Unpublished Draft

Estimating bias and variance from data

Geoffrey I. Webb
Paul Conilione
School of Computer Science and Software Engineering
Monash University
Vic. 3800, Australia

Abstract. The bias-variance decomposition of error provides useful insights into the error performance of a classifier as it is applied to different types of learning task. Most notably, it has been used to explain the extraordinary effectiveness of ensemble learning techniques. It is important that the research community have effective tools for assessing such explanations. To this end, techniques have been developed for estimating bias and variance from data. The most widely deployed of these uses repeated sub-sampling with a holdout set. We argue, with empirical support, that this approach has serious limitations. First, it provides very little flexibility in the types of distributions of training sets that may be studied. It requires that the training sets be relatively small and that the degree of variation between training sets be very circumscribed. Second, the approach leads to bias and variance estimates that have high statistical variance and hence low reliability. We develop an alternative method that is based on cross-validation. We show that this method allows far greater flexibility in the types of distribution that are examined and that the estimates derived are much more stable. Finally, we show that changing the distributions of training sets from which bias and variance estimates are drawn can alter substantially the bias and variance estimates that are derived.

Keywords: bias-variance decomposition of error, estimating classification performance

1. Introduction

The bias plus variance decomposition of error has proved a useful tool for analyzing supervised learning algorithms. While initially developed in the context of numeric regression (specifically, of squared error loss, Geman, Bienenstock, & Doursat, 1992), a number of variants have been developed for classification learning (zero-one loss) (Breiman, 1996b; Kohavi & Wolpert, 1996; Kong & Dietterich, 1995; Friedman, 1997; Domingos, 2000; Webb, 2000; James, 2003). This analysis decomposes error into three terms, derived with reference to the performance of a learner when trained with different training sets drawn from some reference distribution of training sets:

Squared bias: a measure of the error of the central tendency of the learner.

Variance: a measure of the degree to which the learner's predictions differ as it is applied to learn models from different training sets.

Intrinsic noise: a measure of the degree to which the target quantity is inherently unpredictable. This measure equals the expected cost of the Bayes optimal classifier.

This analysis has been used widely to gain insight into the relative performance of alternative algorithms (for example, John, 1995; Breiman, 1996a, 1998; Kohavi, Becker, & Sommerfield, 1997; Kohavi & John, 1997; Bauer & Kohavi, 1999; Zheng, Webb, & Ting, 1999; Webb, 1999, 2000; Gama & Brazdil, 2000; Valentini & Dietterich, 2003; Yang & Webb, 2003).

The machine learning community has a long-running concern for empirical evaluation. It is regarded as important to support theoretical analysis by analysis of experimental outcomes. If this concern is to be extended to theoretical analysis based on the bias-variance decomposition of error, we require reliable methods for evaluating and comparing algorithms' bias-variance performance.

The most widely employed approach to estimating bias and variance from data is the holdout approach of Kohavi and Wolpert (1996). Note that we are interested in their procedure for estimating bias and variance as distinct from their definitions of bias and variance, also provided in the same paper and which we also adopt in the current paper. Their procedure splits the training data into two subsets, a training pool and a holdout test set. The training sets are then formed from random samples drawn from the training pool. We argue that this approach is fundamentally flawed, resulting in undesirably small training sets, providing little if any control over the degree of variation in composition of training sets, and resulting in tremendous instability in the estimates that it derives. All of these problems are serious. The use of small training sets means that bias-variance studies are conducted on fundamentally different distributions of training sets to those to which the learners are to be applied in practice. The absence of control over the degree of variation in the composition of the training sets means that it is not possible to study the bias-variance characteristics of algorithms as they are applied to distributions with alternative levels of variation. The instability of the estimates means that individual estimates are inherently inaccurate.

An alternative approach based on multiple cross-validation trials has been presented by Webb (2000). We herein extend this approach to provide greater control over the size of the training sets and the degree of variation in the composition of training sets. We argue that this new approach overcomes all the identified deficiencies of Kohavi

and Wolpert's procedure, resulting in stable estimates, and allowing precise control over training set sizes and the degree of variation in the composition of training sets.

Using this new bias-variance analysis technique, we derive new insights into the relationship between different types of training set distribution and the bias-variance profiles that are derived therefrom.

While a single definition of bias and variance has been adopted in the context of regression, there has been considerable debate about how this definition might best be extended to classification (Breiman, 1996b; Kohavi & Wolpert, 1996; Kong & Dietterich, 1995; Friedman, 1997; Domingos, 2000; Webb, 2000; James, 2003). Rather than entering into this debate, we here use Kohavi and Wolpert's (1996) definition on the grounds that we believe it is the most widely employed in practice. However, the techniques that we develop are equally applicable to any of the above bias-variance definitions.

2. Evaluating bias and variance

We wish to analyze the classification performance of a classification learning system, $\mathcal{L}$, that can be regarded as a function

$$\mathcal{L}(T) \rightarrow (C(X) \rightarrow Y)$$

from training sets T to models, where each model $C(X) \rightarrow Y$ is a function from objects to classes. T is a multiset of n class–description pairs. Each pair (y, x) associates class $y \in Y$ with description $x \in X$. $\mathcal{L}(T) \rightarrow (C(X) \rightarrow Y)$ is a function from training sets to classifiers, which are in turn functions from descriptions to classes. Bias-variance analyses must be performed with respect to a joint distribution Y, X from which the test set is drawn together with a distribution of training sets $\mathcal{T}$ from which the training sets are drawn. Note that while $\mathcal{T}$ contains (y, x) pairs, these need not be drawn from the same joint distribution Y, X as the test set.

Typically, we have a single sample of data $\mathcal{D}$ from which we wish to derive bias-variance estimates. $\mathcal{D}$ is presumably a sample from some distribution Θ. However, if we generate a distribution of training sets $\mathcal{T}$ by sampling from $\mathcal{D}$, $\mathcal{T}$ will be a different distribution to Θ, unless Θ contains only subsets of $\mathcal{D}$ and the sampling methodology used to draw samples from $\mathcal{D}$ replicates Θ. Further, irrespective of the distribution from which $\mathcal{D}$ is drawn, we may wish to manipulate the distribution of training sets in order to produce quite different distributions for experimental purposes. In consequence, we should not assume that $\mathcal{T}$ replicates Θ.

Clearly it is possible to have many different types of training set distribution $\mathcal{T}$. It is reasonable to expect the bias-variance characteristics of an algorithm to differ depending upon the properties of the distribution. However, this issue has received little attention in the literature. The only property that is usually considered is the size of the training sets in the distribution. For reasons that are not clear to us, it is commonly assumed that there will be a single training set size for any one distribution. That is, it is assumed that a single distribution will not contain training sets of different sizes.

However, the size of the training set is only one of the properties of a distribution that we might expect to affect bias and variance. For example, the relative class frequencies might be expected to affect bias and variance as the more the data is dominated by a single class the less might be the variation in the predictions of a typical classifier. A further property that we expect to be particularly significant for bias-variance estimation is the *inter-training-set variability* (δ). We define δ as the average pairwise proportion of objects in any two training sets of a distribution that are different. This metric provides a measure of the variation between training sets in the distribution. The minimum possible inter-training-set variability of $\delta = 0.0$ indicates that all training sets are identical. A deterministic learning algorithm will have variance of 0.0 for such a training set distribution as it will always learn the same classifier. The maximum possible inter-training-set variability, $\delta = 1.0$ indicates that no two training sets in the distribution share any objects in common. We might expect such a distribution to result in high variance for most learning algorithms.

2.1. Kohavi and Wolpert's definitions of bias and variance

Kohavi and Wolpert (1996) define the bias and variance decomposition of error as follows.

$$bias_x^2 = \frac{1}{2} \sum_{y \in Y} [P_{Y,X}(Y{=}y|X{=}x) - P_{\mathcal{T}}(\mathcal{L}(\mathcal{T})(x){=}y)]^2 \tag{1}$$

$$variance_x = \frac{1}{2} \left(1 - \sum_{y \in Y} P_{\mathcal{T}}(\mathcal{L}(\mathcal{T})(x){=}y)^2 \right) \tag{2}$$

$$\sigma_x = \frac{1}{2} \left(1 - \sum_{y \in Y} P_{Y,X}(Y{=}y|X{=}x)^2 \right) \tag{3}$$

The third term, σ_x, represents the irreducible error.

When estimating these terms from data Kohavi and Wolpert (1996) recommend the use of a correction to unbias the estimates. Thus, the estimate of $bias^2$ is

$$\widehat{bias^2_x} = \frac{1}{2}\sum_{y \in Y}[\hat{P}_{Y,X}(Y{=}y|X{=}x) - \hat{P}_{\mathcal{T}}(\mathcal{L}(\mathcal{T})(x){=}y)]^2$$

$$-\hat{P}_{\mathcal{T}}(\mathcal{L}(\mathcal{T})(x){=}y)\left(1 - \hat{P}_{\mathcal{T}}(\mathcal{L}(\mathcal{T})(x){=}y)\right)/(N-1) \qquad (4)$$

where $\hat{P}(\cdot)$ is the estimate of $P(\cdot)$ derived from the observed frequency of the argument over repeated sample training sets.

Further, as it is infeasible to estimate σ from sample data, this term is aggregated into the bias term by assuming that $P_{\mathcal{T}}(\mathcal{L}(\mathcal{T})(x){=}y)$ is always either 0.0 or 1.0. Hence,

$$\widehat{variance}_x = \hat{P}_{Y,X}(\mathcal{L}(\mathcal{T})(X) \neq Y|X{=}x) - \widehat{bias^2_x}. \qquad (5)$$

That is, the estimate of variance equals the estimate of error minus the estimate of bias.

We draw attention to the manner in which bias and variance are presented as functions with a single parameter, x. From careful analysis, however it can be seen that there are two further parameters, $\mathcal{L}$ and $\mathcal{T}$. That is, the terms should more correctly be written as $bias^2_{x,\mathcal{L},\mathcal{T}}$ and $variance_{x,\mathcal{L},\mathcal{T}}$. In most cases there is little harm in dropping $\mathcal{L}$ as it will usually be very clear from the context. One of our contentions, however, is that the failure to recognize $\mathcal{T}$ as an important parameter has led to a serious failure to understand the significance of the distribution in determining bias-variance results.

Note that bias and variance are defined here with respect to a single test object. In practice we evaluate these terms over all of a set of test objects and present the mean value of each term.

Note also that Kohavi and Wolpert all their bias term $bias^2$, following the convention set in the context of numeric regression (Geman et al., 1992). In this paper we use *bias* to refer to the bias term in a classification context.

2.2. Kohavi and Wolpert's holdout procedure

Kohavi and Wolpert (1996) present the following holdout procedure for estimating the bias and variance of a learner $\mathcal{L}$ from a dataset $\mathcal{D}$.

1. Randomly divide $\mathcal{D}$ into two parts, D, a pool of objects from which training sets are drawn, and E, the test set against which the performance of each model is evaluated.

2. N training sets $d_1 \ldots d_N$ are generated by uniform random sampling without replacement from D. To obtain training sets of size m, the size of D is set to $2m$.

3. Apply $\mathcal{L}$ to each d_i, estimating the required values from the performance of $\mathcal{L}(d_i)$ on E.

This procedure has a number of limitations. For each run of the learner, the available data $\mathcal{D}$ is fragmented into three subgroups, d_i, $D - d_i$, and E. There are good reasons to want each of these to be as large as possible. First, if we are seeking to evaluate the bias and variance characteristics of an algorithm with respect to a certain type of learning scenario, we should examine its behavior when it learns from the quantities of data available in that scenario. This may mean learning in the ideal case from all of $\mathcal{D}$. To learn from less is to study the algorithm's behavior in the context of smaller training sets. On the other hand, however, we also want $D - d_i$ to be large so that there will be variety between the various samples d_i. We additionally want E to be large so that we obtain a stable estimate of the performance of each $\mathcal{L}(d_i)$. If either the pool from which training sets are drawn D or the set of test objects E is small then we can expect the error and hence the bias and variance of the learner to alter substantially depending upon the exact set of objects that happens to be selected for inclusion in either of D or E. This is clearly undesirable as it mitigates against obtaining reliable and consistent results.

Unfortunately, these three demands are contradictory. The sets d_i, $D - d_i$, and E are disjoint and hence the size of any one can only be increased at the expense of at least one of the others.

A further problem is that it allows very little control over the type of distribution from which the training sets are drawn. In Kohavi and Wolpert's technique, the training set of size m is always drawn randomly without replacement from a pool containing $2m$ objects. This means that the inter-training-set variability δ for the distribution that is created will always be $1/2$. If we are to understand bias and variance in general we should want to study alternative rates of variation between training sets. For example, we should want to be able to answer the question *how do bias and variance alter as δ alters?*

2.3. Valentini and Dietterich's out-of-bag technique

Valentini and Dietterich (2003) present an alternative procedure for estimating bias and variance from data. They perform repeated trials in each of which a bootstrapped sample $\mathcal{B}$ is drawn from $\mathcal{D}$. A bootstrapped sample is a sample of the same size as $\mathcal{D}$ that is drawn at

random from $\mathcal{D}$ with replacement. That is, a single object in $\mathcal{D}$ may be represented multiple times in $\mathcal{B}$. On average, 36.8% of objects in $\mathcal{D}$ will not appear in $\mathcal{B}$. This set of objects, $\mathcal{D} - B$ is classified by $\mathcal{L}(B)$, and the resulting classifications recorded. After all of the trials have been performed, bias and variance are then estimated from the record of all classifications of each object.

This approach has a number of advantages over Kohavi and Wolpert's (1996) procedure. First, the training set sizes are larger, providing more representative training set sizes. Second, so long as sufficient trials are performed, all objects in $\mathcal{D}$ will participate in the estimates in both training and test set roles. This can be expected to result in much more stable estimates.

However, this procedure is also subject to the problem that it does not allow much control over the form of distributions that may be studied. Repeated bootstrap samples result in a very particular form of distribution of training sets, a distribution in which each training set contains many objects that are duplicated, some of them many times. While it may be interesting to study the bias and variance characteristics of such distributions, there does not appear to be any strong theoretical reason to wish to restrict bias and variance analyses to such distributions.

A further limitation of this procedure is that because the selection of the test set is random, many trials will need to be performed in order to obtain a guarantee that most objects will be classified sufficient times to obtain a reasonable estimate of the learner's bias and variance when applied to them. The number of times each object is classified will vary substantially. If sufficient trials are performed to ensure sufficient classifications for most objects then many objects will be classified many more times than necessary. For example, with 50 trials, almost 1% of objects will be classified 10 or fewer times while almost 27% of objects will be classified more than 20 times. This suggests that the approach involves some wasted computation in the form of needless classification of some objects.

2.4. Webb's cross-validation procedure

Dietterich (1998) has argued that repeated cross-validation trials provide a superior means of comparing classification error to any of four alternative techniques including random selection of a test set. While we are concerned here with the question of how best to estimate bias and variance from data, rather than the subsequent question of how to compare the bias and variance performance of multiple algorithms, we hypothesize that the characteristics that make cross-validation a

better basis for comparing error also make it a superior process for estimating bias and variance. In particular, we hypothesis that a cross-validation based technique for estimating bias and variance, such as that proposed by Webb (2000), will provide more stable estimates than will a sampling-based technique such as that proposed by Kohavi and Wolpert (1996).

Webb's (2000) procedure repeats k-fold cross validation l times. This ensures that each element x of the dataset $\mathcal{D}$ is classified l times. The $bias_x$ and $variance_x$ can be estimated from the resulting set of classifications. The bias and variance with respect to the distribution from which $\mathcal{D}$ is drawn can be estimated from the average of each term over all $x \in \mathcal{D}$.

This procedure has a number of advantages over Kohavi and Wolpert's. First, like Valentini and Dietterich's (2003) bootstrap procedure, all data are used as both training and test data. This can be expected to lead to far greater stability in the estimates of bias and variance that are derived, as selection of different training and test sets can be expected to substantially alter the estimates that are derived.

A second advantage, that is also an advantage over Valentini and Dietterich's procedure, is that it allows greater control over the training sets sizes and inter-training-set variability. Let $|\cdot|$ represent the size of a data set. In general, k-fold cross validation will result in training sets of size $\frac{(k-1)}{k}|\mathcal{D}|$. Hence changes to k will result in changes to the training set size.

Bias and variance is evaluated with respect to individual objects. Under a single run of k-fold cross-validation, each object is classified once. Therefore, if there are l repetitions of k-fold cross-validation, each object will be classified l times. The training sets used to classify an object o under cross-validation cannot contain o. Hence, the l training sets used to classify o are drawn at random from $\mathcal{D} - o$. In consequence, each object $o' \neq o, o' \in \mathcal{D}$ has $\frac{k-1}{k}|\mathcal{D}|/(|\mathcal{D}| - 1)$ probability of being a member of any training set used to classify o and so the inter-training-set variability of the distribution used to generate the bias and variance estimates is as follows[1].

$$\delta \;=\; \left(1 - \frac{k-1}{k}\right) \times \frac{|\mathcal{D}|}{|\mathcal{D}| - 1} \tag{6}$$

[1] Note that we are here discussing the distribution of training sets used to classify a single object. The training sets generated by successive folds of a single cross-validation run will have a different distribution to the one we describe. However, each test object will only be classified by one test set from this distribution, and hence the distribution of training sets generated by a single cross-validation run is quite distinct from the distribution of interest.

$$= \frac{1}{k} \times \frac{|\mathcal{D}|}{|\mathcal{D}| - 1} \tag{7}$$

$$\approx \frac{1}{k}. \tag{8}$$

The approach has a further advantage over Valentini and Dietterich's procedure. It guarantees that all objects in $\mathcal{D}$ are classified the same number of times. This allows the experimenter to determine the number of classifications needed to obtain the desired level of precision in the estimates and to ensure that all objects are classified exactly that number of times.

However, while this procedure allows more variation in training set sizes and inter-training-set variability than either Kohavi and Wolpert's or Valentini and Dietterich's approach, the control over that variation is nonetheless still constrained. First, it is not possible to alter one of the two parameters without altering the other. Thus it is not possible to study, for example, the effect of altering the training set size while maintaining the inter-training-set variability constant. Second, it is not possible to reduce the training set size below $|\mathcal{D}|/2$ or increase the inter-training-set variability above 0.5. It would be desirable to allow finer-grade control over the training set size and the degree of variation between training sets.

2.5. An extension to the cross-validation procedure

With these considerations in mind we seek to develop a more flexible and reliable bias-variance estimation procedure. Our design objectives are as follows. We wish to have a procedure that takes as inputs a data set $\mathcal{D}$ and parameters m and δ that specify the training set size and the inter-training-set variability, respectively. We wish to derive as stable as possible estimates of distributions with the respective properties drawn from the data without excessive computation.

To maintain stability of estimates we believe that it is desirable to use as much as possible of $\mathcal{D}$ as both training and test data. To do otherwise can be expected to increase variance in the estimates due to variations in the samples that are used for the purposes of estimation. To this end we wish to use cross validation. However, simple cross validation does not allow the precise control over the training set size, m, that we desire. Each training set created by k-fold cross-validation has size $\frac{k-1}{k} \times |\mathcal{D}|$. We overcome this limitation by selecting a k such that $\frac{k-1}{k} \times |\mathcal{D}| > m$, and then sampling m objects from the objects available at each fold.

This mechanism allows us to precisely control m, but not δ. In general, to obtain an inter-training-set variability rate of δ with training set

size m using random sampling without replacement it is necessary to sample each training set from a pool of objects $\mathcal{D}'$ of size $m/(1-\delta)+1$.

From this it follows that we can generate a training set distribution with properties m and δ by selecting a data pool $\mathcal{D}'$ of size $m/(1-\delta)+1$ and selecting training sets of size m therefrom. We can achieve the latter by performing k-fold cross-validation such that $k > |\mathcal{D}'|/m$ and then randomly sampling a training set of size m from each of the non-holdout sets that are so formed. To minimize computation, it is desirable to choose the smallest value of k that satisfies that constraint.

However, to select a single $\mathcal{D}'$ would mean using only a subset of the data for training and testing, those included in $\mathcal{D}'$. More data can be utilized by simply creating a sequence of these subsets of $\mathcal{D}$. Any remaining data can be used for testing by adding it to the test set for one of the folds of one of the data subsets. Note that it is desirable that this additional test data only be added to one fold so that all test data are classified the same number of times. It is necessary that the additional data be classified only from training sets drawn from a single subset of $\mathcal{D}$ in order that the successive training sets used to form the classifiers by which it is classified are drawn from a distribution with the required properties.

The following sub-sampled cross-validation procedure, ssCV, instantiates these desiderata.

$ssCV(\mathcal{D}, l, m, \delta)$

$\mathcal{D}$**:** The data to be used.

l**:** The number of times each test item is to be classified, an integer greater than zero.

m**:** The size of the training sets, an integer $0 < m < |\mathcal{D}|$.

δ**:** The average proportion of objects to be shared in common between any pair of training sets, $m/(|\mathcal{D}|+1) \leq \delta < 1$.

1. Set the training pool size, $tps = \lceil m/\delta + 1 \rceil$.
2. Set the number of cross-validation folds $k = \lceil tps/(tps - m) \rceil$
3. Set the number of segments $q = \lfloor |\mathcal{D}|/tps \rfloor$
4. Randomize $\mathcal{D}$.
5. Partition $\mathcal{D}$ into q segments $E_1 \ldots E_q$ each of size tps with any remaining objects being placed in E_{q+1}
6. Repeat l times

For $i = 1$ to q do

a) Partition E_i into k random subsets $F_1 \ldots F_k$.

b) For $j = 1, k$

i) Select a random sample S of size m from $E_i - F_j$.

ii) For each $x \in F_j$

Record $\mathcal{L}(S)(x)$

iii) If $i = 1$ and $j = 1$ Then

For each $x \in E_{q+1}$

Record $\mathcal{L}(S)(x)$

7. Calculate the estimates of bias, variance, and error from the records of the repeated classifications for each object.

Under this approach, every object is classified once in every k-fold cross-validation for the training pool that contains it, and thus l times in total. The training sets used for successive classifications of an object are drawn from different cross-validation experiments, and hence are independently drawn random samples from the training pool. Thus, any two training sets used to classify an object will on average contain m/tps of their objects in common.

2.6. Comparison of the techniques

This sub-sampled cross-validation procedure is superior to both the holdout and bootstrap procedures in that it provides greater control of the degree of variability between training sets. Under the holdout technique, for any two training sets, each will on average contain half the objects contained by the other. Under the bootstrap approach any two training sets will contain on average 63.2% of objects in common, and some of these will be repeated multiple times in one or both sets. In contrast, the cross-validation technique provides fine-grained control over the degree of variability in the training sets.

A further property of this new procedure that is superior to that of the bootstrap procedure is shared with the holdout procedure. Unlike the bootstrap procedure, all objects are classified the same number of times. This eliminates wasted computation used to classify some objects more times than is necessary to obtain sufficient accuracy in the estimation of the bias and variance measures.

An additional superiority of this new procedure over the holdout procedure is shared with the bootstrap procedure. The training sets that are used can be much larger than those in the holdout approach. For the holdout approach, the training set cannot be more than half

the size of the training pool and the training pool cannot be larger than $\mathcal{D}$. Hence, $|\mathcal{D}|/2$ is an upper limit on the size of the training sets under the holdout approach. In contrast, under cross validation, the training-set size is $|\mathcal{D}| - |\mathcal{D}|/l$ and the maximum possible training-set size is $|\mathcal{D}| - 1$. So, for example, with 5-fold cross-validation and m set to the maximum value possible, each training set will contain 0.8 of the data. We believe that it will often be desirable to have larger training sets as this will more accurately reflect the distribution from which the training sets to which the learner will be applied in practice.

A final respect in which we expect the cross-validation and bootstrap procedures to be superior to the holdout procedure is that they use all objects in $\mathcal{D}$ as test objects rather than a sub-sample. This means that if objects are of varying degrees of difficulty to classify, all will be classified each time. This can be expected to produce much more stable estimates than the holdout method where the once-off selection of a test set E can be expected to impact the estimates obtained. We believe that such stability is a critical feature of a bias-variance estimation procedure. If estimates vary greatly by chance then it follows that any one estimate is unlikely to be accurate. Comparisons between learners on the basis of such unreliable estimates are likely to in turn be inaccurate.

3. Evaluation

All four of the factors identified in Section 2.6 are important. The first three do not require evaluation. From inspection of the procedures it is possible to determine that the specified features of the respective procedures indeed exist. However,the final of these factors does warrant investigation. To what extent, in practice, does the use of all available data in the role of test data lead to more stable estimates than the use of only a single holdout set? A second, and in our assessment more significant issue that we seek to evaluate is what effect if any is there from varying the type of distribution from which bias and variance are estimated.

To these ends we implemented the cross-validation technique in the Weka machine learning environment (Witten & Frank, 2000), which already contains the holdout approach. Unfortunately, we did not have access to an implementation of the bootstrap procedure. However, we assessed it less important to include this approach in our evaluation as its addition could contribute only to further understanding of the first and lesser issue, the relative stability of the estimates.

We applied the two approaches to the nine data sets used by Kohavi and Wolpert (1996). These are described in Table I. We use the hold-

Table I. Data sets

Dataset	**No. features**	**Dataset size**	**Train-set size**
Anneal	38	898	100
Chess	36	3196	250
DNA	180	3186	100
LED-24	24	3200	250
Hypothyroid†	29	3772	250
Segment	19	2310	250
Satimage	36	6435	250
Soybean-large	35	683	100
Tic-Tac-Toe	9	958	100

† Note, this appears to be a different version of the hypothyroid data set to that used by Kohavi and Wolpert. They report 25 attributes describing 3163 objects.

out approach with the training set sizes used by Kohavi and Wolpert. To compare the relative stability of the estimates we apply the cross-validation technique to each data set using the same training set size and inter-training-set variability ($\delta = 0.50$) as Kohavi and Wolpert.

We used $l = 50$, the number of repetitions used by Kohavi and Wolpert. To explore the consequences of reducing l we also applied our procedure with $l = 10$.

Each of the estimation processes was applied to each data set using each of two different learning algorithms, J48 and NB. J48 is a Weka reimplementation of C4.5 (Quinlan, 1993). NB is the Weka implementation of naive Bayes. In order to assess the stability of each estimation process, we repeated each ten times, using a different random seed on each trial. Thus we obtained ten values for each measure (error, bias and variance) for each combination of an evaluation process and a data set. Tables II to VII present the mean and standard deviation of these ten values for every such combination of process and data set.

3.1. Stability

The first hypothesis that we sought to evaluate is whether the cross-validation approaches are more stable than the holdout approach. To assess this we compare the standard deviations of all measures for both holdout and ssCV with $l = 50$. In every single case the standard deviation for the ssCV was lower than the standard deviation of the holdout method. For any one comparison of a method, learner, measure

Table II. Error for J4.8

	ssCV, $l = 10$		**ssCV,** $l = 50$		**holdout,** $l = 50$	
Data set	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.088	0.006	0.087	0.005	0.087	0.023
kr-vs-kp	0.050	0.003	0.049	0.002	0.049	0.008
dna	0.243	0.005	0.243	0.004	0.249	0.019
led-24	0.339	0.004	0.339	0.003	0.346	0.009
hypothyroid	0.021	0.001	0.022	0.001	0.021	0.005
segment	0.099	0.003	0.098	0.003	0.099	0.005
satimage	0.224	0.004	0.224	0.002	0.228	0.005
soybean	0.298	0.012	0.298	0.006	0.298	0.021
ttt	0.305	0.009	0.306	0.009	0.302	0.014

Table III. Bias for J4.8

	ssCV, $l = 10$		**ssCV,** $l = 50$		**holdout,** $l = 50$	
Data set	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.045	0.007	0.044	0.005	0.044	0.019
kr-vs-kp	0.029	0.002	0.029	0.002	0.030	0.005
dna	0.097	0.002	0.097	0.001	0.101	0.012
led-24	0.206	0.001	0.207	0.001	0.205	0.005
hypothyroid	0.012	0.001	0.012	0.001	0.012	0.002
segment	0.047	0.002	0.047	0.002	0.046	0.003
satimage	0.106	0.002	0.106	0.001	0.107	0.004
soybean	0.132	0.006	0.134	0.003	0.134	0.020
ttt	0.179	0.009	0.180	0.007	0.178	0.016

Table IV. Variance for J4.8

	ssCV, $l = 10$		**ssCV,** $l = 50$		**holdout,** $l = 50$	
Data set	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.039	0.006	0.041	0.005	0.042	0.007
kr-vs-kp	0.019	0.001	0.020	0.001	0.019	0.009
dna	0.132	0.004	0.143	0.004	0.146	0.013
led-24	0.119	0.003	0.130	0.002	0.138	0.011
hypothyroid	0.009	0.001	0.009	0.000	0.009	0.004
segment	0.047	0.002	0.050	0.002	0.051	0.004
satimage	0.107	0.002	0.117	0.001	0.119	0.007
soybean	0.149	0.007	0.162	0.004	0.160	0.008
ttt	0.113	0.009	0.123	0.008	0.122	0.023

Table V. Error for naive Bayes

	ssCV, $l = 10$		**ssCV,** $l = 50$		**holdout,** $l = 50$	
Data set	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.127	0.005	0.124	0.004	0.129	0.029
kr-vs-kp	0.159	0.003	0.159	0.002	0.160	0.021
dna	0.145	0.003	0.145	0.002	0.148	0.010
led-24	0.309	0.004	0.310	0.003	0.315	0.004
hypothyroid	0.052	0.001	0.052	0.001	0.054	0.003
segment	0.200	0.006	0.199	0.005	0.192	0.020
satimage	0.209	0.001	0.209	0.001	0.208	0.006
soybean	0.234	0.005	0.235	0.006	0.227	0.015
ttt	0.320	0.008	0.321	0.006	0.322	0.009

Table VI. Bias for naive Bayes

	ssCV, $l = 10$		**ssCV,** $l = 50$		**holdout,** $l = 50$	
Data set	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.077	0.005	0.075	0.004	0.079	0.030
kr-vs-kp	0.109	0.003	0.109	0.003	0.111	0.019
dna	0.082	0.003	0.082	0.002	0.083	0.008
led-24	0.212	0.003	0.213	0.003	0.215	0.005
hypothyroid	0.036	0.001	0.036	0.001	0.037	0.005
segment	0.148	0.007	0.147	0.005	0.139	0.029
satimage	0.186	0.002	0.186	0.002	0.184	0.005
soybean	0.148	0.006	0.149	0.005	0.139	0.014
ttt	0.232	0.008	0.232	0.007	0.236	0.011

Table VII. Variance for naive Bayes

	ssCV, $l = 10$		**ssCV,** $l = 50$		**holdout,** $l = 50$	
Data set	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.045	0.004	0.048	0.003	0.049	0.009
kr-vs-kp	0.046	0.001	0.049	0.001	0.048	0.004
dna	0.057	0.001	0.062	0.001	0.064	0.003
led-24	0.088	0.002	0.095	0.002	0.099	0.005
hypothyroid	0.014	0.001	0.015	0.001	0.017	0.003
segment	0.047	0.004	0.051	0.002	0.052	0.013
satimage	0.021	0.001	0.023	0.001	0.024	0.002
soybean	0.077	0.005	0.084	0.004	0.086	0.006
ttt	0.079	0.006	0.087	0.004	0.084	0.008

combination there are nine outcomes to compare: the outcomes for each data set. The probability of the cross-validation method having a lower value than the holdout method for every one of those comparisons if each method has equal chance of obtaining the lower value is 0.0020 (one-tailed binomial sign test). This result remains significant at the 0.05 level even if a Bonferroni adjustment for multiple comparisons is applied by dividing the alpha by 18. We therefore conclude that there is very strong support for our hypothesis that the cross-validation method that we have proposed is more stable in its estimates than the holdout method. This means that the experimenter can have greater confidence in the values that it produces.

The greater stability is produced at greater computational cost, however, as the cross-validation method learns $l \times k \times q$ models in comparison to the l models learned by the holdout method. To assess whether so much computation is required from the ssCV, we compare it with $l = 10$ against holdout with $l = 50$. The standard deviation for ssCV is still lower than that of the holdout method on all comparisons, demonstrating that more stable estimates can be derived at more modest computational cost.

3.2. Equivalence of results

It would be reassuring to know that there was some degree of consistency between the results of the two approaches. Comparing the results from holdout and ssCV (using $l = 50$) we observe considerable agreement in the error, bias, and variance. It is not possible to perform hypothesis tests to confirm a hypothesis that two population means are identical. Rather, it is only possible to perform hypothesis tests for the hypothesis that the means differ. Two-tailed t-tests with Bonferroni adjustment for multiple comparisons identify none of the 36 pairs of bias and variance means[2] as significantly different at the 0.05 level. Without Bonferroni adjustment, 4 of the 36 pairs of means are assessed as significantly different at the 0.05 level: the variance for J48 on led-24 ($z = -2.3645$), the bias of naive Bayes on soybean ($z = 2.1413$), and the variance of naive Bayes on dna ($z = -1.9755$) and led-24 ($z = -2.2875$). That none of the comparisons is significantly different once allowance is made for multiple comparisons is suggestive that there is not a substantial difference in the central tendency of the estimates of the two approaches.

[2] We consider only the bias and variance means as error is derived therefrom.

3.3. The effects on bias and variance of differing distributions

Having, we believe, obtained overwhelming support for our key hypothesis that the cross-validation technique provides more stable and hence reliable estimates, it is interesting to explore further issues that arise from the work. We have developed a bias-variance estimation process that allows greater control over the types of training set distributions with respect to which estimation can be performed. It is interesting to examine the consequences of this greater control.

Previous research has investigated the effects on bias and variance of altering training set size (Brain & Webb, 2002), using Webb's (2000) simple cross-validation bias-variance estimation technique. The study found that as training set size increases, variance decreases for both J48 and naive Bayes, as does bias for J48, but that the bias of naive Bayes can trend either up or down as training set size increases. Studies with our new bias-variance estimation technique corroborated the findings of that previous study.

However, the new technique also allows us to investigate new questions. Using this new capability we seek to investigate how bias and variance are affected if we hold the training set size constant but vary the inter-training-set variability. If we have two different processes that draw training sets of a given size at random from a single data source, we should expect the average error for a single learner applied to the resulting sequences of training set to be the same for each process. At any given setting of m each training set formed by ssCV is drawn at random from the data set as a whole. Every object in the data set has equal chance of inclusion in any training set. Hence, given a setting of m, we should expect error to remain constant as δ is varied.

However, we should expect changes to δ to affect bias and variance. Consider, for example, the case where $\delta = 1.0$. All training sets contain the same objects. For a non-stochastic learner, all models will therefore be identical. As a result, there will be no variance. The bias of a non-stochastic algorithm will equal the error (less the irreducible error, if allowance is made therefor) for such a distribution. As δ increases we should expect the variation in training sets and hence in models learned to increase, and hence expect variance to rise. Given that we expect error to be constant, it follows that bias must fall as inter-training-set variability increases.

To assess this hypothesis we augmented our study to include runs of ssCV with $\delta = 0.25$ and $\delta = 0.75$. For each data set we again used the training set size used by Kohavi and Wolpert and $l = 50$. The

Table VIII. Summary for error with J48 over differing training set distributions

	δ=0.75		δ=0.50		δ=0.25	
Dataset	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.088	0.004	0.087	0.005	0.087	0.003
kr-vs-kp	0.049	0.002	0.049	0.002	0.049	0.003
dna	0.242	0.004	0.243	0.004	0.244	0.003
led-24	0.339	0.003	0.339	0.003	0.337	0.003
hypothyroid	0.022	0.001	0.022	0.001	0.022	0.001
segment	0.098	0.002	0.098	0.003	0.098	0.003
satimage	0.224	0.001	0.224	0.002	0.224	0.002
soybean	0.301	0.010	0.298	0.006	0.300	0.010
ttt	0.304	0.006	0.306	0.009	0.306	0.008
Average	0.185	0.004	0.185	0.004	0.185	0.004

Table IX. Summary for bias with J48 over differing training set distributions

	δ=0.75		δ=0.50		δ=0.25	
Dataset	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.038	0.004	0.044	0.005	0.054	0.005
kr-vs-kp	0.026	0.001	0.029	0.002	0.034	0.002
dna	0.087	0.002	0.097	0.001	0.115	0.002
led-24	0.194	0.001	0.207	0.001	0.228	0.002
hypothyroid	0.011	0.001	0.012	0.001	0.015	0.001
segment	0.042	0.001	0.047	0.002	0.056	0.002
satimage	0.100	0.001	0.106	0.001	0.115	0.001
soybean	0.120	0.005	0.134	0.003	0.163	0.005
ttt	0.168	0.004	0.180	0.007	0.204	0.007
Average	0.087	0.002	0.095	0.003	0.109	0.003

results are presented in Tables VIII to XIII. For ease of reference we also include in these tables the results presented above for $\delta = .50$.

With respect to our prediction that error will not vary as inter-training-set variability alters for a given training set size, we are again faced with the inability of hypothesis testing to assess a hypothesis that population means are identical. Again we instead take a falsificationist approach and assess whether we can reject our hypothesis. To this end we perform 36 two-tailed t-tests comparing all error means for $\delta = 0.25$ against $\delta = 0.50$ and for $\delta = 0.50$ against $\delta = 0.75$. No value of z exceeded the cutoff of $z = 3.1971$ for significance at the 0.05

Table X. Summary for variance with J48 over differing training set distributions

	δ=**0.75**		δ=**0.50**		δ=**0.25**	
Dataset	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.049	0.003	0.041	0.005	0.032	0.005
kr-vs-kp	0.023	0.001	0.020	0.001	0.015	0.002
dna	0.152	0.002	0.143	0.004	0.127	0.003
led-24	0.142	0.003	0.130	0.002	0.107	0.002
hypothyroid	0.011	0.000	0.009	0.000	0.007	0.001
segment	0.055	0.002	0.050	0.002	0.042	0.002
satimage	0.122	0.001	0.117	0.001	0.107	0.001
soybean	0.178	0.010	0.162	0.004	0.135	0.008
ttt	0.133	0.007	0.123	0.008	0.100	0.008
Average	0.096	0.003	0.088	0.003	0.075	0.003

Table XI. Summary for error with NB over differing training set distributions

δ	**0.75**		**0.50**		**0.25**	
Dataset	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.123	0.004	0.124	0.004	0.125	0.005
kr-vs-kp	0.161	0.003	0.159	0.002	0.160	0.003
dna	0.146	0.002	0.145	0.002	0.144	0.002
led-24	0.310	0.002	0.310	0.003	0.311	0.003
hypothyroid	0.051	0.002	0.052	0.001	0.052	0.002
segment	0.198	0.006	0.199	0.005	0.199	0.008
satimage	0.209	0.001	0.209	0.001	0.209	0.001
soybean	0.229	0.004	0.235	0.006	0.233	0.008
ttt	0.322	0.005	0.321	0.006	0.321	0.008
Average	0.194	0.003	0.195	0.003	0.195	0.005

level with Bonferroni adjustment for multiple comparisons. Only one value $z = -3.0055$ for naive Bayes on the soybean data exceeded the cutoff $z = 1.9600$ for two-tailed significance at the 0.05 level without Bonferroni adjustment. Given the failure to obtain a significant result after allowance for multiple comparisons we are left without reason to reject our initial hypothesis that error will not change if training set size is kept constant while inter-training-set variability is altered.

We predicted that when δ is increased, variance should increase. This can be tested by hypothesis testing. A one-tailed t-test with Bonferroni adjustment for multiple comparisons is significant at the 0.05 level for

Table XII. Summary for bias with NB over differing training set distributions

	δ=0.75		δ=0.50		δ=0.25	
Dataset	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.064	0.005	0.075	0.004	0.090	0.005
kr-vs-kp	0.099	0.003	0.109	0.003	0.124	0.004
dna	0.069	0.002	0.082	0.002	0.098	0.002
led-24	0.196	0.002	0.213	0.003	0.240	0.003
hypothyroid	0.032	0.001	0.036	0.001	0.041	0.002
segment	0.135	0.006	0.147	0.005	0.162	0.010
satimage	0.181	0.001	0.186	0.002	0.192	0.002
soybean	0.127	0.006	0.149	0.005	0.174	0.007
ttt	0.213	0.003	0.232	0.007	0.259	0.007
Average	0.124	0.003	0.137	0.004	0.153	0.005

Table XIII. Summary for variance with NB over differing training set distributions

	δ=0.75		δ=0.50		δ=0.25	
Dataset	**Mean**	**s**	**Mean**	**s**	**Mean**	**s**
anneal	0.058	0.003	0.048	0.003	0.034	0.003
kr-vs-kp	0.061	0.002	0.049	0.001	0.035	0.002
dna	0.075	0.001	0.062	0.001	0.045	0.001
led-24	0.112	0.001	0.095	0.002	0.070	0.002
hypothyroid	0.018	0.001	0.015	0.001	0.011	0.001
segment	0.062	0.002	0.051	0.002	0.036	0.003
satimage	0.028	0.000	0.023	0.001	0.016	0.000
soybean	0.100	0.005	0.084	0.004	0.058	0.004
ttt	0.106	0.004	0.087	0.004	0.061	0.003
Average	0.069	0.002	0.057	0.002	0.041	0.002

35 of the 36 comparisons. The exception is for J48 on ttt for which the mean variance increases from 0.1234 to 0.1327 as δ increases from 0.50 to 0.75, but the t-test $z = 2.8106$ which falls short of the critical $z = 2.9913$ for the 0.05 significance level adjusted for 36 comparisons.

If error remains constant and variance increases then bias must decrease. A one-tailed t-test evaluation of this prediction is significant at the 0.05 level with adjustment for 36 comparisons in all 36 cases.

These effects on error, bias and variance of varying inter-training-set variability are illustrated in Figures 1 and 2 where the average of error, bias and variance is taken across all data sets. For both classifiers, as δ

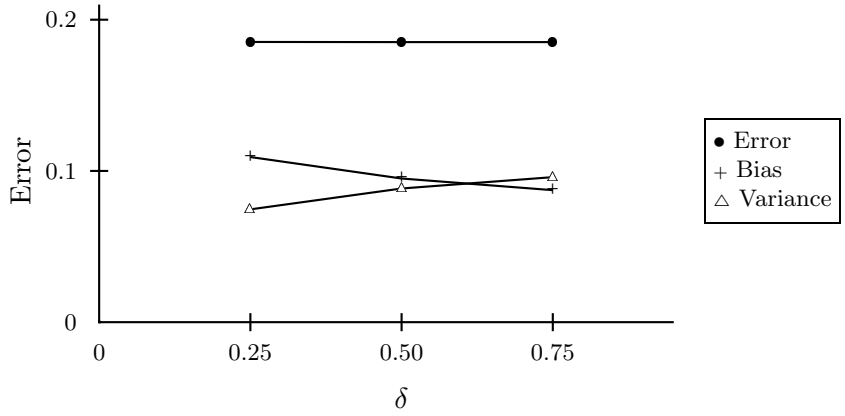

Figure 1. Comparison of error, bias and variance as δ is changed for J48.

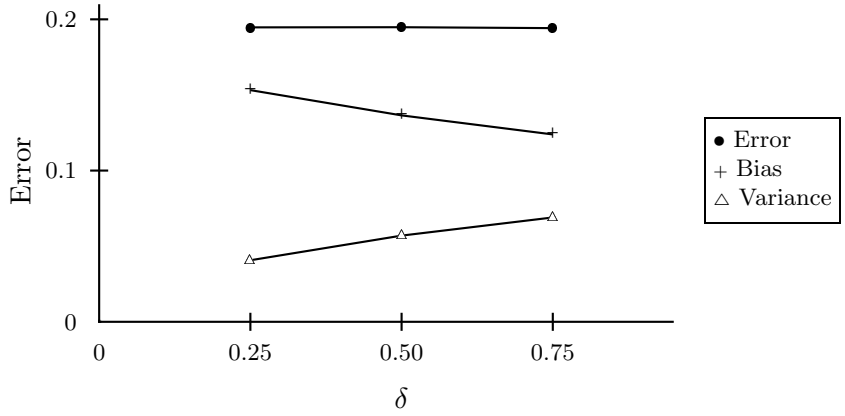

Figure 2. Comparison of error, bias and variance as δ is changed for Naive Bayes.

increases, the average error remains constant, the average bias decreases and the average variance increases.

4. Discussion

We believe that we have demonstrated that varying the type of distribution from which bias and variance are estimated will alter the results that are obtained. This has far reaching consequences for both explanations of classifier performance that rely on bias and variance and for how bias and variance experiments should be conducted. Bias-variance experiments to date have overwhelmingly been conducted using the

holdout procedure and hence have overwhelmingly come from distributions with 50% commonality between training set pairs. But we have shown that altering the distribution can alter the bias-variance results. This creates a cloud of doubt over the generality of any conclusions that have been drawn from previous bias-variance studies.

Turning to explanations of classifier performance based on bias-variance characteristics, we believe that the implications of our findings are quite profound. Consider Breiman's (1996a) account of bagging as a variance reduction mechanism. If bagging does indeed operate by reducing variance, then it seems reasonable to expect its efficacy to vary as the variance of the base learner to which it is applied alters. However, as we have seen, holding training set size constant while altering the inter-training-set variability can be expected to alter variance without altering error. Hence if we apply bagging to a learner under altering values of δ we should expect the variance of the base learner to alter, but the error of bagging to remain constant. There appears to be a deficiency in a straight forward account of bagging solely in terms of variance reduction.

As bias and variance characteristics can be expected to vary with variations in training set distribution it appears critical that bias and variance experiments should attempt to map out the dimensions of such variability by exploring a range of different distributions. We hope that our procedure might be of value for such purposes.

We accept, however, that there is value in having a single standardized point of comparison for the bias-variance characteristics of different algorithms, if for no other reason than that it is not feasible to compute, present or interpret the full space of possible distributions. For this purpose we propose the use of bias-variance evaluation using two-fold cross-validation (ssCV with $m = |\mathcal{D}|/2$ and $\delta = \frac{|\mathcal{D}|/2}{|\mathcal{D}|-1} \approx 0.5$). This approach is computationally efficient due to the use of the smallest possible number of folds, uses substantially larger training sets than can be supported by the holdout approach, and provides the reliability of estimation inherent in the ssCV approach.

5. Conclusions

We have argued that the holdout approach to bias-variance estimation has a number of serious drawbacks. First, it necessitates the use of relatively small training sets, meaning that typical bias-variance studies must examine quite different types of learning scenarios to those that occur in practice. Second, it provides no control over the type of distribution that is formed. All distributions will be such that any two

training sets will on average contain 50% of objects in common. Third, it is unstable, random variations resulting a large differences in the estimates produced.

We have presented an alternative procedure that address all of these problems, allowing fine-level control over training set size and the types of distributions while producing much more stable estimates.

Using this procedure we have demonstrated that altering the distribution of training sets with respect to which bias and variance are estimated alters the bias-variance performance that is observed. We believe that this has serious implications for studies that perform empirical evaluation of bias-variance profiles as the use of different distributions may result in different conclusions. In particular we have demonstrated that there are serious difficulties with any account of the efficacy of an algorithm that explains its error performance in terms of its ability to control variance. We hope that the procedure we have developed will enable to research community to obtain better understanding of the factors that influence classification bias and variance.

Acknowledgements

We are very grateful to Kai Ming Ting, Alexander Nedoboi and Shane Butler for valuable comments on drafts of this paper.

References

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, *36*, 105–139.

Brain, D., & Webb, G. I. (2002). The need for low bias algorithms in classification learning from large data sets. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002)*, pp. 62–73, Berlin. Springer-Verlag.

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, Berkeley, CA.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, *26*(3), 801–849.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Domingos, P. (2000). A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 231–238, Stanford University, USA.

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

Gama, J., & Brazdil, P. (2000). Cascade generalization. *Machine Learning*, *41*(3), 315–343.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–48.

James, G. (2003). Variance and bias for general loss functions. *Machine Learning*, *51*(2), 115–135.

John, G. H. (1995). Robust linear discriminant trees. In *AI & Statistics-95*, pp. 285–291.

Kohavi, R., & Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 275–283, San Francisco. Morgan Kaufmann.

Kohavi, R., Becker, B., & Sommerfield, D. (1997). Improving simple Bayes. In *ECML97 Poster Proceedings*.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324.

Kong, E. B., & Dietterich, T. G. (1995). Error-correcting output coding corrects bias and variance. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 313–321, San Francisco. Morgan Kaufmann.

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

Valentini, G., & Dietterich, T. G. (2003). Low bias bagged support vector machines.. In *Accepted for publication, International Conference on Machine Learning, ICML-2003*, Washington, DC.

Webb, G. I. (1999). Decision tree grafting from the all-tests-but-one partition. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 702–707, San Francisco, CA. Morgan Kaufmann.

Webb, G. I. (2000). Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, *40*(2), 159–196.

Witten, I. H., & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA.

Yang, Y., & Webb, G. I. (2003). Discretization for naive-Bayes learning: Managing discretization bias and variance. Tech. rep. 2003/131, School of Computer Science and Software Engineering, Monash University.

Zheng, Z., Webb, G. I., & Ting, K. M. (1999). Lazy Bayesian Rules: A lazy semi-naive Bayesian learning technique competitive to boosting decision trees. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, pp. 493–502. Morgan Kaufmann.