A Unified Bias-Variance Decomposition and its Applications

Pedro Domingos PEDROD@CS.WASHINGTON.EDU
Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, U.S.A.

Abstract

This paper presents a unified bias-variance decomposition that is applicable to squared loss, zero-one loss, variable misclassification costs, and other loss functions. The unified decomposition sheds light on a number of significant issues: the relation between some of the previously-proposed decompositions for zero-one loss and the original one for squared loss, the relation between bias, variance and Schapire et al.'s (1997) notion of margin, and the nature of the trade-off between bias and variance in classification. While the bias-variance behavior of zero-one loss and variable misclassification costs is quite different from that of squared loss, this difference derives directly from the different definitions of loss. We have applied the proposed decomposition to decision tree learning, instance-based learning and boosting on a large suite of benchmark data sets, and made several significant observations.

1. Introduction

The bias-variance decomposition is a key tool for understanding machine-learning algorithms, and in recent years its use in empirical studies has grown rapidly. The notions of bias and variance help to explain how very simple learners can outperform more sophisticated ones, and how model ensembles can outperform single models. The bias-variance decomposition was originally derived for squared loss (see, for example, Geman et al. (1992)). More recently, several authors have proposed corresponding decompositions for zero-one loss. However, each of these decompositions has significant shortcomings. Kong and Dietterich's (1995) decomposition allows the variance to be negative, and ignores the noise component of misclassification error. Breiman's (1996b) decomposition is undefined for any given example (it is only defined for the instance space as a whole), and allows the variance to be zero or undefined even when the learner's predictions fluctuate in response to the training set. Tibshirani (1996) defines bias and variance, but decomposes loss into bias and the "aggregation effect," a quantity unrelated to his definition of variance. James and Hastie (1997) extend this approach by defining bias and variance but decomposing loss in terms of two quantities they call the "systematic effect" and "variance effect." Kohavi and Wolpert's (1996) decomposition allows the bias of the Bayes-optimal classifier to be nonzero. Friedman's (1997) decomposition relates zero-one loss to the squared-loss bias and variance of class probability estimates, leaving bias and variance for zero-one loss undefined. In each of these cases, the decomposition for zero-one loss is either not stated in terms of the zero-one bias and variance, or is developed independently from the original one for squared loss, without a clear relationship between them.

In this paper we propose a single definition of bias and variance, applicable to any loss function, and show that the resulting decomposition for zero-one loss does not suffer from any of the shortcomings of previous decompositions. Further, we show that notions like order-correctness (Breiman, 1996a) and margin (Schapire et al., 1997), previously proposed to explain why model ensembles reduce error, can be reduced to bias and variance as defined here. We also provide what to our knowledge is the first bias-variance decomposition for variable misclassification costs. Finally, we carry out a large-scale empirical study, measuring the bias and variance of several machine-learning algorithms in a variety of conditions, and extracting significant patterns.

2. A Unified Decomposition

Given a training set $\{(x_1, t_1), \ldots, (x_n, t_n)\}$, a learner produces a model f. Given a test example x, this model produces a prediction $y = f(x)$. (For the sake of simplicity, the fact that y is a function of x will remain implicit throughout this paper.) Let t be the true value of the predicted variable for the test example x. A *loss function* $L(t, y)$ measures the cost of predicting y when the true value is t. Commonly used loss functions are squared loss ($L(t, y) = (t-y)^2$), absolute loss ($L(t, y) = |t - y|$), and zero-one loss ($L(t, y) = 0$ if $y = t$, $L(t, y) = 1$ otherwise). The goal of learning can be stated as producing a model with the smallest possible loss; i.e., a model that minimizes the average $L(t, y)$ over all examples, with each example weighted by its probability. In general, t will be a nondeterministic function of x (i.e., if x is sampled repeatedly, different values of t will be seen). The *optimal prediction* y_* for an example x is the prediction that minimizes $E_t[L(t, y_*)]$, where the subscript t denotes that the expectation is taken with respect to all possible values of t, weighted by their probabilities given x. The optimal model is the model for which $f(x) = y_*$ for every x. In general, this model will have non-zero loss. In the case of zero-one loss, the optimal model is called the *Bayes classifier*, and its loss is called the *Bayes rate*.

Since the same learner will in general produce different models for different training sets, $L(t, y)$ will be a function of the training set. This dependency can be removed by averaging over training sets. In particular, since the training set size is an important parameter of a learning problem, we will often want to average over all training sets of a given size. Let D be a set of training sets. Then the quantity of interest is the expected loss $E_{D,t}[L(t, y)]$, where the expectation is taken with respect to t and the training sets in D (i.e., with respect to t and the predictions $y = f(x)$ produced for example x by applying the learner to each training set in D). Bias-variance decompositions decompose the expected loss into three terms: bias, variance and noise. A standard such decomposition exists for squared loss, and a number of different ones have been proposed for zero-one loss.

In order to define bias and variance for an arbitrary loss function we first need to define the notion of main prediction.

Definition 1 *The* main prediction *for a loss function L and set of training sets D is* $y_m^{L,D} = \text{argmin}_{y'} E_D[L(y, y')]$.

When there is no danger of ambiguity, we will represent $y_m^{L,D}$ simply as y_m. The expectation is taken with respect to the training sets in D, i.e., with respect to the predictions y produced by learning on the training sets in D. Let Y be the multiset of these predictions. (A specific prediction y will appear more than once in Y if it is produced by more than one training set.) In words, the main prediction is the value y' whose average loss relative to all the predictions in Y is minimum (i.e., it is the prediction that "differs least" from all the predictions in Y according to L). The main prediction under squared loss is the mean of the predictions; under absolute loss it is the median; and under zero-one loss it is the mode (i.e., the most frequent prediction). For example, if there are k training sets in D, we learn a classifier on each, $0.6k$ of these classifiers predict class 1, and $0.4k$ predict 0, then the main prediction under zero-one loss is class 1. The main prediction is not necessarily a member of Y; for example, if $Y = \{1, 1, 2, 2\}$ the main prediction under squared loss is 1.5.

We can now define bias and variance as follows.

Definition 2 *The* bias *of a learner on an example x is $B(x) = L(y_*, y_m)$.*

In words, the bias is the loss incurred by the main prediction relative to the optimal prediction.

Definition 3 *The* variance *of a learner on an example x is $V(x) = E_D[L(y_m, y)]$.*

In words, the variance is the average loss incurred by predictions relative to the main prediction. Bias and variance may be averaged over all examples, in which case we will refer to them as *average bias* $E_x[B(x)]$ and *average variance* $E_x[V(x)]$.

It is also convenient to define noise as follows.

Definition 4 *The* noise *of an example x is $N(x) = E_t[L(t, y_*)]$.*

In other words, noise is the unavoidable component of the loss, incurred independently of the learning algorithm.

Definitions 2 and 3 have the intuitive properties associated with bias and variance measures. y_m is a measure of the "central tendency" of a learner. (What "central" means depends on the loss function.) Thus $B(x)$ measures the systematic loss incurred by a learner, and $V(x)$ measures the loss incurred by its fluctuations around the central tendency in response to different training sets. If the loss function is nonnegative then bias and variance are also nonnegative. The bias is independent of the training set, and is zero for a learner that always makes the optimal prediction. The variance is independent of the true value of the predicted variable, and is zero for a learner that always makes the same prediction regardless of the training set. The only property that the definitions above require of the loss function is that its expected value be computable. However, it is not necessarily the case that the expected loss $E_{D,t}[L(t, y)]$ for a given loss function L can be decomposed into bias and variance as defined above. Our approach will be to propose a decomposition and then show that it applies to each of several different loss functions. Even when it does not apply, it may still be worthwhile to investigate how the expected loss can be expressed as a function of $B(x)$ and $V(x)$.

Consider an example x for which the true prediction is t, and a learner that predicts y given a training set in D. Then, for certain loss functions L, the following decomposition of $E_{D,t}[L(t,y)]$ holds:

$$\begin{aligned} & E_{D,t}[L(t,y)] \\ & \quad = c_1 E_t[L(t,y_*)] + L(y_*,y_m) + c_2 E_D[L(y_m,y)] \\ & \quad = c_1 N(x) + B(x) + c_2 V(x) \end{aligned} \quad (1)$$

c_1 and c_2 are multiplicative factors that will take on different values for different loss functions. It is easily seen that this decomposition reduces to the standard one for squared loss with $c_1 = c_2 = 1$, considering that for squared loss $y_* = E_t[t]$ and $y_m = E_D[y]$ (Geman et al., 1992):

$$\begin{aligned} E_{D,t}[(t-y)^2] &= E_t[(t-E_t[t])^2] + (E_t[t]-E_D[y])^2 \\ &\quad + E_D[(E_D[y]-y)^2] \end{aligned} \quad (2)$$

We now show that the same decomposition applies to a broad class of loss functions for two-class problems, including zero-one loss. (Below we extend this to multiclass problems for zero-one loss.) Let $P_D(y=y_*)$ be the probability over training sets in D that the learner predicts the optimal class for x.

Theorem 1 *In two-class problems, Equation 1 is valid for any real-valued loss function for which* $\forall_y\, L(y,y) = 0$ *and* $\forall_{y_1 \neq y_2}\, L(y_1,y_2) \neq 0$, *with* $c_1 = P_D(y=y_*) - \frac{L(y_*,y)}{L(y,y_*)} P_D(y \neq y_*)$ *and* $c_2 = 1$ *if* $y_m = y_*$, $c_2 = -\frac{L(y_*,y_m)}{L(y_m,y_*)}$ *otherwise.*

Proof. We begin by showing that

$$L(t,y) = L(y_*,y) + c_0 L(t,y_*) \quad (3)$$

with $c_0 = 1$ if $y = y_*$ and $c_0 = -\frac{L(y_*,y)}{L(y,y_*)}$ otherwise. If $y = y_*$ Equation 3 is trivially true with $c_0 = 1$. If $t = y_*$, $L(t,y) = L(y_*,y) - \frac{L(y_*,y)}{L(y,y_*)} L(t,y_*)$ is true because it reduces to $L(t,y) = L(t,y) - 0$. If $t = y$, $L(t,y) = L(y_*,y) - \frac{L(y_*,y)}{L(y,y_*)} L(t,y_*)$ is true because it reduces to $L(t,t) = L(y_*,y) - L(y_*,y)$, or $0 = 0$. But if $y \neq y_*$ and we have a two-class problem, either $t = y_*$ or $t = y$ must be true. Therefore if $y \neq y_*$ it is always true that $L(t,y) = L(y_*,y) - \frac{L(y_*,y)}{L(y,y_*)} L(t,y_*)$, completing the proof of Equation 3. We now show in a similar manner that

$$L(y_*,y) = L(y_*,y_m) + c_2 L(y_m,y) \quad (4)$$

with $c_2 = 1$ if $y_m = y_*$ and $c_2 = -\frac{L(y_*,y_m)}{L(y_m,y_*)}$ otherwise. If $y_m = y_*$ Equation 4 is trivially true with $c_2 = 1$. If $y = y_m$, $L(y_*,y) = L(y_*,y_m) - \frac{L(y_*,y_m)}{L(y_m,y_*)} L(y_m,y)$ is true because it reduces to $L(y_*,y_m) = L(y_*,y_m) - 0$. If $y = y_*$, $L(y_*,y) = L(y_*,y_m) - \frac{L(y_*,y_m)}{L(y_m,y_*)} L(y_m,y)$ is true because it reduces to $L(y_*,y_*) = L(y_*,y_m) - L(y_*,y_m)$, or $0 = 0$. But if $y_m \neq y_*$ and we have a two-class problem, either $y = y_m$ or $y = y_*$ must be true. Therefore if $y_m \neq y_*$ it is always true that $L(y_*,y) = L(y_*,y_m) - \frac{L(y_*,y_m)}{L(y_m,y_*)} L(y_m,y)$, completing the proof of Equation 4. Using Equation 3, and considering that $L(y_*,y)$ and c_0 do not depend on t and $L(t,y_*)$ does not depend on D,

$$\begin{aligned} E_{D,t}[L(t,y)] &= E_D[E_t[L(t,y)]] \\ &= E_D[L(y_*,y) + c_0 E_t[L(t,y_*)]] \\ &= E_D[L(y_*,y)] + E_D[c_0] E_t[L(t,y_*)] \end{aligned} \quad (5)$$

Substituting Equation 4 and considering that $E_D[c_0] = P_D(y=y_*) - \frac{L(y_*,y)}{L(y,y_*)} P_D(y \neq y_*) = c_1$ results in Equation 1. □

In particular, if the loss function is symmetric (i.e., $\forall_{y_1,y_2} L(y_1,y_2) = L(y_2,y_1)$), c_1 and c_2 reduce to $c_1 = 2P_D(y=y_*) - 1$ and $c_2 = 1$ if $y_m = y_*$ (i.e., if $B(x) = 0$), $c_2 = -1$ otherwise (i.e., if $B(x) = 1$). Specifically, this applies to zero-one loss, yielding a decomposition similar to that of Kong and Dietterich (1995). The main differences are that Kong and Dietterich ignored the noise component $N(x)$ and defined variance simply as the difference between loss and bias, apparently unaware that the absolute value of that difference is the average loss incurred relative to the most frequent prediction. A side-effect of this is that Kong and Dietterich incorporate c_2 into their definition of variance, which can therefore be negative. Kohavi and Wolpert (1996) and others have criticized this fact, since variance for squared loss must be positive. However, our decomposition shows that the subtractive effect of variance follows from a self-consistent definition of bias and variance for zero-one and squared loss, even if the variance itself remains positive. The fact that variance is additive in unbiased examples but subtractive in biased ones has significant consequences. If a learner is biased on an example, increasing variance decreases loss. This behavior is markedly different from that of squared loss, but is obtained with the same definitions of bias and variance, purely as a result of the different properties of zero-one loss. It helps explain how highly unstable learners like decision-tree and rule induction algorithms can produce excellent results in practice, even given very limited quantities of data. In effect, when zero-one loss is the evaluation criterion, there is a much higher tolerance for variance than if the bias-variance decomposition was purely additive, because the increase in average loss caused by variance on unbiased examples is partly offset (or more than offset) by its decrease on biased ones. The average loss over all examples is the sum of noise, the average bias and what might be termed the *net variance*, $E_x[c_2V(x)]$:

$$E_{D,t,x}[L(t,y)] = E_x[c_1N(x)] + E_x[B(x)] + E_x[c_2V(x)] \quad (6)$$

by averaging Equation 1 over all test examples x, with c_2 positive for unbiased examples and negative for biased ones.

The c_1 factor (see Theorem 1) also points to a key difference between zero-one and squared loss. In squared loss, increasing noise always increases error. In zero-one loss, for training sets and test examples where $y \neq y_*$, increasing noise decreases error, and a high noise level can therefore in principle be beneficial to performance.

The general case of Theorem 1 is also important. In many practical applications of machine learning, loss is highly asymmetric; for example, classifying a cancerous patient as healthy is likely to be more costly than the reverse. In these cases, Theorem 1 essentially shows that the loss-reducing effect of variance on biased examples will be greater or smaller depending on how asymmetric the costs are, and on which direction they are greater in.

Equation 1 does not apply if $L(y,y) \neq 0$; in this case the decomposition contains an additional term corresponding to the cost of the correct predictions. Whether it applies in the general multiclass case is an open problem. However, it applies to the general multiclass problem for zero-one loss, as described in the following theorem.

Theorem 2 *Equation 1 is valid for zero-one loss in multiclass problems, with* $c_1 = P_D(y = y_*) - P_D(y \neq y_*)\, P_t(y = t \mid y_* \neq t)$ *and* $c_2 = 1$ *if* $y_m = y_*$, $c_2 = -P_D(y = y_* \mid y \neq y_m)$ *otherwise.*

We omit the proof in the interests of space; see Domingos (2000). Theorem 2 means that in multiclass problems not all variance on biased examples contributes to reducing loss; of all training sets for which $y \neq y_m$, only some have $y = y_*$, and it is in these that loss is reduced.

3. Properties of the Unified Decomposition

One of the main concepts Breiman (1996a) used to explain why the bagging ensemble method reduces zero-one loss was that of an *order-correct* learner. A learner is order-correct on an example x iff $\forall_{y \neq y_*}\ P_D(y) < P_D(y_*)$. Breiman showed that bagging transforms an order-correct learner into a nearly optimal one. Order-correctness and bias are closely related: a learner is order-correct on an example x iff $B(x) = 0$ under zero-one loss. (The proof of this is immediate from the definitions, considering that y_m for zero-one loss is the most frequent prediction.)

Schapire et al. (1997) proposed an explanation for why the boosting ensemble method works in terms of the notion of *margin*. For algorithms like bagging and boosting, which generate multiple hypotheses by applying the same learner to multiple training sets, their definition of margin can be stated as follows.

Definition 5 *(Schapire et al., 1997) In two-class problems, the* margin *of a learner on an example* x *is* $M(x) = P_D(y = t) - P_D(y \neq t)$.

A positive margin indicates a correct classification by the ensemble, and a negative one an error. Intuitively, a large margin corresponds to a high confidence in the prediction. D here is the set of training sets to which the learner is applied. For example, if 100 rounds of boosting are carried out, $|D| = 100$. Further, for algorithms like boosting where the different training sets (and corresponding predictions) have different weights that sum to 1, $P_D(.)$ is computed according to these weights. Definitions 1–4 apply unchanged in this situation. In effect, we have generalized the notions of bias and variance to apply to any training set selection scheme, not simply the traditional one of "all possible training sets of a given size, with equal weights."

Schapire et al. (1997) showed that it is possible to bound an ensemble's generalization error (i.e., its zero-one loss on test examples) in terms of the distribution of margins on training examples and the VC dimension of the base learner. In particular, the smaller the probability of a low margin, the lower the bound on generalization error. The following theorem shows that the margin is closely related to bias and variance as defined above.

Theorem 3 *The margin of a learner on an example* x *can be expressed in terms of its zero-one bias and variance as* $M(x) = \pm[2B(x) - 1][2V(x) - 1]$, *with positive sign if* $y_* = t$ *and negative sign otherwise.*

Proof. When $y_* = t$, $M(x) = P_D(y = y_*) - P_D(y \neq y_*) = 2P_D(y = y_*) - 1$. If $B(x) = 0$, $y_m = y_*$ and $M(x) = 2P_D(y = y_m) - 1 = 2[1 - V(x)] - 1 = -[2V(x) - 1]$. If $B(x) = 1$ then $M(x) = 2V(x) - 1$. Therefore $M(x) = [2B(x)-1][2V(x)-1]$. The demonstration for $y_* \neq t$ is similar, with $M(x) = P_D(y \neq y_*) - P_D(y = y_*)$. □

Conversely, it is possible to express the bias and variance in terms of the margin: $B(x) = \frac{1}{2}[1 \pm \text{sign}(M(x))]$, $V(x) = \frac{1}{2}[1 \pm |M(x)|]$, with positive sign if $y_* \neq t$ and negative sign otherwise. The relationship between margins and bias/variance expressed in Theorem 3 implies that Schapire et al.'s theorems can be stated in terms of the bias and variance on training examples. Bias-variance decompositions relate a learner's loss on an example to its bias and variance on that example. However, to our knowledge this is the first time that *generalization* error is related to bias and variance on *training* examples.

Theorem 3 also sheds light on the polemic between Breiman (1996b, 1997) and Schapire et al. (1997) on how the success of ensemble methods like bagging and boosting is best explained. Breiman has argued for a bias-variance explanation, while Schapire et al. have argued for a margin-based explanation. Theorem 3 shows that these are two faces of the same coin, and helps to explain why the bias-variance explanation sometimes seems to fail when applied to boosting. Maximizing margins is a combination of reducing the number of biased examples, decreasing variance on unbiased examples, and increasing it on biased ones (for examples where $y_* = t$; the reverse, otherwise). Without differentiating between these effects it is hard to understand how boosting affects bias and variance.

4. Experiments

We applied the bias-variance decomposition of zero-one loss proposed here in a series of experiments with classification algorithms. To our knowledge this is the most extensive such study to date, in terms of the number of data sets and number of algorithms/parameter settings studied. This section summarizes the results. We used the following 30 data sets from the UCI repository (Blake & Merz, 2000): annealing, audiology, breast cancer (Ljubljana), chess (king-rook vs. king-pawn), credit (Australian), diabetes, echocardiogram, glass, heart disease (Cleveland), hepatitis, horse colic, hypothyroid, iris, labor, LED, lenses, liver disorders, lung cancer, lymphography, mushroom, post-operative, primary tumor, promoters, solar flare, sonar, soybean (small), splice junctions, voting records, wine, and zoology.

As the noise level $N(x)$ is very difficult to estimate, we followed previous authors (e.g., Kohavi & Wolpert (1996)) in assuming $N(x) = 0$. This is not too detrimental to the significance of the results because we are mainly interested in the variation of bias and variance with several factors, not their absolute values. We estimated bias, variance and zero-one loss by the following method. We randomly divided each dataset into training data (two thirds of the examples) and test data (one third). For each dataset, we generated 100 different training sets by the *bootstrap* method (Efron & Tibshirani, 1993): if the training data consists of n examples, we create a *bootstrap replicate* of it by taking n samples *with replacement* from it, with each example having a probability of $1/n$ of being selected at each turn. As a result, some of the examples will appear more than once in the training set, and some not at all. The 100 training sets thus obtained were taken as a sample of the set D, with D being the set of all training sets of size n. A model was then learned on each training set. We used the predictions made by these models on the test examples to estimate average zero-one loss, average bias and net variance, as defined in Section 2. We also measured the total contribution to average variance from unbiased examples $V_u = \frac{1}{n}[\sum_{i=1}^{n}(1-B(x_i))V(x_i)]$ and the contribution from biased examples $V_b = \frac{1}{n}[\sum_{i=1}^{n} cB(x_i)V(x_i)]$, where x_i is a test example, n is the number of test examples, $c = 1$ for two-class problems, and $c = P_D(y = y_*|y \neq y_m)$ for multiclass problems (see Theorem 2), estimated from the test set. The net variance is the difference of the two: $V = V_u - V_b$.

We carried out experiments with decision-tree induction, boosting, and k-nearest neighbor; their results are reported in turn. Space limitations preclude presentation of the complete results; see Domingos (2000). Here we summarize the main observations, and present representative examples.

4.1 Decision-Tree Induction

We used the C4.5 decision tree learner, release 8 (Quinlan, 1993). We measured zero-one loss, bias and variance while varying C4.5's pruning parameter (the confidence level CF) from 0% (maximum pruning) to 100% (minimum) in 5% steps. The default setting is 25%. Surprisingly, we found that in most data sets CF has only a minor effect on bias and variance (and therefore loss). Only at the CF=0% extreme, where the tree is pruned all the way to the root, is there a major impact, with very high bias and loss; but this disappears by CF=5%. These results suggest there may be room for improvement in C4.5's pruning method (cf. Oates & Jensen (1997)).

In order to obtain a clearer picture of the bias-variance trade-off in decision tree induction, we replaced C4.5's native pruning scheme with a limit on the number of levels allowed in the tree. (When a maximum level of m is set, every path in the tree of length greater than m is pruned back to a length of m.) The dominant effect observed is the rapid decrease of bias in the first few levels, after which it typically stabilizes. In 9 of the 25 data sets where this occurs, bias in fact increases after this point (slightly in 6, markedly in 3). In 5 data sets bias increases with the number of levels overall; in 3 of these (echocardiogram, post-operative and sonar) it increases markedly. Variance increases with the number of levels in 26 data sets; in 17 of these the increase is generally even, and much slower than the initial decrease in bias. Less-regular patterns occur in the remaining 9 data sets. V_u and V_b tend to be initially similar, but V_b increases more slowly than V_u, or decreases. At any given level, V_b typically offsets a large fraction of V_u, making variance a smaller contributor to loss than would be the case if its effect was always positive. This leads to the hypothesis that higher-variance algorithms (or settings) may be better suited to classification (zero-one loss) than regression (squared loss). Perhaps not coincidentally, research in classification has tended to explore higher-variance algorithms than research in regression.

Representative examples of the patterns observed are shown in Figure 1, where the highest level shown is the highest produced by C4.5 when it runs without any limits. Overall, the expected pattern of a trade-off in bias and variance leading to a minimum of loss at an intermediate level was observed in only 10 data sets; in 6 a decision stump was best, and in 14 an unlimited number of levels was best.

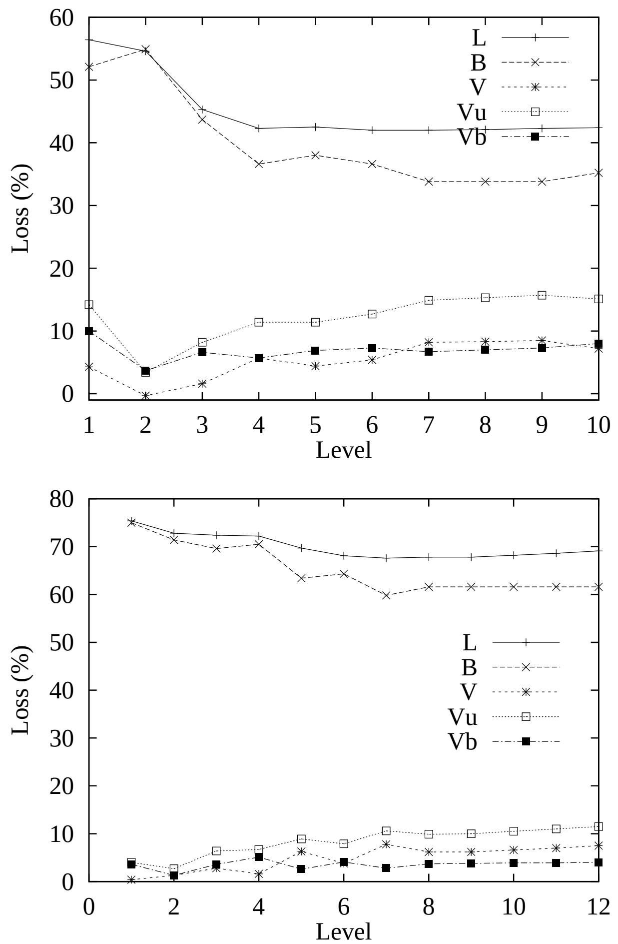

Figure 1. Effect of varying the number of levels in C4.5 trees: glass (top) and primary tumor (bottom).

4.2 Boosting

We also experimented with applying AdaBoost (Freund & Schapire, 1996) to C4.5. We allowed a maximum of 100 rounds of boosting. (In most data sets, loss and its components stabilized by the 20th round, and only this part is graphed.) Boosting decreases loss in 21 data sets and increases it in 3; it has no effect in the remainder. It decreases bias in 14 data sets and increases it in 7, while decreasing net variance in 18 data sets and increasing it in 7. The bulk of bias reduction typically occurs in the first few rounds. Variance reduction tends to be more gradual. On average (over all data sets) variance reduction is a much larger contributor to loss reduction than bias reduction (2.5% vs. 0.6%). Over all data sets, the variance reduction is significant at the 5% level according to sign and Wilcoxon tests, but bias reduction is not. Thus variance reduction is clearly the dominant effect when boosting is applied to C4.5; this is consistent with the notion that C4.5 is a "strong" learner. Boosting tends to reduce both V_u and V_b, but it reduces V_u much more strongly than V_b (3.0% vs. 0.5%). The ideal behavior would be to reduce V_u and increase V_b; it may be possible to design a variant of boosting that achieves this, and as a result further reduces loss. Examples of the boosting behaviors observed are shown in Figure 2.

4.3 K-Nearest Neighbor

We studied the bias and variance of the k-nearest neighbor algorithm (Cover & Hart, 1967) as a function of k, the number of neighbors used to predict a test example's class. We used Euclidean distance for numeric attributes and overlap for symbolic ones. k was varied from 1 to 21 in increments of 2; typically only small values of k are used, but this extended range allows a clearer observation of its effect. The pattern

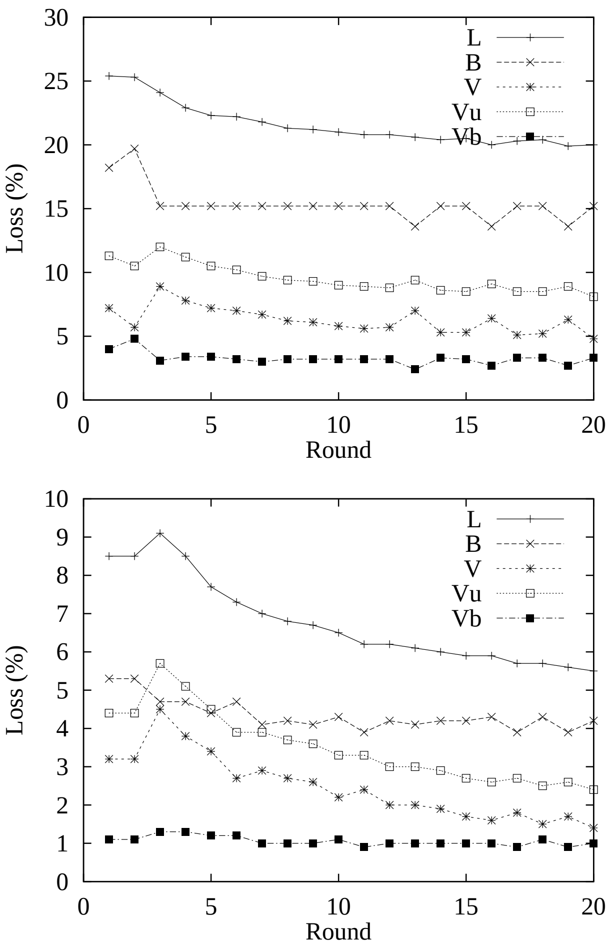

Figure 2. Effect of boosting on C4.5: audiology (top) and splice junctions (bottom).

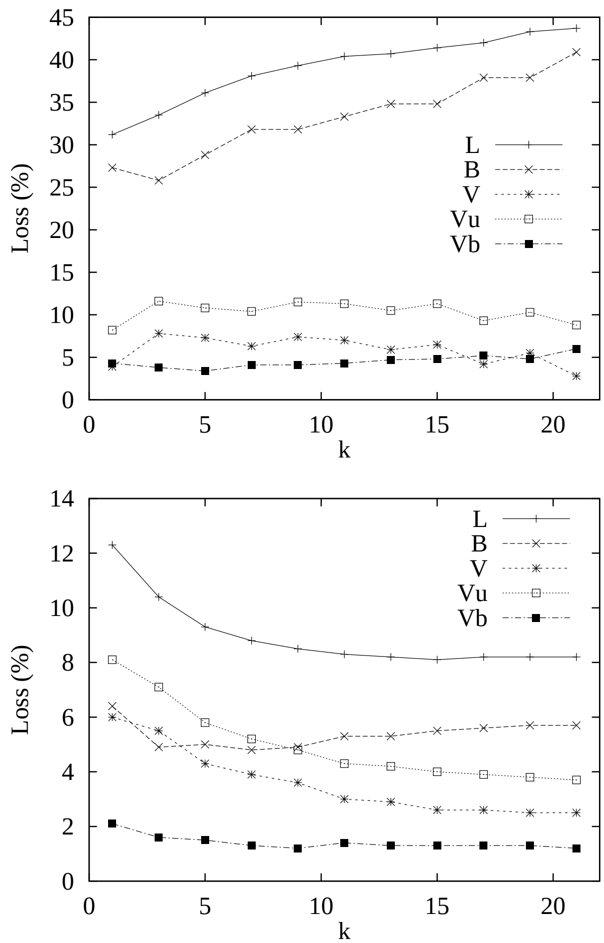

Figure 3. Effect of varying k in k-nearest neighbor: audiology (top) and chess (bottom).

of an increase in bias and a decrease in variance with k producing a minimal loss at an intermediate value of k is seldom observed; more often one of the two effects dominates throughout. In several cases bias and variance vary in the same direction with k. In 13 data sets, the lowest loss is obtained with $k = 1$, and in 11 with the maximum k. On average (over all data sets) bias increases markedly with k (by 4.9% from $k = 1$ to k= 21), but variance decreases only slightly (0.8%), resulting in much increased loss. This contradicts Friedman's (1997) hypothesis (based on approximate analysis and artificial data) that very large values of k should be beneficial. This may be attributable to the fact that, as k increases, what Friedman calls the "boundary bias" changes from negative to positive for a majority of the examples, wiping out the benefits of low variance. Interestingly, increasing k in k-NN has the "ideal" effect of reducing V_u (by 0.5% on average) while increasing V_b (0.3%). Figure 3 shows examples of the different types of behavior observed.

5. Future Work

The main limitation of the definitions of bias and variance proposed here is that many loss functions cannot be decomposed according to them and Equation 1 (e.g., absolute loss). Since it is unlikely that meaningful definitions exist for which a simple decomposition is always possible, a central direction for future work is determining general properties of loss functions that are necessary and/or sufficient for Equation 1 to apply. Even when it does not, it may be possible to usefully relate loss to bias and variance as defined here. (For example, in Domingos (2000) we show that, as long as the loss function is a metric, it can be bounded from above and below by linear functions of the bias, variance and noise.)

Another major direction for future work is applying the decomposition to a wider variety of learners, in order to gain insight about their behavior, both with respect to variations within a method and with respect to comparisons between methods. We would also like to study experimentally the effect of different domain characteristics (e.g., sparseness of the data) on the bias and variance of different learning algorithms. The resulting improved understanding should allow us to design learners that are more easily adapted to a wide range of domains.

6. Conclusion

This paper proposed unified definitions of bias and variance, applicable to any loss function. The resulting decomposition specializes to the conventional one for squared loss, avoids the difficulties of previous ones for zero-one loss, and is also applicable to variable misclassification costs. While the decomposition is not always purely additive, we believe that more insight is gained from this approach—formulating consistent definitions and investigating what follows from them—than from crafting definitions case-by-case to make the decomposition purely additive. For example, uncovering the different role of variance on biased and unbiased examples in zero-one loss leads to an improved understanding of classification algorithms, and of how they differ from regression ones. This was illustrated in an extensive empirical study of bias and variance in decision tree induction, boosting, and k-nearest neighbor.

The bias-variance decomposition proposed in this paper is available in C code at http://www.cs.washington.edu/homes/pedrod/bvd.c.

Acknowledgments

This research was partly supported by a PRAXIS XXI grant. The author is grateful to all those who provided the data sets used in the experiments.

References

Blake, C., & Merz, C. J. (2000). *UCI repository of machine learning databases*. Department of Information and Computer Science, University of California at Irvine, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (1996b). *Bias, variance and arcing classifiers* (Technical Report 460). Statistics Department, University of California at Berkeley, Berkeley, CA.

Breiman, L. (1997). *Arcing the edge* (Technical Report 486). Statistics Department, University of California at Berkeley, Berkeley, CA.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Domingos, P. (2000). *A unified bias-variance decomposition* (Technical Report). Department of Computer Science and Engineering, University of Washington, Seattle, WA.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 148–156). Bari, Italy: Morgan Kaufmann.

Friedman, J. H. (1997). On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*, 55–77.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

James, G., & Hastie, T. (1997). *Generalizations of the bias/variance decomposition for prediction error* (Technical Report). Department of Statistics, Stanford University, Stanford, CA.

Kohavi, R., & Wolpert, D. H. (1996). Bias plus variance decomposition for zero-one loss functions. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 275–283). Bari, Italy: Morgan Kaufmann.

Kong, E. B., & Dietterich, T. G. (1995). Error-correcting output coding corrects bias and variance. *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 313–321). Tahoe City, CA: Morgan Kaufmann.

Oates, T., & Jensen, D. (1997). The effects of training set size on decision tree complexity. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 254–262). Madison, WI: Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 322–330). Nashville, TN: Morgan Kaufmann.

Tibshirani, R. (1996). *Bias, variance and prediction error for classification rules* (Technical Report). Department of Statistics, University of Toronto, Toronto, Canada.