



Active Learning with Scarcely Labeled Data via Bias Variance Reduction

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

Abstract.

In many occasions in real life, we are faced with the problem of classification of partially labeled data, or semi-supervised learning. We consider the special case of scarcely labeled data or when the labeled data is insufficient, and present a principled method which implements active learning in scarcely labeled data to enhance the performance of the learner. This method is based on the recent bias variance decomposition work for a 0-1 loss function. We focus on bias and variance reduction to reduce 0-1 loss by first selecting a random pool from the unlabeled data, and then using the most-informative instances from that pool to reduce the variance, bias, and thereby overall loss of the learner via active learning. Our empirical results show that this technique can decrease the loss of the learner significantly.

Keywords: *Active learning, Bias variance reduction, Semi-supervised learning, Scarcely-labeled data, 0-1 loss.*

1. Introduction and Motivation

An active learner seeks instance(s) which maximize its performance. A passive learner, however, receives instances from the provider or randomly draws out of a distribution. Active learning can be used in different types of learning environments.

Naturally, active learning can be invaluable where we have limited amount of labeled data, and labeling instances are expensive or difficult. This situation is common in the learning from labeled and unlabeled data [12] where the learner is presented with a large pool of unlabeled data and a small set of labeled data. In active semi-supervised learning, first we train a classifier from the labeled data, then from the set of unlabeled data we select instances that if labeled and added to the labeled data will be most beneficial to the performance of the learner. Next we ask a human or an oracle to label these instances and add them to the labeled data to retrain the classifier. This procedure can be repeated, and our goal is

to label as little data as possible to achieve a certain performance. In contrast, in passive semi-supervised learning, we can train a classifier using the initial labeled data as well as the unlabeled data, and the labels for the unlabeled data are typically assigned based on the current state of the classifier. The new augmented labeled data is used to retrain the classifier [11].

Generalization error is a metric to measure the performance of the learner but since true future error rates are unknown, the active learning algorithm of Roy and McCallum [18] attempts to reduce the generalization error probability in selecting its next example to be labeled.

Another approach in reducing the generalization error, is to decompose the generalization error and attempt to reduce its components. Bias and variance are two such components. Bias variance decomposition in its original formulation for squared error loss [13], though useful, is not readily applicable to *classification* problems with 0-1 loss functions. Recently several authors have proposed corresponding decompositions for zero-one loss [9][19][10]. In this paper, we use the bias-variance decomposition proposed by Domingos [2] for variable misclassification costs. In his definition, the bias of a learner on an example is the loss incurred by the main prediction relative to the optimal prediction while the variance on an example is the average loss incurred by predictions relative to the main prediction. The main prediction of a learner on an example is the most frequent prediction that the learner makes. The optimal prediction for a model space is the prediction that minimizes the expected value of loss (risk) taken over all possible values of true classes weighted by their probabilities given x , and is irrespective of a learner. This definition of bias variance decomposition is applicable to any loss function; and we use the variance and bias associated with each example to guide active learning.

Sometimes features describing the data are redundant for a given task, so that we can classify an instance using only one set of features or another. Blum and Mitchell proposed an algorithm called co-training [1] which is a

semi-supervised algorithm applicable to problems with two separate but redundant views of the data. This algorithm, though successful, is not applicable to problems with no obvious feature splits. As a result, researchers have begun to investigate the co-training procedures that use two different learning algorithms in lieu of the multiple views required by standard co-training [4].

In this paper, we use two different learning algorithms to obtain two different views and learn from their differences; we chose naïve Bayes as the base learner in our method.

We begin the paper by bias variance decomposition for classification loss; next we explain how we select the most informative instances, and use these instances to reduce the bias and variance of the learner via active learning. Then we will discuss the details of our algorithm followed by empirical results to verify the idea, and finally discuss the associated issues and future work.

2 Preliminaries

One of the measures of quality of a learner in classification is the amount of loss. A loss function measures the cost associated with the misclassification. The goal of learning can be stated as producing a model with the smallest possible loss, i.e. a model that minimizes loss over all examples, with each example weighted by its probability [2].

Bias variance decomposition is complicated in classification problems, since here we are interested in calculation of misclassification loss/error rather than the squared error, but by considering the expected value we can recast the classification loss into the frame of mean squared error.

First we define commonly used loss function. Suppose we have a training set of pairs $\{(x_i, t_i), i = 1, \dots, n\}$, and a model which produces an estimate y_i for x_i . Let t be the true value (most probable value) of the estimated variable for the test example x . It is important to realize that Domingos treats the instance labeling using a uniform noise model, hence there is a chance of mislabeling of an instance. Commonly used loss functions are absolute loss in which $L(t, y) = |t-y|$, squared loss ($L(t, y) = (t-y)^2$) and the zero-one loss. The zero-one loss is defined to be zero if $y = t$, and one otherwise.

There are many definitions to bias and variance in classification, for example to understand Domingos definition, we need to define the notions of optimal prediction and main prediction. The optimal prediction for a specific example x is the lowest loss prediction irrespective of our model or formally:

$$y_* = \arg \min_{y_i} E_t[L(t, y_i)] \quad (1)$$

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 .

$$y_m = \arg \min_{y'} E_D[L(y_i, y')] \quad (2)$$

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

$$B(x_i) = L(y_*, y_m) \quad (3)$$

And the variance of the learner on an example as:

$$V(x_i) = E_D[L(y_i, y_m)] \quad (4)$$

Noise is defined as:

$N(x_i) = E_t[L(t, y_*)]$, and based on all these the following decomposition holds:

$$E_{D,t}[L(t, y_i)] = c_1 E_t[L(t, y_*)] + L(y_*, y_m) + c_2 E_D[L(y_m, y_i)] \quad (5)$$

or:

$$E_{D,t}[L(t, y_i)] = c_1 N(x_i) + B(x_i) + c_2 V(x_i) \quad (6)$$

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

Therefore, given a training set in D if a learner predicts y for an example x for which the true prediction is t , the average loss over all the examples will be:

$$E_{D,t,x}[L(t, y_i)] = E_x [c_1 N(x_i)] + E_x [B(x_i)] + E_x [c_2 V(x_i)] \quad (7)$$

Based on these definitions, we propose a method that reduces the expected value of bias and variance of the learner via active learning. We will explain this method in the following section.

3 Active Learning for Bias Variance Reduction

We focus on learning from labeled and unlabeled data problem, also known as semi-supervised learning. In this situation the learner has a relatively small collection of labeled data points $D_l = \{(x_1, y_1) \dots (x_m, y_m)\}$ and a large amount of unlabeled data $D_u = \{x_{m+1}, \dots, x_{n+m}\}$. It is assumed that both labeled and unlabeled data are drawn from the same population, but it is too costly or difficult to label every single instance so only a small subset are given labels.

The aim is to learn a function to predict the dependent variable y from the independent variables x . Interestingly it is possible to obtain a more accurate classifier by using the unlabeled and labeled data than by just using the labeled data [12].

To reach our goal of increasing the classification accuracy while we have a small set of labeled data, and a large set of unlabeled data, we select active learning approach.

Active learning approach assumes that the learner has some way of labeling any instance but the process is possibly expensive to do. Our challenge then, is to determine which unlabeled instances, if labeled and added to the training set, would be most-informative or would improve the classifier the most. These are the

