td>1.78</td>
<td>1.12</td>
<td>92.41 (22)</td>
</tr>
<tr>
<td>Breast</td>
<td>0.7</td>
<td>0.1</td>
<td>96.7 (37)</td>
</tr>
<tr>
<td>Auto.</td>
<td>1.83</td>
<td>1.26</td>
<td>63.12 (40)</td>
</tr>
</tbody>
</table>

4 Computing Bias

Formally, the calculation that BOC is performing is:

$$\arg\max_{\theta\epsilon\Theta} \int P(y_i,x|\theta)P(\theta|D)d\theta$$

However, integration over the entire model space which could be high dimensional is a very time consuming process, so we have to come up with an approximation to BOC. We can use bootstrap model averaging which is discussed in detail in (Davidson, 2004) to approximate the performance of the BOC.

Using this approach, from the labeled data we can create multiple bootstrap samples to have the bootstrap averaging to estimate BOC. Next we can pick a random pool from the unlabeled data and run both the classifiers on that pool, then we can compute bias based on equation (1) for that random pool and select instances for which the “optimal model” predicts a different label (biased instances). Next we label the biased instances and add them to the training data to train a new classifier.

5 Empirical Results

We applied our method to reduce zero-one loss in a series of experiments with Naive Bayes classifier. We used a number of datasets from the UCI repository (cpu-performance, wine, labor, iris, breast, and auto.) and recorded the expected loss over the test data before and after applying the algorithm.

In these experiments, the BOC estimator was trained only once on the small labeled data set, but our model was retrained in each iteration after adding the newly labeled data (from the BOC) to the training set until our learner’s accuracy is that of the BOC. An example of a performance graph is shown in Figure 1. Table 2 shows the comparative performance of passively choosing instances and actively choosing the same number of instances.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Loss for cpu-performance dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BOC Loss</td>
</tr>
<tr>
<td></td>
<td>Active Learning Loss</td>
</tr>
<tr>
<td></td>
<td>Passive Learning Loss</td>
</tr>
</tbody>
</table>

Figure 1: Reduction of 0-1 loss for Cpu

6 Discussion and Future Work

As mentioned before the role of the optimal classifier is crucial in this analysis as it is used to label instances which our model has low confidence in.

The paper used a BOC approximation (Davidson, 2004) and we will investigate other approximations to the optimal classifier in future for other learners. We also plan to explore how to actively choose instances to reduce variance.

References
