Small Sample Statistics for Classification Error Rates
II: Confidence Intervals and Significance Tests

J. Kent Martin and D. S. Hirschberg
(jmartin@ics.uci.edu) (dan@ics.uci.edu)
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425

Technical Report No. 96-22
July 7, 1996

Abstract

Several techniques for estimating the reliability of estimated error rates and for estimating the significance of observed differences in error rates are explored in this paper. Textbook formulas which assume a large test set, *i.e.,* a normal distribution, are commonly used to approximate the confidence limits of error rates or as an approximate significance test for comparing error rates. Expressions for determining more exact limits and significance levels for small samples are given here, and criteria are also given for determining when these more exact methods should be used. The assumed normal distribution gives a poor approximation to the confidence interval in most cases, but is usually useful for significance tests when the proper mean and variance expressions are used. A commonly used $\pm 2\sigma$ significance test uses an improper expression for σ, which is too low and leads to a high likelihood of Type I errors. Common machine learning methods for estimating significance from observations on a single sample may be unreliable.

1 Introduction

There is a substantial body of literature on estimating classifier error rates, and a clear consensus that some type of resampling technique is necessary to obtain unbiased estimates. In the companion paper [32] we dealt with methods for estimating a classifier's accuracy and the bias and variance of the estimates obtained from various methods. In this paper, we deal with confidence intervals, *i.e.,* the range of likely values of a classifier's true error rate given an estimated value, and with significance tests for the difference in the estimated error rates of alternative classifiers for the same population. The thesis of both papers is that *"...the traditional machinery of statistical processes is wholly unsuited to the needs of practical research ...the elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data." — R. A. Fisher [14] (1925)*

Among the significant findings reported in this paper are: (1) that the traditional formulas for error rate confidence intervals commonly found in introductory statistics textbooks assume an asymptotically large sample and are not accurate enough for machine learning research (an alternative formula is given), (2) that textbook formulas for significance tests are generally accurate enough, provided that the proper expression for the variance is used, (3) that there are many pitfalls in estimating the variance, leading to common mistakes in significance testing, and (4) that the common practice in machine learning research of estimating significance from observations on a single sample is unreliable.

Throughout this paper, the terms error and error rate (meaning misclassification rate) will be used interchangeably. The term bias, rather than error, is used to refer to a systematic difference between an error rate estimate and the true error (non-zero average difference). Also, the function $E(\cdot)$ denotes the expected (mean) value of a random variable, $\Phi(\cdot)$ the cumulative standard (zero mean, unity variance) normal distribution, $P\{\cdot\}$ a probability, and $f\{\cdot\}$ a probability density function.

1.1 Hypothesis Testing

In this section we provide a brief tutorial on the statistical inference issues relating to confidence intervals and significance tests, and on their common foundation, statistical hypothesis testing. We also give a short outline of the organization of the paper.

Given a classifier and an estimate of its error, the true error might be substantially higher or lower than the estimate. In view of this, the point estimate (single value) is of little utility unless its *reliability* is also somehow indicated. One way to do this is to give the standard deviation of the estimate's sampling distribution. Another way is to specify a *confidence interval*, a region which contains the relatively plausible values of the true error. When the sampling distribution is skewed (asymmetric), as is usually the case for error rates, a correctly defined confidence interval is more informative than the standard deviation.

Given two unbiased estimators, if one has a lower variance it has a greater *power* to discriminate between different classifiers and is the preferred estimator for that reason. If two unbiased estimators have equal power, the least expensive method is preferred. An unbiased estimator may sometimes be less powerful than a biased estimator if the bias is the same for all of the classifiers being compared and the biased estimator has lower variance than the unbiased estimator.

The reliability and power of the various estimators have received relatively little attention in the machine learning literature, as compared to the literature on estimating error rates. In the first paper [32], we were concerned with the applications of statistical inference for *estimation*: using sample characteristics to infer population characteristics, such as inferring a classifier and estimating its true error. The topics dealt with in this paper concern a different aspect of statistical inference, *hypothesis testing*: using sample information to answer questions about the population and the inferred classifier.

One such question is whether the classifier correctly predicts the classes. The various methods for estimating error can be thought of as alternative methods for assessing the truth of the hypothesis that the classifier's predictions are correct. If we knew or assumed that the population data were free of any measurement, observation, or labeling errors, then the occurrence of a single prediction error would serve to refute the hypothesis. If we know or can reasonably assume that the population data are imperfect, as is typically the case, then a single prediction error is not sufficient to refute the hypothesis (it could be that the prediction is right and the data are wrong). In the latter circumstance, we must accept or reject the hypothesis based on an inference regarding the strength of the contradictory evidence relative to the reliability of our data.

Another hypothesis that we frequently wish to test is that the true errors of two alternative classifiers are different, *i.e.*, that one classifier predicts more accurately than the other. This question is more conveniently posed as a test of the null hypothesis that the true errors are equal. Again, typically we must accept or reject the hypothesis based on an inference regarding the strength of the contradictory evidence relative to the reliability of our data.

Thus, the ability to answer the following two questions is particularly important: (1) how reliable is our estimated error, *e.g.,* within what interval is the true error to be found with a 95% (or 99%) likelihood? and (2) given another classifier having a different estimated error, how confident can we be that its true error is different from that of the first classifier?

We deal with the first of these questions, confidence intervals, in Section 2, dealing separately with traditional, textbook methods in Section 2.1 and with more exact methods derived from Bayesian analysis in Section 2.2 (a brief tutorial on the Bayesian methods is provided in Appendix A).

We deal with the second question, significance tests, in Section 3; presenting first, in Section 3.1, a common mistake which confuses the confidence level of a significance test with the confidence interval for an estimate. Sections 3.2 and 3.3 present more correct formulations: a traditional, textbook method for independent error estimates and a paired comparison method appropriate when estimates are not independent. Section 3.4 discusses the particular difficulties encountered in single-sample significance tests and illustrates the tendency to over-estimate significance inherent in common approaches to applying such tests. Section 3.5 discusses an argument proposed as a justification for single-sample tests, which we call the "dataset equals population" fallacy.

Section 4 briefly describes a related topic, overfitting avoidance[1], also known as pruning or the subset selection problem. Section 5 discusses ongoing efforts in machine learning to formulate more robust approaches to choosing classifiers, *i.e.,* methods which are more trustworthy than traditional significance tests. A summary of our significant findings and recommended methods is given in Section 6.

We present empirical data regarding significance tests in Appendix B. Section B.1 presents an extended example, comparing nearest neighbor and three nearest neighbor classifiers, which also

[1] The apparent error can be made arbitrarily low by considering very complex, *ad hoc* classifiers. This is called *overfitting* [48], described by CART [10] as inferring classifiers that are larger than the information in the data warrant, and by ID3 [39] as increasing the classifier's complexity to accomodate a single noise-generated special case.

illustrates several pitfalls in designing and analyzing such experiments (notably, use of biased error estimates and improper expressions for the variance). Section B.2 summarizes experiments using CART-style decision trees which test the generality of results in Section B.1. Section B.3 discusses experiments using iterated paired cross-validation as a possible single-sample significance test.

2 Confidence Intervals

As we have said, a classifier's true error might be somewhat higher or lower than the estimated rate, ϵ. We quantify this by specifying a confidence interval (τ_a, τ_b) such that this interval is expected to contain the true error (τ) with high likelihood (in at least 95% of our experiments, for instance). Typically, we also balance the risk on either side of the interval, *i.e.,*

$$P\{\tau_a < \tau < \tau_b \mid \epsilon\} \approx 0.95 \qquad P\{\tau \leq \tau_a \mid \epsilon\} \approx 0.025 \qquad P\{\tau \geq \tau_b \mid \epsilon\} \approx 0.025$$

These equations can be solved only if we specify a probability relationship between the true error and our observations. All of the confidence intervals given here assume that the number of errors is binomially distributed, *i.e.,* that the probability of m misclassified items in a random sample of size N from a distribution with a true error of τ is given by:

$$P\{m \mid \tau, N\} \;=\; \frac{N!}{m!\;(N-m)!}\;\tau^m\;(1-\tau)^{N-m} \tag{1}$$

such that the expected number of errors is $E(m) = N\tau$ and the variance is $\mathrm{Var}(m) = N\tau(1-\tau)$.

This assumption is commonly made (*e.g.,* [10, 27, 53, 54]), even though cross-validation methods are somewhat different from the simple random sampling scheme from which the binomial is derived. CART [10, pp. 78,307-308] notes that this binomial assumption works reasonably well, but is heuristic when applied to cross-validation.

Kohavi [27] gives a proof that k-fold cross-validation is unbiased and binomial if the classifier induction method is stable under cross-validation (*i.e.,* the k induced classifiers all make the same predictions). Commonly used induction methods such as decision trees should be reasonably stable, except in pathological cases, and stability should increase with the number of folds, k. Empirical data (Kohavi [27] and the companion paper [32], for instance) show that k-fold cross-validation estimates have a small bias that decreases rapidly with increasing k and increasing sample size N, and that their variance is nearly independent of k (a corollary prediction under stability [27]).

We performed a simple experiment as a further check that the binomial assumption is reasonable for 10-fold cross-validation of small samples: 1,000 samples of size $N = 100$ leading to discriminant classifiers with virtually identical true error ($\tau = 0.0203 \pm 0.0001$) were accumulated by repeatedly simulating samples from a population having two equally likely classes, each normally distributed on a single attribute, with the same standard deviation (σ) and with 2.053σ distance between the class means, until 1,000 classifiers in the target range were obtained (many simulated samples led to classifiers with true errors outside the narrow target, which were not included). The number of errors in 10-fold cross-validation of these selected samples was compared to the frequencies expected for a binomial with $N = 100$ and $\tau = 0.0203$. The differences were small ($\chi^2 = 9.05$, with 7 degrees of freedom, which is not statistically significant). Thus, for independent classifiers which have identical true errors, our assumption that Equation 1 describes the distribution of their 10-fold cross-validation error rates seems valid. We note that this says nothing about the error distribution when the classifiers are not independent or when their true errors differ.

2.1 Limits from a Normal Distribution Approximation

Throughout this section, τ is a classifier's true error, M the test set size, m the number of errors on the test set, and $\epsilon = m/M$ the estimated error. For k-fold cross-validation, take m to be the total number of errors on the k test sets and M to be the sample size, N.

Approximate confidence intervals for error rates are often derived by assuming that the binomial distribution of the integer number of errors (m) is approximately a normal distribution, with mean $M\tau$ and variance $M\tau(1-\tau)$. Or, equivalently, that ϵ is normally distributed with mean τ and variance $\tau(1-\tau)/M$. See *e.g.,* [4, pp. 340-341] or [27].

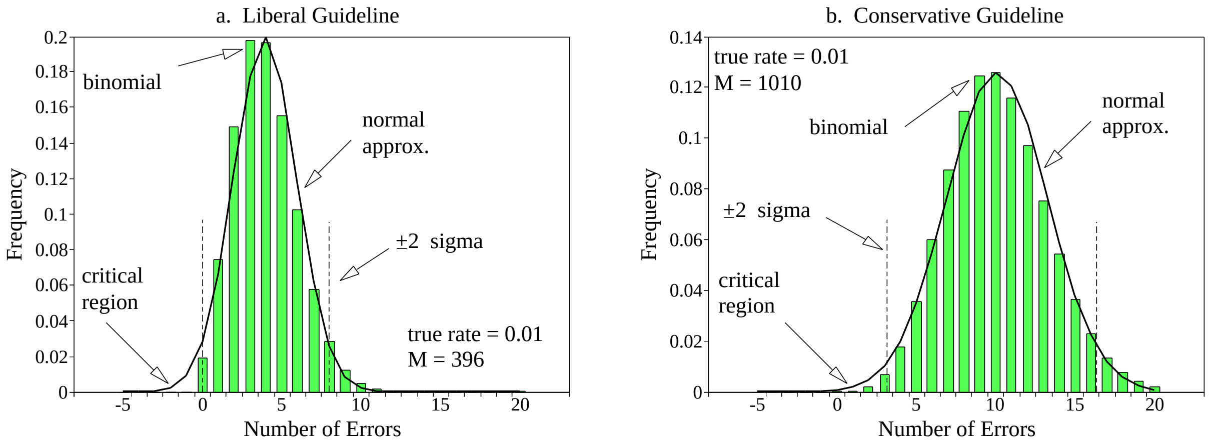
This normal approximation to the discrete binomial is valid (Hodges & Lehman [22, p. 187]) whenever $M\epsilon(1-\epsilon) \geq 10$, but can be quite misleading for small samples or low error rates. Since the usual goal is to achieve the smallest possible error, it is questionable whether these textbook limits are an adequate approximation to a good (less than 5% error) classifier's $(1-\alpha)$ confidence limits unless the test set is very large. In much machine learning research, test set sizes of 200 or less are the rule, and this is certainly borderline regarding validity of the normal approximation when the error is low. Breiman, *et al.* [10, p. 308], for instance, report that noticeably more than 5% of the data fall outside $\pm 2\sqrt{\epsilon(1-\epsilon)/M}$ limits.

Somewhat different guidelines for the validity of the normal approximation are given by different authors: Anderson & Sclove [4, p. 322] give this criterion as $M\epsilon \geq 5$ and $M(1-\epsilon) \geq 5$, while Mendenhall, *et al.* [33, p. 326] give the rule of thumb that the approximation is valid provided that $0 < \epsilon \pm 2\sqrt{\epsilon(1-\epsilon)/M} < 1$ and also that, in estimating the probability of an error of ϵ or less, we use the area under the normal curve below $\epsilon + 0.5/M$ (*i.e.,* a continuity correction). Note that the various guidelines are all functions of two variables (M and ϵ), not solely of the test set size M. The controlling factor is the number of errors observed, m, and we note that the binomial is symmetric and bell-shaped when $m \approx M/2$, but increasingly skewed as $m \to 0$ or $m \to M$.

The rule given by Hodges & Lehman is the more conservative, but requires about twice the minimum sample size than is implied by Anderson's rule (2.5× Mendenhall's minimum). The more conservative guidelines are more precise in the crucial tail of the distribution. This is illustrated in Figure 1: in Figures 1a and 1b, we see that the normal approximation is good in an overall sense under either rule, but better in the critical tail of the distribution under the more conservative rule. The x-axes ranges shown in Figure 1 include the absurdity of a negative number of errors. Use of the normal approximation implies that such a thing is possible; in fact, under the more liberal guidelines, that it has an appreciable (about 1 in 40) likelihood.

It is clearly inappropriate to use these approximate limits without first applying the tests for determining whether they are applicable, yet this is commonly the case. Three likely causes for this are that many texts and handbooks omit or do not stress criteria for applicability, that methods for estimating confidence limits when the normal approximation is not valid are beyond the scope of introductory texts, so the user is given no alternative limits, and that the $\pm 2\sigma$ 95% confidence limits rule is ingrained and its underlying assumptions are rarely recalled or questioned.

Figure 1: Liberal *vs.* Conservative Guidelines



2.1.1 Textbook Confidence Limits

A commonly used expression for the approximate $(1-\alpha)$ confidence interval is derived by Anderson and Sclove [4, pp. 340-341]:

$$\begin{aligned} P\{a < m\} &\approx \quad (\alpha/2) \approx \Phi\left(\frac{a + 0.5 - M\epsilon}{\sqrt{M\epsilon(1-\epsilon)}}\right) = \Phi(-z) \\ P\{m < b\} &\approx (1-\alpha/2) \approx \Phi\left(\frac{b - 0.5 - M\epsilon}{\sqrt{M\epsilon(1-\epsilon)}}\right) = \Phi(z) \end{aligned}$$

where (a, b) is the approximate $(1-\alpha)$ interval for m. Note that the interval is centered at $M\epsilon$, that the empirical variance $M\epsilon(1-\epsilon)$ has been substituted for the theoretical binomial variance $M\tau(1-\tau)$, and also that a 'continuity correction' of 0.5 has been made. Letting $z = \Phi^{-1}(1-\alpha/2)$ (*e.g.,* $z = 1.96$ for 95% confidence) and solving for a/M and b/M gives

$$\tau \approx \epsilon \pm \left(\frac{0.5}{M} + z\ s\right) \qquad \text{where } s = \sqrt{\epsilon\,(1-\epsilon)/M} \tag{2}$$

Equation 2 is commonly given in introductory texts [4, 10, 22, 42], and henceforth we refer to these limits as the *textbook limits.* The $0.5/M$ 'continuity correction' term is often omitted.

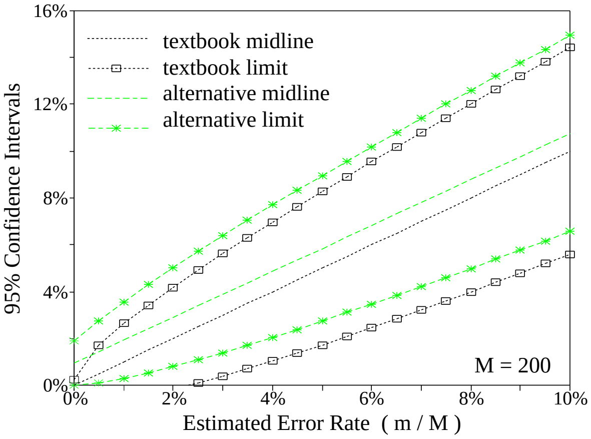
2.1.2 An Alternative Normal Approximation

Kohavi [27] describes an alternative derivation of $(1-\alpha)$ confidence limits from the normal approximation to the binomial (assuming a reasonably large test set[2]):

$$\frac{|\,\epsilon - \tau\,|}{\sqrt{\tau(1-\tau)/M}} < \Phi^{-1}(1-\alpha/2)$$

[2] Kohavi provides no specific guidelines for applicability. From this assumption and the assumption of approximate normality, guidelines similar to those discussed at the beginning of this section likely also apply here.

Figure 2: Textbook and Alternative Limits Compared



Note that, in contrast to the textbook limits, this formulation does not substitute the empirical variance $\epsilon(1-\epsilon)/M$ for the theoretical $\tau(1-\tau)/M$, nor does it apply a continuity correction. Letting $z = \Phi^{-1}(1-\alpha/2)$ and solving for τ gives

$$\tau \approx \left(\epsilon + \frac{k_1}{2M}\right) \pm z s' \qquad s' = \sqrt{\frac{\epsilon(1-\epsilon)}{M} + k_2} \qquad k_1 = \frac{(1-2\epsilon)z^2}{1+z^2/M} \tag{3}$$

$$k_2 = \left(\frac{z}{2(M+z^2)}\right)^2 \left[1 - 4\epsilon\,(1-\epsilon)\left(2+\frac{z^2}{M}\right)\right]$$

The numeric differences between these alternative limits and the textbook limits of Equation 2 are on the order of $1/M$, and lie in the sign and magnitude (k_1) of the $1/2M$ term, which emerges naturally here in solving the quadratic equation, and in the k_2 adjustment in the s' term.

The qualitative differences between the limits are profound. First, even though the normal approximation implicitly allows absurd negative error rates, these alternative confidence intervals never include the absurd values, even for very small samples where the guidelines for the textbook approximation are not met[3]. Secondly, both the midpoint and the width of the intervals are significantly different, even for moderately large samples when the more conservative guidelines are met. The alternative intervals are shifted upwards relative to the textbook intervals[4], as shown in Figure 2.

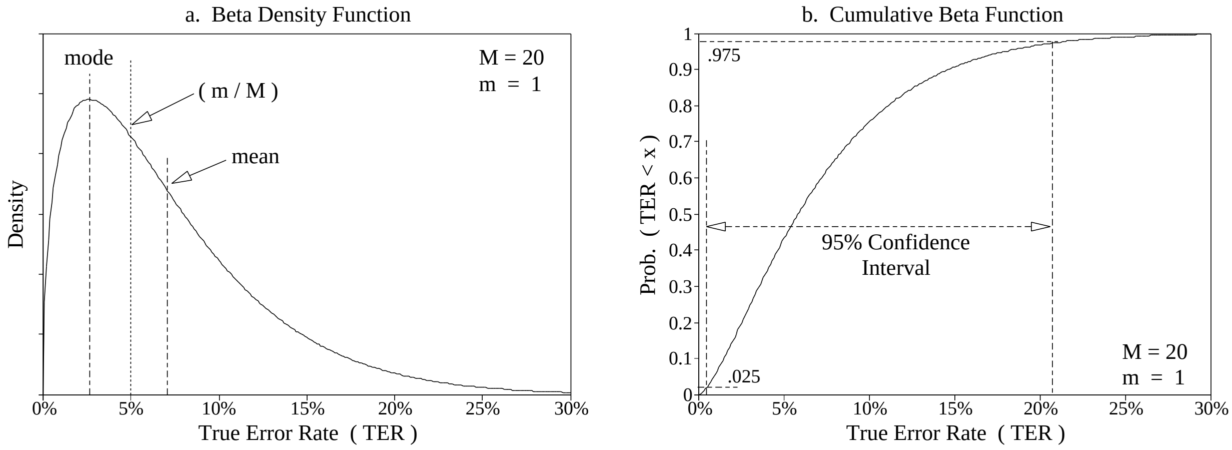
2.2 More Exact Confidence Limits

The binomial distribution (Equation 1) expresses the probability, $P\{m \mid M, \tau\}$, of m errors given the test set size, M, and true error, τ. Confidence intervals for τ require that one be able to answer such questions as how likely is it that τ is less than some particular value x, given M and m? That is, what is $P\{\tau < x \mid M, m\}$, the *posterior distribution* of τ? This posterior distribution depends on the likelihood of finding various values of τ, regardless of m/M. This *a priori*, unconditional probability

[3] Mendenhall's criteria for applicability of the textbook approximation share this feature. Note that this does not mean that the alternative intervals are necessarily accurate. The inclusion of absurd values by the textbook formulas in a particular case is a strong clue that the textbook intervals are not accurate in that case, but the absence of such absurdities does not guarantee correctness.

[4] Above 50% error they are shifted downwards (the midlines cross at 50%), but this has no practical significance.

Figure 3: The Jeffreys Beta Distribution



a. Beta Density Function

b. Cumulative Beta Function

function $f\{\tau\}$ is known as the prior distribution or simply the *prior* of τ. The relationship between the prior and posterior distributions is given by Bayes' Theorem[5]:

$$P\{\tau < x \mid M, m\} = \int_0^x P\{m \mid M, \tau\} f\{\tau\} d\tau \Bigg/ \int_0^1 P\{m \mid M, \tau\} f\{\tau\} d\tau$$

or by the derivative of this expression evaluated at $x = \tau$, the *posterior density function*:

$$f\{\tau \mid M, m\} = P\{m \mid M, \tau\} f\{\tau\} \Bigg/ \int_0^1 P\{m \mid M, \tau\} f\{\tau\} d\tau$$

Note that we use $P\{\cdot\}$ and $f\{\cdot\}$ generically to denote a probability or a probability density, respectively, without any intent to suggest that different arguments have the same distribution.

For a binomial proportion τ, Jeffreys' prior, $f\{\tau\} = \mathrm{Be}(\tau, 0.5, 0.5)$, has been shown to be the best choice for estimating confidence intervals in the absence of problem-specific knowledge (see Appendix A for a discussion of this prior, the uniform prior, and other priors). Jeffreys' prior gives a Jeffreys' Beta distribution, illustrated in Figure 3, as the posterior of τ:

$$\begin{aligned} P\{\tau < x \mid M, m\} &= I(x, m{+}0.5, M{-}m{+}0.5) \equiv \int_0^x \mathrm{Be}(\tau, m{+}0.5, N{-}m{+}0.5)\; d\tau \qquad (4) \\ \text{where} \qquad & \mathrm{Be}(\tau, u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}\; \tau^{u-1}(1-\tau)^{v-1} \end{aligned}$$

is the Beta probability distribution and $I(\tau, u, v)$ the Incomplete Beta function, with parameters u and v. $\Gamma(x)$ is the Gamma function, a generalization of the factorial. See Appendix A and the cited sources for information on these functions. The $(1{-}\alpha)$ interval can be found by solving (inverting) the Incomplete Beta function as shown in Figure 3b. For this particular Beta distribution, the posterior mean (μ) and variance (σ^2) of the true error are $\mu = (m{+}0.5)/(M{+}1)$ and $\sigma^2 = \mu(1 - \mu)/(M{+}2)$, and the mode (most likely value) is

$$\text{mode} = \begin{cases} 0 & \text{if } m = 0 \\ (m - 0.5)/(M - 1) & \text{if } 0 < m < M \\ 1 & \text{if } m = M \end{cases}$$

[5] See Appendix A. Iversen [21] provides a fairly non-formal introduction to Bayesian analysis. More formal treatments can be found in Box & Tiao [7] or Hartigan [19].

Note the apparent paradox that, while the expected value of the estimated error $\epsilon = m/M$ is equal to the true error, $E(m/M \mid \tau) = \tau$, the expected value of the true error given m/M is slightly different from m/M, $E(\tau \mid m, M) = (m+0.5)/(M+1)$. See the appendix for more discussion of this point. The variance σ^2 is larger than $s^2 = \epsilon(1-\epsilon)/M$ for low error rates, $m/M < 0.1$, and for very high error rates, $m/M > 0.9$, and is less than s^2 for $0.15 < m/M < 0.85$ (provided that $M \geq 4$).

Precise numeric inversion of the Incomplete Beta function can be very difficult in practice; see the appendix, where we also give a very accurate approximation to the inverse function. A reasonably close and computationally simpler approximation to the Beta distribution's 95% confidence limits is given by Center$(M, m) \pm$ Half-width(M, m), where we have found the following empirical formulas for Center(M, m) and Half-width(M, m):

$$\begin{aligned}
\text{Center}(M,m) &= A_M + (1-2A_M)\,\epsilon \qquad (5)\\
\text{Half-width}(M,m) &= \begin{cases} \text{Center}(M,m) & \text{if } m \leq 1 \\ 2B_M\sqrt{\epsilon(1-\epsilon)} & \text{otherwise} \end{cases}\\
\text{where} \qquad A_M &= \frac{1.96\sqrt{0.5}}{M+3} \quad \text{and} \quad B_M = \frac{1.96}{2\sqrt{M+2.5}}
\end{aligned}$$

We have compared these formulas to our precise numeric solutions for the inverse function at $M = 10, 20 \ldots 200$ and $m = 0, 1 \ldots M/2$. The largest absolute difference found was 0.017 at $M = 10$ and $m = 3$, and the absolute difference appeared to be $< 0.27/M$.

For $m \geq 2$ this expression is similar in form to the textbook and alternative limits, but has a different center and width:

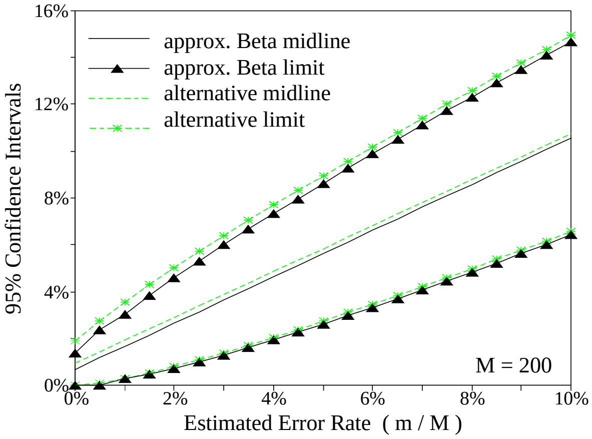
$$\begin{aligned}
\text{textbook:} &\quad \epsilon \pm \left(\frac{1}{2M} + 1.96\ s\right) \qquad \text{(see Equation 2)} \qquad (6)\\
\text{alternative:} &\quad \left(\epsilon + \frac{k_1}{2M}\right) \pm 1.96\ s' \qquad \text{(see Equation 3)}\\
\text{approx. Beta:} &\quad \left(\epsilon + \frac{g(M,m)}{2M}\right) \pm 1.96\ s''\\
\text{where} &\quad g(M,m) = 2(M-2m)A_M \quad \text{and} \quad s'' = \sqrt{\epsilon(1-\epsilon)/(M+2.5)}
\end{aligned}$$

The differences between the three expressions are on the order of $1/M$, and lie in the sign and magnitude of the $1/2M$ terms and in the magnitude of the half-width terms. The alternative and approximate Beta limits are compared in Figure 4 (and see Figure 2).

The preceding discussions illustrate the important point that the confidence interval for an error rate varies according to the prior $f\{\tau\}$, that is, according to one's knowledge, beliefs, prejudices, or assumptions as to the likely values of τ prior to having inferred a classifier or estimated its error. We note that the textbook and alternative limits implicitly assume a uniform prior. The confidence interval also depends very strongly on the test set size M, such that the interval becomes narrower as M increases and, importantly, such that the particular assumed prior becomes less important as M increases. For small samples, however, the influence of the assumed prior is very strong. Given the wide diversity possible even among the textbook methods, reported confidence intervals for error rates are of little use unless the method used to calculate them is explicitly stated (and, preferably, the underlying assumptions, as well).

For small samples, we recommend the approximate limits from the Beta distribution using Jeffreys' prior given in Equation 5.

Figure 4: Alternative and Approximate Beta Limits Compared



3 Significance Tests

Having established the range of plausible values for the true error given an estimate, we now shift focus to the second question posed in the introduction: given alternative classifiers for a population and estimates of their error, how confident can we be in asserting that one classifier predicts more accurately than the other?

The null hypothesis for comparing error rates is the hypothesis that the classifiers' true errors are equal. The *level of significance* α is the probability, given that the null hypothesis is true, of obtaining the observed difference or a more extreme value (in a two-sided test, a difference having greater magnitude). The level of significance is commonly expressed as its converse, $(1-\alpha)$, the *confidence level.* If the confidence level is sufficiently high (typically 95%), we reject the null hypothesis and assert that the true errors are different. If the confidence level is lower than our critical value, we accept the null hypothesis and assert that the true errors are equal. There is a risk associated with either assertion:

- In *Type I error* we reject the null hypothesis when it is true (wrongly assert that the true errors are different). α is a measure of the Type I risk.
- In *Type II error* we accept the null hypothesis when it is false (wrongly assert that the true errors are equal). The Type II risk is neither α nor $(1-\alpha)$, because the assessment of α explicitly assumes the proposition being asserted; see [22, pp. 370-376] for a discussion of this point. The Type II risk is usually not assessed.

The null hypothesis asserts only that the true errors are equal. In order to assess α, it is necessary to specify either the value or probability distribution of this common true error (τ) and, typically, neither of these is known. Any value obtained for α is thus conditional on whatever assumptions are made concerning τ and its distribution. All subsequent use of the symbol α should be understood to represent a conditional estimate of significance.

3.1 A Common Mistake

A common mistake in testing significance is to check whether one estimated error is outside the confidence interval of the other. This kind of comparison is invited by tabulated results such as "Method A: 58.1 ± 0.7%, Method B: 57.3 ± 0.7%". The difference of 0.8% between methods A and B may be significant at the 95% level, or it may not be, depending on the experimental conditions (*e.g.,* if the experiments used independent data, the difference is not significant even at a lower 90% level).

Comparing the higher estimate to the upper bound for the lower (by analogy to CART's 1-SE rule [10, pp. 78-80], which was developed for a more specialized use) is logically inappropriate for the following reasons:

1. It is a one-sided test — a one-sided test is appropriate only if it is *known* that one classifier has a lower true error than the other. If that were known, of course, there would be no point in making the comparison unless one were willing to accept a slightly higher error in exchange for, say, reduced complexity (which, to be fair to CART, is part of the context in which their 1-SE rule was proposed).

2. If the textbook confidence interval is used, this interval is too narrow for low error rates, which leads to a high likelihood of Type I error. The problem is particularly severe when the samples are small and the continuity correction is not used.

3. Even if the improved confidence interval given in the previous section is used, this is still not the proper formulation for this significance test, because the quantity being tested is the difference between the estimates. The variance of the difference of two random variables is the sum of their variances less twice their covariance. A very unique relationship between the two estimates is implied by the 1-SE rule[6]. While this relationship might be assumed to hold in 1-SE's narrow context (it seems a reasonable heuristic there), that is certainly not a valid assumption for all contexts.

In addition to the inappropriate analogy to the 1-SE rule, this mistake might also arise from confusing the 95% *confidence interval* or *confidence limits* with the 95% *confidence level* for the difference in two estimates.

The 1-SE rule was developed in the narrow context of selecting which members of a set of trees (derived by differently pruning a larger tree) have error rates comparable to the candidate which appears to be best, and should be evaluated for their complexity. The errors of this series of related trees are not independent, and it is difficult to know the distribution of the differences in estimated rates. Including a pruned tree in the set to be studied when its estimated error is within 1 standard deviation of the lowest rate found is a heuristic which should be judged empirically in the narrow domain for which it was intended [7]. The difficulty arises when something like the 1-SE

[6] CART applies the 1-SE rule in a series of n correlated estimates. Here, we restrict attention to $n = 2$. Under the null hypothesis, the two estimates, ϵ_1 and ϵ_2, are drawn from the same distribution, assumed normal with mean τ and variance $\sigma^2 = \tau(1-\tau)/M$. The difference between the two estimates is then also normal, with mean zero and variance $s^2 = 2\sigma^2(1-\rho)$, where ρ is their correlation coefficient. The normalized absolute difference, $|\,\epsilon_1 - \epsilon_2\,|\,/s$, is less than $Z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ with confidence level $1-\alpha$. For a k-SE (absolute difference $< k\sigma$) rule, we have $Z_{\alpha/2} = k\sqrt{2(1-\rho)}$, *i.e.,* the confidence level for this rule is a function of the correlation ρ. Or, assuming that the CART authors intended a uniformly high confidence level, their choice of a constant k implies a very restricted range of relatively high values for ρ, *e.g.,* $k = 1$ implies $\rho \approx 0.9$ for 95% confidence or $\rho \approx 0.8$ for 90% confidence.

[7] See Section 5 and the cited sources regarding the search for more robust classifier selection methods.

Table 5: Exact *vs.* Approximate Significance Levels

Normal Approx. Level	Nominal 99% Test for N=10			
	Exact Level from the Binomial			
	<90%	[90,95)	[95,99)	[99,100]
< 90%	66	2	0	0
[90,95)	0	0	0	0
[95,99)	4	1	12	0
[99,100]	1	1	10	24

Approx. Level	Nominal 95% Test for N=10			
	<90%	[90,95)	[95,99)	[99,100]
< 90%	66	2	0	0
[90,95)	0	0	0	0
[95,99)	4	1	12	0
[99,100]	1	1	10	24

Approx. Level	Conspicuous Errors for N=10			
	<90%	[90,95)	[95,99)	[99,100]
< 90%	66	2	0	0 *II
[90,95)	0	0	0	0
[95,99)	*I 4	1	12	0
[99,100]	1	1	10	24

Sample Size	Decision Accuracy			Conspicuous Errors	
	Nominal Level			Type	Type
N	90%	95%	99%	*I	*II
10	94%	94%	90%	4.1%	0
20	97	93	96	1.8	0
30	97	94	96	1.2	0
50	98	96	98	0.8	0
100	98	97	99	0.5	0

rule (specifically the +1.645-SE and ±2-SE rules), where $\mathrm{SE} = \sqrt{\epsilon(1-\epsilon)/M}$, is used as a one- or two-sided significance test.

3.2 A Textbook Significance Test

An approximate $(1-\alpha)$ confidence level test for the difference between independent error estimates ($\epsilon_1 = m_1/M_1$ and $\epsilon_2 = m_2/M_2$) is given [4, pp 412-415] by: $\alpha \approx 2\Phi(-Z)$, where $Z = |\epsilon_1 - \epsilon_2| / s$. Here, $s^2 = \tau_0(1-\tau_0)(1/M_1 + 1/M_2)$, where $\tau_0 = (m_1+m_2)/(M_1+M_2)$ is a weighted average error, M_i is the test set size for estimate i, and m_i is the number of errors found in test set i. When the null hypothesis is true, this textbook normal approximation is fairly good even when the underlying binomial distribution of the two estimates is far from normal. The approximation is poorer when the sample sizes are unequal and one of them is small.

The likelihood of various values of $(\epsilon_1 - \epsilon_2)$ under the null hypothesis can be calculated directly from the binomial probabilities for (M_1, τ_0) and (M_2, τ_0). This provides a means for estimating $(1-\alpha)$ for small test sets, even when the normal approximation is not good. We have compared confidence levels calculated in this way to those calculated from the textbook formula, as summarized in Table 5 for equal sample sizes. Partitioning the 121 m_1 and m_2 combinations for $N = 10$ as in a test at the 99% level, the approximate level leads to a different decision regarding the null hypothesis than would be made using the exact binomial calculation in 12 cases[8], *i.e.,* the decisions agree in 90% of the cases. At the 90% and 95% levels for $N = 10$, only 7 decisions are reversed, a 94% decision accuracy. Similar results were found for other sample sizes with increasing decision accuracy as the sample size increased. Table 5 also highlights conspicuous errors, *i.e.,* those where the approximate test asserts significance at the 95% level or above when the binomial calculation is less than 90%, or where the approximate test asserts non-significance at the 90% level when the binomial level is 95% or greater. The percentage of conspicuous errors is low and decreases rapidly with increasing sample size (it appears to be $O(N^{-1})$). The differences between the approximate and exact levels are predominantly in the direction of increased Type I risk.

Thus, the Type I risk is slightly higher than the nominal level when the textbook approximation is used. If we keep in mind that even the 'exact' calculations are conditional on our assumptions

[8] Summing the cases in the lower left and upper right submatrices of the 99% partitioning in Table 5.

that m_i is binomially distributed with true error τ_0, then the textbook formula seems accurate enough[9] for most purposes, provided that the significance levels are reported as being only approximate (strictly, we should report simply that we accept or reject the null hypothesis, based on an approximate test at the 0.05 level).

3.3 Paired Comparison Significance Tests

The methods in Section 3.2 are conditional on several assumptions in addition to normality and the null hypothesis: that the classifiers and their error estimates are independent (*i.e.,* inferred from independent data) and that the estimates are binomially distributed. As pointed out in Section 3.1, these methods are not appropriate when the classifiers and estimates are not independent. In this section we present a method which is appropriate whether or not the independence and binomial assumptions hold. In this and following sections, TER denotes a classifier's true error (the rate which would be observed were the classifier tested on the entire population), and subscripts 1 and 2 denote any two competing classifiers, *e.g,* a nearest neighbor (1-NN) classifier and a 3 nearest neighbors (3-NN) classifier.

Under the null hypothesis, the statistic $t=\overline{x}/s(\overline{x})$ is distributed approximately as Student's t, where $\overline{x}=\overline{\text{TER}}_2-\overline{\text{TER}}_1$ and $s(\overline{x})$ is the estimated standard deviation of $\overline{x}$. The method for estimating $s(\overline{x})$ and the number of degrees of freedom (dof) of the appropriate Student's t distribution depend on the experimental conditions (see [42, pp. 348-377], for instance). In the simplest experiments, $\overline{\text{TER}}_2$ and $\overline{\text{TER}}_1$ are each based on classifiers inferred from η equal-size random samples from the population, and there are η–1 degrees of freedom. When two different, independent sets of samples are used to infer the two types of classifiers, then $s(\overline{x})=\sqrt{(s_1^2+s_2^2)/\eta}$, where s_1 and s_2 are the unbiased[10] standard deviations of the TER's. When both classifier types are inferred from the same set of samples, a paired t-test is more appropriate, $s(\overline{x})=s(x)/\sqrt{\eta}$, where $s^2(x)=\sum_i(x_i-\overline{x})^2/(\eta-1)$, and $x_i=\text{TER}_{2,i}-\text{TER}_{1,i}$ is the observed difference for the i^{th} of η samples.

An extended example of a paired comparison is given in Appendix B.1, testing whether 3-NN classifiers are significantly more accurate than 1-NN classifiers. Similar tests comparing unpruned and pre-pruned (stopped) decision trees using nominal attributes data are discussed in Appendix B.2.

In a paired test, the paired classifiers and their error rates are not independent, since they are inferred from the same data. Assuming that the correlation is positive, the paired variance is lower than would be calculated using the unpaired formula[11]. It is crucially important that the variance be calculated properly; for example, in the 3-NN *vs.* 1-NN example given in Appendix B.1, mis-applying the unpaired formula to a paired comparison resulted in making the wrong decision regarding significance in 7 of 35 cases.

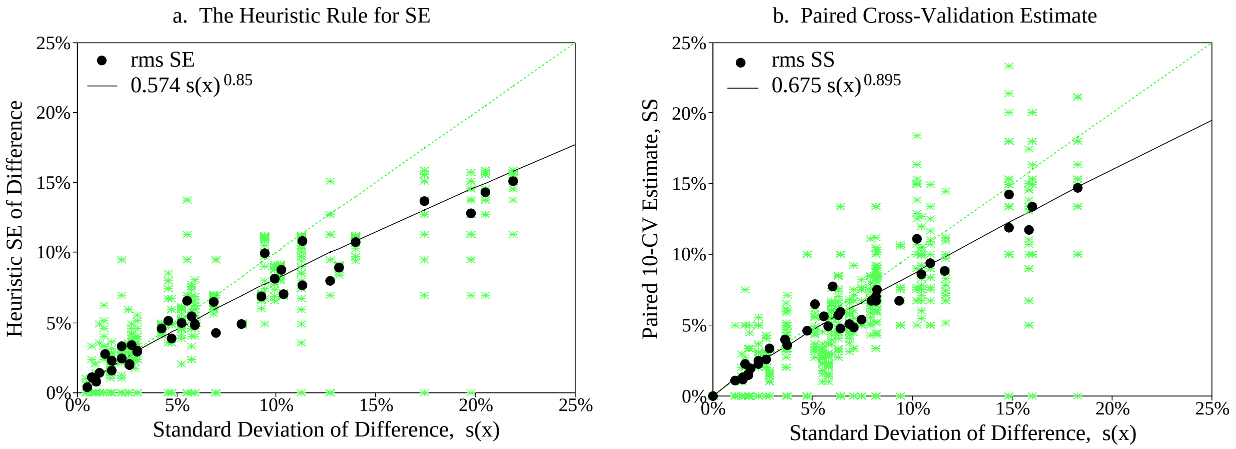
It is also very important that the error rate estimator be suitable for the comparisons being made. The 632b bootstrap estimator [13], for example, may be highly biased for nearest neighbors, and the bias is different for 3-NN than for 1-NN (see [24] [27] and [50] regarding the breakdown of 632b for nearest neighbors). In our comparison of 3-NN and 1-NN, use of this estimator would have led to the wrong decision regarding significance in half of the cases, and would have led to the wrong conclusion regarding the sign of the difference in 60% of the cases.

[9] That is, the textbook formula is *robust* (approximately correct, even under departures from normality).

[10] The unbiased standard deviation of η observations of a random variable x is given by $s_x=\sqrt{\sum_i(x_i-\bar{x})^2/(\eta-1)}$.

[11] The paired variance is $(s_1^2+s_2^2-2s_1s_2\rho)/\eta$, where ρ is the correlation coefficient. In the unlikely event that ρ were negative, the paired variance would actually be higher than would be calculated from the unpaired formula.

Figure 6: Single-Sample Estimates of Variance


a. The Heuristic Rule for SE

b. Paired Cross-Validation Estimate

3.4 Single-Sample Tests for Significance

In the analysis in Section 3.3, we assumed that we have the luxury of drawing many independent random samples from the population under study. In most real situations, there is but one small sample. In such a case, we can certainly infer two different classifiers and estimate the error of each (though the classifiers and their error estimates are hardly independent), but we cannot obtain from this single sample any direct measurement of the variance of the estimated error rates in general, nor of the variance of their difference, in particular. Can we, then, test whether the difference is or is not significant?

One approach would be to assume a value for the sampling variance. Weiss & Indurkhya [53, 54], for instance, adopt this approach in a ±2-SE test for pruned *vs.* unpruned decision trees. When two error rates (ϵ_1 and ϵ_2, with variances s_1^2 and s_2^2) are not independent, the variance of the difference between the rates is given by $s^2(x) = s_1^2 + s_2^2 - 2s_1 s_2 \rho$, where ρ is the correlation coefficient of the two rates, and the lack of independence means that $\rho \neq 0$. Under the null hypothesis, $\epsilon_1 \approx \epsilon_2 \approx \epsilon$, $s_1^2 \approx s_2^2 \approx s^2$, and $s^2(x) \approx 2s^2(1-\rho)$. Since it is not at all clear how to obtain valid estimates of s^2 and ρ from a single sample [10, p. 307], any heuristic for $s^2(x)$ must tacitly assume particular values for s^2 and ρ. The ±2-SE test assumes that $s^2 = \epsilon(1-\epsilon)/N$ and $\rho = +0.5$.

Is SE a reasonable estimate of the sampling standard deviation, $s(x)$? In our nearest neighbor comparisons, we used 20 independent samples for each paired comparison. Let $\epsilon_{1,i}$ and $\epsilon_{3,i}$ represent the 10-CV estimates for 1-NN and 3-NN, respectively, for the i^{th} sample,

$$\epsilon_i = (\epsilon_{1,i} + \epsilon_{3,i})/2 \qquad x_i = (\epsilon_{3,i} - \epsilon_{1,i}) \qquad \mathrm{SE}_i^2 = \epsilon_i(1-\epsilon_i)/N$$

and $s^2(x) =$ the variance of x_i. Figure 6a is a scatter plot of SE_i *versus* $s(x)$ for each of 40 different tests (varying the sample size and population inherent error). SE_i is highly variable and, as shown by the rms SE $= (\mathrm{mean}\{\mathrm{SE}_i^2\})^{1/2}$ values, is a biased estimator of $s(x)$. The rms SE values are well fit by a simple power function of $s(x)$, which is shown as the smooth curve in Figure 6a. Because SE_i is an optimistically biased estimate of $s(x)$, the $\pm 2\mathrm{SE}_i$ test entails a significantly greater Type I risk than the intended 0.05 level.

Examination of the data indicates that the assumption $\rho \approx +0.5$ is reasonable, but optimistically biased in this case, (*i.e.*, $s_1^2 + s_3^2 - s_1 s_3 \approx 0.76 s^2(x)$). An unbiased estimate of $s^2(x)$ (for these data) is given by $s_1^2+s_3^2-1.25s_1s_3$, or $\rho \approx 0.375$. The magnitude of the bias, $s(x)$ − rms SE, increases as SE increases. The individual SE$_i$ estimates are also highly variable, and their variance about rms SE is approximately proportional to N^{-1}.

While it is possible to infer a heuristic rule for estimating $s(x)$ given SE$_i$ from the power function in Figure 6a, we caution that the data underlying Figure 6a are all very similar and very simple — while the shape of the curve probably captures a general, qualitative relationship, the coefficients of a particular fitted function might not adequately describe the relationship for situations involving more complex data.

It is sometimes suggested that one might simply raise the threshold for rejecting the null hypothesis when using this SE$_i$ heuristic formula (*e.g.,* $|t| > 2.5$, rather than $|t| > 2.0$, for 95% confidence). If this is done, however, we feel that it would be misleading to report a 95% significance level or that $|t| > 2.5$. Rather, the result should simply be stated as apparently (heuristically) significant without quantifying it. How is the value 2.5 to be justified? Why not 4.0? Reporting a level or t-value under these circumstances would lend the analysis an undeserved aura of rigor.

A less biased estimate of $s(x)$ can be obtained from a single sample if a paired cross-validation is done. Though the data in Figure 6a are paired comparisons, the cross-validations for each sample are unpaired, because the cross-validation for 3-NN was done separately from that for 1-NN, using a different random partitioning. If only one partitioning is done, and the 1-NN and 3-NN error rates for the j^{th} train/test combination are paired ($\epsilon_{1,j}$, $\epsilon_{3,j}$), then the weighted variance of $\delta_j = (\epsilon_{3,j} - \epsilon_{1,j})$ provides a single-sample estimator for the sampling variance of the k-CV estimate, $\mathrm{SS}^2 \approx \sum_j (\delta_j - x)^2 m_j/(k-1)N$, where $x = \sum_j m_j \delta_j / N$. Figure 6b shows the individual and rms values of SS for 20 samples each for the 40 population/sample size combinations (as in Figure 6a). This estimator is less biased than SE, but biased nonetheless[12], and more variable than SE.

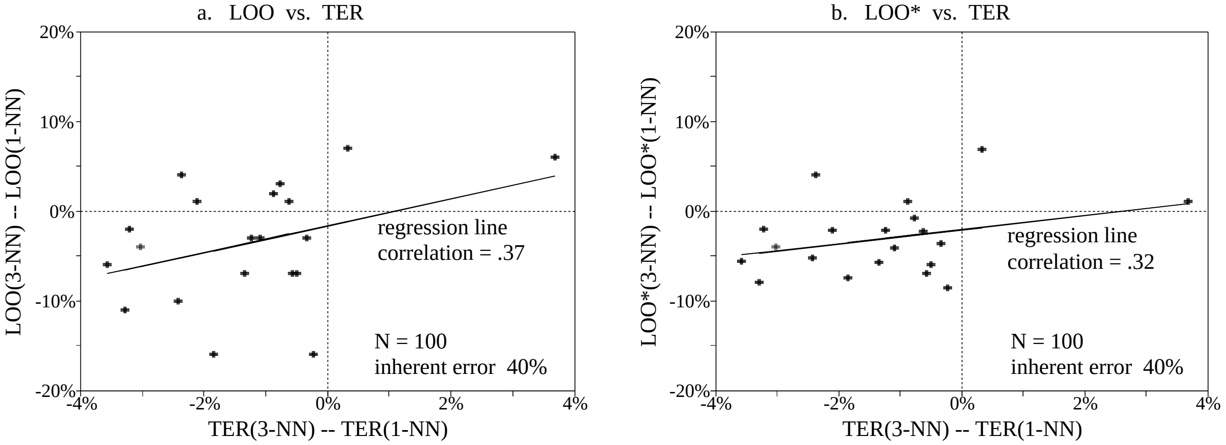
There are many other problems with these single-sample approaches, all deriving from the fallacy that conclusions about the differences between inference methods based on observations from a single sample are representative of the differences for the problem population at hand. Even for a single population and a fixed sample size, classifier error rates and the differences in paired error rates are so variable from one sample to another that we cannot draw a reliable inference even as to the sign of the difference from only one sample.

We illustrate this last point in Figure 7, where we show the paired differences for 20 samples of the same size from the same population ($N = 100$, inherent error 40%, where the average difference is negative and significant at the 95% level). In these figures we plot the paired difference in TER along the horizontal axis, and the paired difference of estimated error along the vertical axis. Only leave-one-out (LOO) and Weiss' LOO* estimator are shown, 10-fold cross-validation (10-CV) is very similar to both of these, but even more variable and less correlated. We can see that even here, where the average TER difference is strongest, the difference in estimated rates is poorly correlated with the difference in TER, highly variable, and apt to reverse the sign of the difference. LOO* is less variable[13] and less likely to reverse the sign or be grossly wrong as to the magnitude than LOO or 10-CV, but still poorly correlated with the difference in TER. Regardless of how we approach estimating $s(x)$, single-sample tests are apt to be misleading if we cannot be confident that at least the sign of the single-sample difference is correct. In the case of 3-NN *vs.* 1-NN (and, by analogy, pruned *vs.* unpruned decision trees), we cannot be sure of even that much, even for a population and sample size where the average difference is highly significant.

[12] This estimator is biased because, for any pair $(\delta_p, \delta_q) q \neq p$, 89% of the items in the two training sets are identical. These values simply are not free to vary as widely within a single sample as they would be from sample to sample.

[13] And, therefore, $\mathrm{SE} = \epsilon(1-\epsilon)/N$ is not an appropriate estimate of LOO*'s variance.

Figure 7: Single-Sample Differences Between 1-NN and 3-NN



It is known [32] that iterating the k-fold cross-validation of a sample many (typically 100) times significantly reduces the variance of the k-CV estimate for smaller k values, though at higher computational cost and with an increasing pessimistic bias as k decreases. This has naturally led to speculation that pairing the interated train/test partitioning of k-CV for competing classifiers might give both a lower variance for the estimated difference and a more reliable estimate of that variance.

Appendix B.3 gives details of experiments which iterated paired cross-validation 100 times (k-CV*) for 3-NN and 1-NN using both 2-CV* and 10-CV*. These methods did not lead to more reliable variance estimates or to more reliable 'significance' tests, despite the lower variance of 2-CV*. These methods also hold several pitfalls for the unwary user:

1. 2-CV* is strongly biased, and the bias may be different for different classifiers, which renders the comparison meaningless. This is certainly the case for 1-NN and 3-NN for very small samples and low error rates (see Appendix B.3 and Weiss [50]). This differing bias causes both the apparent difference between classifiers and the paired cross-validation estimate of the variance of the difference to be exaggerated.

2. Though the variance of k-CV* iterated m times is lower than that of uniterated k-CV, it is not reduced by a factor of $1/m$, as would be the case for m independent estimates. The variance of the k-CV* difference, $S^2(\overline{x})$, is lower than the variance, $S^2(x)$, of the mk individual test set differences, $S(x)/\sqrt{m} \leq S(\overline{x}) \leq S(x)$, but we know of no analysis that predicts where in this range $S(\overline{x})$ is to be found in a particular case. It appears to vary with the problem population and with the classifiers being compared. There is a tempting trap here: by using the incorrect formula $S(x)/\sqrt{m}$ one could, by making m sufficiently large, 'prove' significance for any case.

We have confirmed empirically (see Appendix B.2) that these concerns regarding the variance and lack of predictive correlation of error estimates also apply for decision trees using nominal attributes. Bailey & Elkan [5] have also noted this problem of high variation and poor correlation and suggested that it might be problematic as to the current machine learning approaches to comparing inference methods. Our experiments show that their misgivings are absolutely correct.

We know of no reliable method for projecting a measure of the internal (within-sample) variance of an error estimate to predict the sampling variance of the error rate. Nor do we believe such a method possible. The repeated internal estimates are not independent, nor are they free to vary as widely as the estimated error itself, even if the true error remained constant from sample-to-sample. This latter observation is the driving force behind the well-known $m-1$ degrees-of-freedom correction in for an unbiased estimate of the variance of a set of m independent observations. When observations are not independent; when we add the complications of possibly biased estimators, looking at the difference between classifiers, and possibly iterating the estimation many times, it is unlikely that any simple *a priori* correction will give an unbiased estimate of the sampling variance.

Thus, assertions from observations on a single sample such as "classifier X predicts this population more accurately than classifier Y, and the difference in accuracy is significant at the 95% confidence level" are very apt to be wrong on one or on both accounts.

3.5 The "Dataset Equals Population" Fallacy

Despite the problems raised in the last section (high variance, lack of correlation, and biased estimates of sampling variance), there is a persistent effort to somehow justify single-sample 'significance' tests. The motivations are understandable, the honest desire to do the best we can with what we have or the necessity to present some statement of significance to support our claims.

However, we feel that these single-sample statements of significance should not be presented as though they were even approximate statistical tests of significance at a certain (*e.g.,* 95%) confidence level. They are merely heuristic, and they typically entail a conspicuously higher Type I risk than is suggested by the nominal 'confidence level'.

Perhaps the most egregious arguments are the "dataset equals population and sampling variance is irrelevant" hypotheses, sometimes described as 'conditional significance'. Certainly, if we take the counter-factual position that the dataset 'defines' the population, then the sampling variance is irrelevant and all of the concerns we have raised disappear. If we take this specious argument literally, however, the whole business of resampling to estimate error and pruning to avoid overfitting is absurd. Under this hypothesis, the apparent accuracy of the classifier trained using the entire dataset *IS* the true accuracy, and the classifier-induction branch of machine learning research is largely out of business.

Of course, the proponents probably do not intend for this hypothesis to be taken that literally. Perhaps what they really propose is that we simply ignore the biases of SE and the paired cross-validation estimates of variance. Thus, the 'significance level' is contingent on the particular dataset (and ±2-SE is also contingent on the $\rho = +0.5$ assumption). In our opinion, significance claims of this kind should be clearly indicated as contingent, and the assumptions clearly stated.

These 'contingent significance' claims simply ignore the high variance and lack of predictive correlation of error estimates, *i.e.,* they ignore the fact that the claimed significance may not be replicable. For example, given two classifiers trained on the Iris data, if someone had returned to the Gaspé peninsula and measured another 150 flowers, then by testing the classifiers on this independent data we might well reach the opposite conclusion as to which classifier performs better than we reached using only Andersen's [3, 38] data.

There is currently great interest in machine learning in finding classifier selection methods that are more robust than traditional 'significance' tests (see Section 5). We hope that these methods will, when they have matured, obviate the need for heuristic contingent significance claims.

4 Pruning, or the Subset Selection Problem

All of the significance tests we have presented have focused on comparing the error rates of the final classifiers produced by different algorithms, rather than on the internal selection processes by which an algorithm arrives at its final classifier. In machine learning, probably the most familiar of these processes is decision tree post-pruning. Post-pruning is but one facet of a more general process known as *subset selection*, which also embraces processes as diverse as selection of splits during decision tree construction, stepwise discriminant analysis, and stepwise forward selection or backward elimination in fitting a logistic regression model. For an overview of subset selection and pruning, see the monograph by Miller [34], the Cart [10] and C4.5 [40] texts, and the articles by John, *et al.* [26], Mingers [35, 36], Schaffer [43, 44, 45, 46], and Weiss [51, 52, 53, 54].

Overfitting avoidance (pruning) is controversial. It has long been recognized that the "significance" tests used have a limited theoretical foundation and that they are biased [2, 34]. In machine learning, Schaffer [45] has described pruning as simply a form of bias. In applied statistics, Miller [34] cites descriptions of pruning or subset selection as 'unclean', 'distasteful', 'fishing expeditions', or 'torturing the data until they confess'.

Miller describes the available methods for empirical model selection as being based largely on folklore, with a dearth of respectable theory, or even of trustworthy advice. The so-called significance tests are part of that folklore. A related bit of folklore concerns variable interactions[14]; a common heuristic excludes the interaction of two variables if neither variable taken alone is to be included, on the empirical grounds that such relationships, *e.g.,* parity, are rare. In machine learning, this exclusion is structurally, rather than heuristically, a part of common (Cart [10] and ID3 [39]) decision tree algorithms. This exclusion renders it very difficult, if not impossible, to learn concepts similar to the exclusive-or. One consequence of this problem has been the adoption of various forms of look-ahead in decision tree algorithms. Another consequence has been a blanket condemnation of forward inclusion (stopping, or pre-pruning) methods and widespread adoption of post-pruning (backward elimination from an almost-certainly overfitted model). Look-ahead, the consideration of multivariate decisions, has much to recommend it. Trying to rationally prune a deliberately overfitted model is more questionable.

In agreement with Miller, we do not consider pruning or subset selection to be a well-posed question of statistical significance and, for that reason and because of the great variety of the pruning methods that have been proposed and the controversies they have sometimes stirred, detailed consideration of these topics is not within the scope of this paper. The so-called significance tests typically used in pruning are merely heuristic and, in our opinion, elaborate post-pruning schemes may be an over-embellishment of these heuristics. We believe that the sometimes considerable resources involved [30, 31] would perhaps be better utilized in look-ahead or in otherwise considering multiple, more diverse models.

5 Towards More Robust Classifier Selection

It is evident from the discussions in Sections 3 and 4 that traditional methods for choosing a classifier based on cross-validation estimates are very uncertain, and that many assertions of statistical significance for differences between competing classifiers entail a markedly greater Type I risk

[14] Two variables interact if the effect of changing one variable while holding the second constant depends on the value of the second, or *vice-versa.*

(wrongly asserting that the error rates are different) than indicated by the nominal significance level. This is so both because of the high variance of cross-validation and because commonly used methods for estimating variance from a single sample underestimate the sampling variance.

There is currently much research in machine learning focused on more robust methods for choosing a classifier. Central to all of this work is the concept of bias-variance tradeoff, *i.e.,* of averaging or aggregating predictions and estimates from many varied classifiers in order to reduce variance and, hence, increase confidence in the predictions and estimates. Given only a single sample, each of these constituent classifiers is necessarily inferred from a subsample and they tend to be increasingly biased as the subsample size decreases. On the other hand, the variance of the averaged or aggregated classifier tends to decrease as the subsample size decreases. Hence, the tradeoff between variance and bias.

The conventional thinking on bias-variance tradeoff, based on analogy to model selection in regression analysis, has been centered on a bias^2 + variance, or squared error loss, formulation. A recent paper by Friedman [16] casts this tradeoff in terms of a 0/1-loss function, *i.e.,* a classifier's prediction is either right or wrong, rather than smoothly varying. In this formulation, Friedman shows that the tradeoff is sometimes counter-intuitive relative to conventional squared error tradeoff, and that certain highly biased methods, *e.g.,* nearest neighbor, are nonetheless often highly competitive (see, for instance, Holte [20]).

A line of research being pursued by Breiman [8, 9] (see also [49, 57]) aggregates classifiers rather than averaging. In the approach known as 'bagging' (bootstrap aggregation), a single instance is held out as in leave-one-out and the remaining instances are bootstrapped (sampled with replacement to provide a training set) many times. The resulting classifiers 'vote' on the classification of the holdout instance. The outcome of these 'elections' varies far less from sample to sample than does the single prediction derived from training on the entire sample. In a similar approach [9], known as 'arcing' (adaptively resample and combine), proposed by Freund & Schapire [15], a weighted resampling is used, with an instance's weight, *i.e.,* its likelihood of being included in a training set, increasing as it is more frequently misclassified. At termination, the classifiers' predictions are combined by voting.

Wolpert [56] has proposed 'stacking' cross-validation, *i.e.,* taking the output predictions from the classifers inferred during cross-validation, together with the correct classifications, as input for cross-validating another inducer. These cross-validated inducers may be nested or stacked as deeply as desired, *e.g.,* until the predictions converge. This idea is readily generalized to combining output predictions from dissimilar classifiers, *e.g.,* neural nets and decision trees.

A related approach, but from a Bayesian point of view, is found in Raftery's [41] BIC (Bayesian Information Criterion) approach, which averages logistic regression classifiers over a small set (known as *Occam's window*) of best-fitting, *i.e.,* most likely, models.

All of these approaches, as well as earlier approaches to using multiple models (*e.g.,* [12, 23, 28]), seem to be useful for reducing variance and for selecting a better (more accurate) model. There is no clear winner among these approaches nor, given Schaffer's [47] conservation law for generalization, is it likely that any single method will be 'best' for all situations.

There are many open questions at this early stage for research in utilizing multiple models. One of these concerns aggregation methods, such as bagging, where it is not clear what the final model is (*i.e.,* how previously unseen instances are to be mapped onto a class prediction). Of more immediate interest to the concerns of this paper is that none of these methods appears, as yet, to have directly addressed questions of confidence intervals and significance tests, nor the issues of lack of independence, sample-to-sample correlation with true error, or underestimating sampling variance which we have identified in connection with cross-validation. While premature at present, we believe that these questions will become more important as this research area matures.

6 Conclusions and Recommendations

1. The textbook formula based on the normal approximation to the binomial is not a good approximation to the confidence interval of an error rate estimate for small samples or low error rates, even if a 'continuity adjustment' is made. When the number of observed errors is less than 10, the more exact limits calculated from the Beta distribution should be used.

2. The confidence interval for a single estimate (even the more exact Beta distribution limit) does not provide a good significance test for the difference between two estimated error rates. These limits, especially the ±2-SE limits, have a high additional Type I risk.

3. For comparing two independent rates on small samples, the textbook normal approximation entails a slightly greater Type I risk than the exact calculation of significance level from the binomial distribution. The textbook method is computationally simpler and its decision accuracy is typically 90% or better and, thus, it appears to be accurate enough for most purposes.

4. For paired comparison of classifier inference methods, Student's t test for the average difference over several independent samples is appropriate. The 10-CV estimator is recommended (see also Breiman & Spector [11]) because it seems to correspond most closely to the significance of the differences in TER. The 632b bootstrap and iterated 2-fold cross-validation (2-CV*) methods are not recommended because they are biased, and the bias of either method may differ for different inference methods. Weiss' LOO* estimator showed no advantage in our tests, and it is not recommended because of its high cost.

5. Paired comparison of inference methods based on a single sample may be misleading, even as to the sign of the difference, because the difference in error estimates is highly variable and poorly correlated with the difference in true error (see also Bailey & Elkan [5]), and because commonly used heuristics for the variance of the differences are biased and lack rigor. Thus, assertions such as "classifier X predicts this population more accurately than classifier Y, and the difference in accuracy is significant at the 95% confidence level" based on observations on a single sample are very apt to be wrong on either or on both accounts.

7 Acknowledgement

The authors are indebted to Dr. Sholom Weiss (Rutgers University) and Dr. Leo Breiman (University of California, Berkeley) for their assistance in sorting out issues relevant to single-sample significance tests. We are also grateful to the editors and reviewers of two earlier versions of these papers, whose suggestions have been invaluable in correcting many of the weaknesses and oversights.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, New York, 1972.

[2] A. Agresti. *Categorical Data Analysis.* Wiley, New York, 1990.

[3] E. Andersen. The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935. (cited by Morrison [37, p 468]).

[4] T. W. Anderson and S. L. Sclove. *The Statistical Analysis of Data.* Scientific Press, Palo Alto, 2nd edition, 1986.

[5] T. L. Bailey and C. Elkan. Estimating the accuracy of learned concepts. In *Proceedings of IJCAI-93*, pages 895–900, San Mateo, CA, 1993. Morgan Kaufmann.

[6] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society*, B41:113–147, 1979.

[7] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA, 1973.

[8] L. Breiman. Bagging predictors. Technical Report 421, Dept. of Statistics, University of California, Berkeley, 1994.

[9] L. Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Dept. of Statistics, University of California, Berkeley, 1996.

[10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Pacific Grove, CA, 1984. (Cart).

[11] L. Breiman and P. Spector. Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, 60:291–319, 1992.

[12] W. Buntine. Classifiers: A theoretical and empirical study. In *Proceedings of IJCAI-91*, pages 638–644, San Mateo, CA, 1991. Morgan Kauffman.

[13] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[14] R. A. Fisher. *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, 14th edition, 1970. (the quotation is from the preface to the first (1925) edition).

[15] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Machine Learning Conference*, 1996. (to appear, cited in Breiman [9]).

[16] J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. Technical report, Dept. of Statistics, Stanford University, 1996.

[17] J. A. Hartigan. The asymptotically unbiased prior distribution. *Annals of Mathematical Statistics*, 36:1137–1152, 1965.

[18] J. A. Hartigan. Note on the confidence-prior of Welch and Peers. *Journal of the Royal Statistical Society*, B28:55–56, 1966.

[19] J. A. Hartigan. *Bayes Theory.* Springer, New York, 1983.

[20] R. C. Holte. Very simple classification rules perform well on most commonly used data sets. *Machine Learning*, 11:63–91, 1993.

[21] G. R. Iversen. *Bayesian Statistical Inference.* Sage University Paper series on Quantitative Applications in the Social Sciences, no. 07-043. Sage Publications, Beverly Hills, 1984.

[22] J. J. L. Hodges and E. L. Lehman. *Basic Concepts of Probability and Statistics.* Holden-Day, Oakland, CA, 1970.

[23] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[24] A. K. Jain, R. C. Dubes, and C. Chen. Bootstrap techniques for error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9:628–633, 1987.

[25] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, A*, 186:453–461, 1946.

[26] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the Machine Learning Conference, 1994*, pages 121–129, San Francisco, 1994. Morgan Kaufmann.

[27] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of IJCAI-95*, pages 1137–1143, San Mateo, CA, 1995. Morgan Kaufmann.

[28] S. W. Kwok and C. Carter. Multiple decision trees. In R. D. Schacter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence, 4*, pages 327–335, Amsterdam, 1990. North-Holland.

[29] J. K. Martin. An exact probability metric for decision tree splitting. In D. Fisher and H-J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, volume 112 of *Lecture Notes in Statistics*, pages 399–410. Springer, New York, 1996.

[30] J. K. Martin and D. S. Hirschberg. The time complexity of decision tree induction. Technical Report 95-27, Dept. of Information & Computer Science, University of California, Irvine, 1995.

[31] J. K. Martin and D. S. Hirschberg. On the complexity of learning decision trees. In *Proceedings of the 4th International Symposium on Artificial Intelligence and Mathematics (AI/MATH-96)*, pages 112–115, Fort Lauderdale, 1996.

[32] J. K. Martin and D. S. Hirschberg. Small sample statistics for classification error rates, I: error rate measurements. Technical Report 96-21, Dept. of Information & Computer Science, University of California, Irvine, 1996.

[33] W. Mendenhall, D. D. Wackerly, and R. L. Scheaffer. *Mathematical Statistics with Applications.* PWS-KENT Publishing, Boston, 4th edition, 1990.

[34] A. J. Miller. *Subset Selection in Regression.* Monographs on Statistics and Applied Probability, series no. 40. Chapman & Hall, London, 1990.

[35] J. Mingers. An empirical comparison of pruning measures for decision tree induction. *Machine Learning*, 4:227–243, 1989.

[36] J. Mingers. An empirical comparison of selection measures for decision tree induction. *Machine Learning*, 3:319–342, 1989.

[37] D. F. Morrison. *Multivariate Statistical Methods.* McGraw-Hill, New York, 3rd edition, 1980.

[38] P. M. Murphy and D. W. Aha. *UCI Repository of Machine Learning Databases.* Dept. of Information and Computer Science, University of California, Irvine.

[39] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986. (ID3).

[40] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[41] A. E. Raftery. Bayesian model selection in social research. Technical report, University of Washington, 1994.

[42] S. Rasmussen. *An Introduction to Statistics with Data Analysis.* Brooks/Cole, Pacific Grove, CA, 1992.

[43] C. Schaffer. When does overfitting decrease prediction accuracy in induced decision trees and rule sets? In *Proceedings of the European Working Session on Learning (EWSL-91)*, pages 192–205, Berlin, 1991. Springer.

[44] C. Schaffer. Sparse data and the effect of overfitting avoidance in decision tree induction. In *Proceedings of AAAI-92*, pages 147–152, Cambridge, MA, 1992. MIT Press.

[45] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10:153–178, 1993.

[46] C. Schaffer. Selecting a classification method by cross-validation. *Machine Learning*, 13:135–143, 1993.

[47] C. Schaffer. A conservation law for generalization performance. In *Proceedings of the Machine Learning Conference, 1994*, pages 259–265, San Francisco, 1994. Morgan Kaufmann.

[48] J. W. Shavlik and T. G. Dietterich, editors. *Readings in Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1990.

[49] R. Tibshirani. Bias, variance and prediction error for classification rules. Technical report, Dept. of Statistics, University of Toronto, 1996.

[50] S. M. Weiss. Small sample error rate estimation for k-nearest neighbor classifers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:285–289, 1991.

[51] S. M. Weiss and N. Indurkhya. Reduced complexity rule induction. In *Proceedings of IJCAI-91*, pages 678–684, San Mateo, CA, 1991. Morgan Kaufmann.

[52] S. M. Weiss and N. Indurkhya. Optimized rule induction. *IEEE Expert*, 8:61–69, 1993.

[53] S. M. Weiss and N. Indurkhya. Decision tree pruning: Biased or optimal? In *Proceedings of AAAI-94*, pages 626–632, Menlo Park, CA, 1994. AAAI Press.

[54] S. M. Weiss and N. Indurkhya. Small sample decision tree pruning. In *Proceedings of the Machine Learning Conference, 1994*, pages 335–342, San Francisco, 1994. Morgan-Kaufman.

[55] B. L. Welch and H. W. Peers. On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society*, B25:318–329, 1963.

[56] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[57] D. H. Wolpert and W. G. Macready. An efficient method to estimate bagging's generalization error. Technical Report SFI-TR-96-03, Santa Fe Institute, 1996.

Appendices

A Bayesian Analysis for Confidence Intervals

In this appendix, we provide a review of Bayesian analysis applied to estimating confidence intervals. Following a brief introduction to Bayesian methods in Section A.1, Section A.2 introduces the family of Beta distribution priors for the binomial. Section A.3 presents Jeffreys' Beta prior, which has been shown to be the best choice for estimating confidence intervals. Section A.4 summarizes arguments against use of the uniform prior, and presents empirical data on the prior distribution of problems frequently used to evaluate classification algorithms in machine learning research.

A.1 Bayesian Analysis for the Binomial Distribution

What is the *posterior* distribution of the true error τ, $P\{\tau < x \mid M, m\}$, given a test set of M items and that the observed number of errors m is binomially distributed:

$$P\{m \mid M, \tau\} = \frac{M!}{m!\,(M-m)!}\ \tau^m\,(1-\tau)^{M-m}$$

The expected value (mean) of m/M is $E(m/M) = \tau$, and its variance is $\sigma^2 = \tau(1-\tau)/M$. From this, one might surmise that the expected value of τ given m/M would be m/M, but this is not so because it is possible to obtain the result $m = 0$ when $\tau \neq 0$:

$$\begin{aligned}
P\{m = 0 \mid M = 10, \tau = 0.0\} &= 1.0000\\
P\{m = 0 \mid M = 10, \tau = 0.1\} &= 0.3487\\
P\{m = 0 \mid M = 10, \tau = 0.2\} &= 0.1074\\
&\vdots\\
P\{m = 0 \mid M = 10, \tau = 1.0\} &= 0.0000
\end{aligned}$$

The expected value of τ given m/M (in particular, and the posterior distribution in general) depends on the assumed *prior* distribution $f\{\tau\}$ via Bayes' Theorem:

$$P\{\tau < x \mid M, m\} = \int_0^x P\{m \mid M, \tau\} f\{\tau\} d\tau \ \Big/ \int_0^1 P\{m \mid M, \tau\} f\{\tau\} d\tau$$

$$f\{\tau \mid M, m\} = P\{m \mid M, \tau\} f\{\tau\} \ \Big/ \int_0^1 P\{m \mid M, \tau\} f\{\tau\} d\tau$$

Note that the normalizing factor (the denominator) of these expressions can also be expressed as $P\{m \mid M\}$, the prior unconditional probability of m errors in a test set of size M.

Many analysts find the dependence of the results on an assumed prior $f\{\tau\}$ troubling, as it affords the opportunity to interject subjective opinion into the analysis. However, as noted in the companion paper [32], the problem of drawing inferences about τ (such as confidence intervals) is ill-posed, and it is necessary to assume some prior distribution for the kinds of problems one is likely to encounter. We also note that the textbook and alternative approximate confidence intervals discussed in the paper implicitly assume a uniform (rectangular) prior distribution for τ in addition to their normality assumptions. This implicit uniform prior assumption is common in classical statistical inference. Obviously, an important topic in Bayesian analysis has been the formulation of priors which add little information to the sample information [6], *i.e., non-informative priors*[15], or express our relative ignorance of the distribution of τ [17], *i.e., prior densities on ignorance.*

A key principle of Bayesian analysis [19, pp. 34-39] [21, pp. 59-70] is that the impact of the assumed prior diminishes as more evidence is accumulated (as M increases) and that, for non-informative priors, Bayesian and classical analyses should asymptotically converge to the same numerical results (the results may be interpreted somewhat differently). Hartigan [17] emphasizes *asymptotically unbiased priors*, *i.e.,* those for which the associated estimator converges asymptotically to the true value of the quantity being estimated. Bernardo [6] and Hartigan [17] both note that the choice of the relevant prior differs according to the quantity of interest, *e.g.,* according to whether one is estimating a mean, a variance, or a binomial proportion.

For the small samples which are the topic of this paper, asymptotic convergence and unbiasedness, while crucial properties of a prior, are not sufficient in themselves. We must also be concerned with the relative efficiency (rate of convergence) and non-informativeness of the prior; *i.e.,* following Bernardo [6], a reference prior should maximize the expected information about τ to be provided by the sample data and, thus, minimize the impact of the prior assumptions.

A.2 Beta Distributions

It is convenient if $f\{\tau\}$ can be expressed in such a form that Bayes' Theorem is easily integrated; such priors are sometimes called *conjugate priors.* The Beta distribution with parameters u and v, $\mathrm{Be}(\tau, u, v)$, is a family of conjugate priors (sometimes called *Beta* or *Dirichlet priors*) for the binomial which are capable of expressing, to at least a very good approximation, a very wide variety of plausible priors for τ (see Iversen [21, pp. 18-33]).

$$\begin{aligned} \mathrm{Be}(\tau, u, v) &= \tau^{u-1}(1-\tau)^{v-1} / B(u,v) \\ \text{where} \qquad & B(u,v) = \int_0^1 \tau^{u-1}(1-\tau)^{v-1} d\tau = \Gamma(u)\Gamma(v) \,/\, \Gamma(u+v) \quad \text{is the Beta function} \end{aligned}$$

$\Gamma(z)$ is the Gamma function, a generalization of the factorial. For positive integers n, $\Gamma(n) = (n-1)!$. In general, $\Gamma(z) = (z-1)\Gamma(z-1)$ and, in particular, $\Gamma(1/2) = \sqrt{\pi}$. See [1, pp. 255-258,944-945] for information on these functions. These Beta priors give a Beta distribution as the posterior:

$$\begin{aligned} f\{\tau \mid M, m\} &= \mathrm{Be}(\tau, m+u, M-m+v) \\ P\{\tau < x \mid M, m\} &= I(x, m+u, M-m+v) = \int_0^x \mathrm{Be}(\tau, m+u, M-m+v)\, d\tau \end{aligned}$$

where $I(x, m+u, M-m+v)$ is the Incomplete Beta function. For this Beta distribution, the posterior mean (expected value, μ) and variance (σ^2) of the true error rate τ are

$$\mu = (m+u)/(M+u+v) \qquad \sigma^2 = \mu(1-\mu)/(M+u+v+1)$$

and the most likely value (the mode) is

$$\text{mode} = \begin{cases} 0 & \text{if } m=0 \\ (m+u-1)/(M+u+v-2) & \text{if } 0<m<M \\ 1 & \text{if } m=M \end{cases}$$

[15] Bernardo [6] terms these *reference priors*, emphasizing their possible use as a standard to assess the relative importance of a particular assumed prior.

A.3 The Jeffreys' Beta Distribution

Assuming only quite general regularity conditions[16], Bernardo [6], Hartigan [19], and Welch & Peers [55], among many others, justify the family of prior distributions known as *Jeffreys' [25] prior.* Bernardo [6] shows that this prior maximizes the expected information provided by the sample, Hartigan [17] shows that this prior is asymptotically unbiased, and Welch & Peers [55] show that confidence intervals generated from Jeffreys' prior are asymptotically closer to providing the targeted confidence level than those of any other prior.

For binomial confidence limits, Jeffreys' prior is a Beta distribution, $f\{\tau\} = \mathrm{Be}(\tau, 1/2, 1/2)$, and leads to an Incomplete Beta function as the posterior:

$$P\{\tau < x \mid M, m\} = I(x,\ m{+}0.5,\ M-m{+}0.5) \tag{7}$$

$$= \left(\frac{x^{m+0.5}(1-x)^{M-m+0.5}}{(m{+}0.5)\ B(m{+}0.5, M-m{+}0.5)}\right) \left(1 + \sum_{i=0}^{\infty} \frac{B(m{+}1.5, i{+}1)}{B(M{+}1, i{+}1)}\ x^{i+1}\right)$$

The derivation by Welch & Peers [55] shows this solution to asymptotically provide an actual confidence level of $(1-\alpha) \pm O(M^{-1})$ for Jeffreys' prior, while other priors converge only to $(1-\alpha) \pm O(M^{-1/2})$. Thus, for sufficiently large M, we expect the Jeffreys' prior intervals to be both more accurate and more efficient (converging more quickly) than those for any other prior[17].

We caution that this is not a proof, since the orders of magnitude $O(\cdot)$ here should be qualified as "in the probability sense" and, in particular, because these asymptotic results say little or nothing about small-sample behavior. Our confidence in the small-sample behavior of Jeffreys' prior stems from Bernardo's [6] results showing that it minimizes the contribution of the prior. Also, following Welch & Peers [55], we emphasize that we do not wish to imply that only formal Bayesian solutions are allowable for calculating confidence intervals, or that they are necessarily "best". For binomial confidence intervals, however, they are certainly better founded than the highly questionable (*e.g.*, for $m = 0$) assumption of normality.

For large values of M and m, the series in Equation 7 converges very slowly. Solutions to determine the confidence interval are very sensitive to numerical precision errors, including the use of Stirling's [1, p. 257] approximation for $\Gamma(z)$. For $m > M/2$, advantage can taken of the symmetry of the Incomplete Beta function, $I(x, u, v) = 1 - I(1-x, v, u)$. For $m \gg 0$, advantage can be taken of the recurrence relation $I(x, u, v) = x\ I(x, u-1, v) + (1-x)\ I(x, u, v-1)$.

For our 'precise' calculations of $I(\cdot)$ we used IEEE-standard double floating point, a look-up table for $\ln\Gamma(i + 0.5), i \leq 500$, and Stirling's approximation for $\ln\Gamma(i + 0.5), i > 500$. To estimate the inverses $I^{-1}(\alpha/2)$ and $I^{-1}(1-\alpha/2)$, we used binary search to locate $x = I^{-1}(y)$ to within $y \pm 1.0 \times 10^{-8}$.

Abramowitz & Stegun [1, p. 945] give an approximation for the inverse of the incomplete Beta function which translates into the $(1-\alpha)$ ± confidence limits for (M, m) as:

$$\tau \quad = \quad \frac{a}{a + b\exp(\omega_2 \mp \omega_1)} \qquad \text{where}$$

[16] Described by Bernardo [6] as "the usual regularity conditions for asymptotic normality of the posterior", these are enumerated by Hartigan [17, pp. 1141-1142], and are principally that m is bounded, $f\{\tau\}$ is smooth, and $P\{m \mid M, \tau\}$ is continuous in τ.

[17] Hartigan [18] shows that for two-sided confidence intervals, *assuming symmetric posteriors*, other priors also lead to $O(M^{-1})$ asymptotic accuracy. Since posteriors for the binomial are sometimes not symmetric, in particular for $m = 0$, his result cannot be applied generally for the binomial.

$$\begin{aligned}
a &= m + 0.5 & b &= M - m + 0.5 \\
\omega_1 &= \frac{z\sqrt{h+\lambda}}{h} & \omega_2 &= \left(\lambda + \frac{5}{6} + \frac{2}{3h}\right) t \\
t &= \frac{M - 2m}{2m(M-m)} & h &= \frac{4m(M-m)}{M} \\
\lambda &= \frac{z^2 - 3}{6} & z &= \Phi^{-1}(1 - \alpha/2) \qquad (z = 1.96, \text{ for } 95\% \text{ confidence})
\end{aligned}$$

We found this approximation to be very accurate for $5 < m < M-5$. At the extremes, $m \leq 5$ or $m \geq M-5$, we note the symmetry of the confidence limits, $\text{lcl}(M, m, \alpha) = 1 - \text{ucl}(M, M-m, \alpha)$, where $\text{lcl}(\cdot)$ is the lower and $\text{ucl}(\cdot)$ the upper (1–α) limit, and we have found the following empirical formulas for the 95% limits:

$$\text{lcl}(M, m) \;=\; \begin{cases} 0 & \text{if } m = 0 \\ \exp[\,k_0 + k_1(\ln m) - (1 + k_2 m)(\ln M)\,] & 0 < m \leq 5 \end{cases}$$

$$k_0 = -2.03763 \quad k_1 = 1.874143 \quad k_2 = 0.01545$$

$$\text{ucl}(M, m) \;=\; \begin{cases} \exp[\,b_0 - b_1(\ln M) - b_2(\ln M)^2 + b_3(\ln M)^3\,] & \text{if } m = 0 \\ \exp[\,A_m - B_m(\ln M) - C_m(\ln M)^2 + D_m(\ln M)^3\,] & 0 < m \leq 5 \end{cases}$$

$$\begin{aligned}
&b_0 = 0.132893 \quad b_1 = 0.539755 \quad b_2 = -0.093773 \quad b_3 = 0.00657 \\
&A_m = 0.237901 + 0.207148\,m - 0.05118\,m^2 + 0.002488\,m^3 \\
&B_m = 0.500124 - 0.148240\,m \\
&C_m = 0.099266 + 0.032337\,m \\
&D_m = 0.006670 + 0.002430\,m
\end{aligned}$$

We compared the combination of these empirical formulas (for $m \leq 5$ or $m \geq M$–5) and Abramowitz & Stegun's formula (for $5 < m < M-5$) to the precise calculation at $M = 10, 20 \ldots 200$ and $m = 0, 1 \ldots N/2$. The greatest absolute deviation found for the upper limit $\text{ucl}(\cdot)$ was 0.0046 at $(M = 10, m = 4)$, and the absolute deviation appeared to be $< 0.00035 + 0.07/M$. The greatest deviation found for the the lower limit $\text{lcl}(\cdot)$ was 0.0033 at $(M = 20, m = 5)$, and the absolute deviation appeared to be $< 0.00078 + 1/M^2$.

A.4 The Uniform Prior and Other Beta Distributions

The *uniform prior*, $f\{\tau\} = 1$, is a special case, $f\{\tau\} = \text{Be}(\tau, 1, 1)$, of the Beta priors, nominally expressing complete ignorance as to τ. A qualitative *a priori* argument against the uniform prior for classification problems takes note of the following facts:

- The true error rates being estimated are those of classifiers inferred from the sample. The inferred classifier always correctly predicts the class of some of the items in its training set. Those items are members of the population, therefore the error rate tested on the entire population cannot be 100%.

- Populations involving actual measurements and observations, as opposed to hypothetical populations, always involve measurement, observation, and recording errors, and frequently have missing or inconsistent data. The inherent error of these real populations is unlikely to be zero; even if it were zero, the inference method entails a language-intrinsic error which is usually non-zero[18]. Therefore, it is very unlikely that the true error will be zero.

- We know, *a priori*, that we can construct a classifier, the *majority inducer* (always predict whichever class has the largest frequency in the sample) which makes minimal use of the sample data and typically has an expected true error of less than 50%. Our *ex post*, more informed inference methods should certainly be able to do at least this well. Therefore, true errors larger than 50% are relatively unlikely.

Note that the expected error of the majority inducer depends crucially on the assumptions made concerning the prior distributions of the number of classes and of the frequency of the larger class. For a population with C classes and fraction p in the larger class, $1/C \leq p \leq 1$, the expected error is approximately[19] $1 - p$. If p is assumed to be uniformly distributed, then the overall expected error of the majority inducer is $0 \leq E(\epsilon) \leq 0.5(1 - 1/C) < 50\%$. However, if the problem domain tends to have approximately equally frequent classes, then error rates near $1 - 1/C$ will dominate and the overall expected error would be greater than 50%. Conversely, in domains where one of the classes tends to dominate, *e.g.,* quality control or medicine where the majority of the product is good and the majority of patients are healthy, the overall expected error would be very low.

Figure 8 summarizes some empirical data for the majority inducer gathered from 118 ‘real-world’ problems found in the UCI databank [38], a textbook on categorical data analysis [2], and a textbook on multivariate statistical analysis [37]. These data do not necessarily represent the distribution of problems likely to be found in nature, but are fairly representative of a broad range of datasets likely to be used in evaluating classifier induction methods. In Figure 8a we show a histogram for the apparent error of the majority inducer on these problems and a best fitting (least squares) Beta distribution approximation, Be(0.74,1.44). In Figure 8b we show the empirical distributions for the number of classes, which is roughly an exponential distribution, $\exp(-0.83C)$, and for the average error rate for a given number of classes, which is fairly described by $0.86(1 - 1.39/C)$. The overall average error for the data in Figure 8 is 39%.

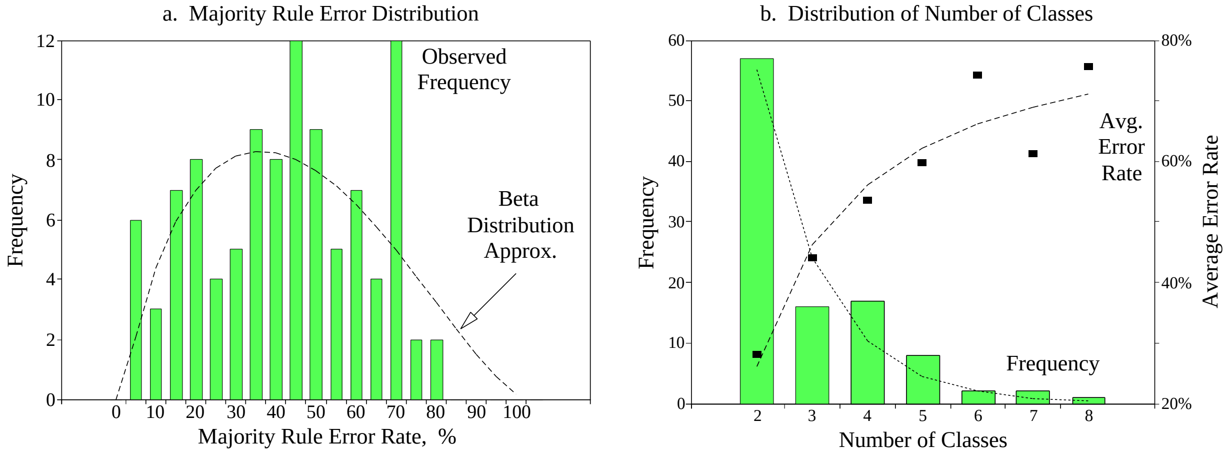
A non-uniform Beta distribution is consistent with these qualitative observations concerning the prior for τ. Figure 9 shows some data on the relative frequency of various error rates compiled from a study [29] of 16 data sets from several problem areas, from the survey of error rates by Holte [20], and from the study of decision tree pruning for small samples by Weiss & Indurkhya [54]. Altogether, these data cover 35 different datasets, several different induction algorithms, and several variants for some of the algorithms. While we do not purport that these data are representative of the distribution of problems in the ‘real world’, they are fairly representative of the datasets frequently used to evaluate classification algorithms.

As shown in Figure 9, these empirical data are consistent with a Beta distribution, Be$(1, 3.67)$, which is also consistent with our qualitative observations. For this prior, $E(\tau) = 21\%$, $\sigma_\tau = 16\%$, the median is 16%, the mode is zero, less than 8% of the error rates exceed 50%, and only about

[18] The inherent error is the hypothetical lowest possible error for any deterministic classifier due to the inherent ambiguity of the data, also known [10] as the Bayes optimal error rate. If the correct classifier is a quadratic discriminant, then the linear discriminant which has the lowest possible error can only approximate the correct classifier. Its error, though minimal for this kind of classifier, is greater than the inherent error. We term this hypothetical minimum error for the chosen inference method the *language-intrinsic error*, denoting its dependence on the language used to represent a classifier.

[19] It is exactly $1 - p$ when $p = 1$ or $p = 1/C$, and slightly greater than $1 - p$ between these limits (because the most frequent class in the sample is occasionally not the most frequent class in the population). The expected error converges to $1 - p$ quickly as either N or p increases.

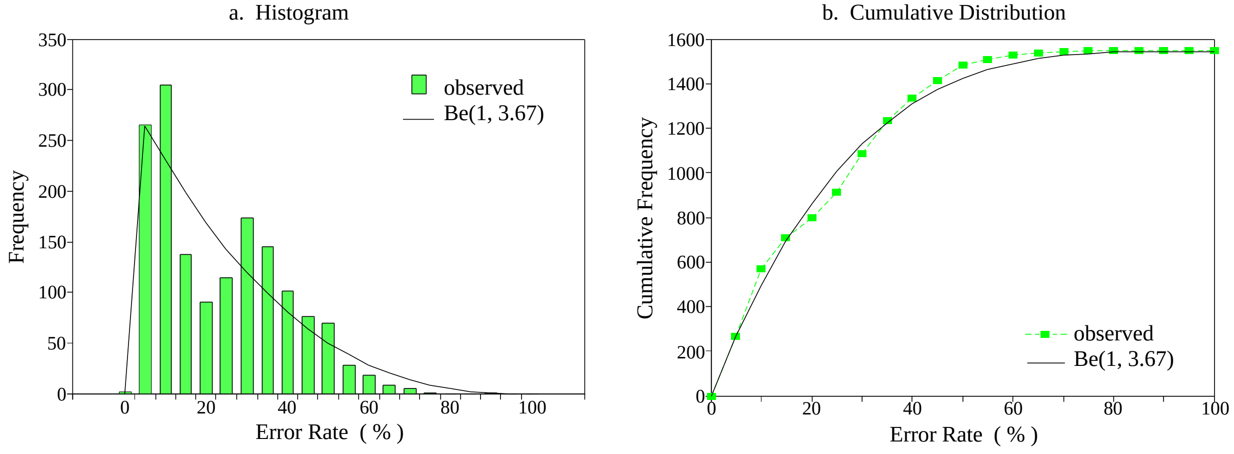
Figure 8: Majority Inducer Error Rate Distribution



a. Majority Rule Error Distribution

b. Distribution of Number of Classes

Figure 9: Empirical Error Rate Distribution



a. Histogram

b. Cumulative Distribution

Table 10: Paired t-test of TER's for 3-NN *vs.* 1-NN

values of t, two-sided test with 19 dof

$|t| > 2.093$ is significant at the 95% level

$t < 0$ indicates 3-NN is the more accurate classifier

Sample	Population Inherent Error Rate						
Size	0.1%	1%	2%	5%	10%	25%	40%
10	.9	2.3	.5	– .5	1.0	1.0	.1
20	2.0	–1.2	– .2	–1.0	–1.7	–2.9	–3.0
30	3.0	.2	– .4	–3.6	–3.7	–6.5	– .8
50	3.1	–1.7	–2.1	–3.0	–2.9	–5.4	–6.2
100	3.2	–1.0	–3.2	–4.4	–9.0	–7.3	–3.5

one-third exceed 25%. Also, the confidence intervals are lower and narrower using this prior than those obtained using either the uniform or Jeffreys' priors. We caution that the variety and range of the problems summarized in Figure 9 are much too sparse to accept this prior without reservation.

B Empirical Study of Significance Tests

In this section we present results of empirical studies of significance for classifier error differences. An extended example is given in Section B.1, studying differences between nearest neighbor and three nearest neighbors classifiers. Section B.2 summarizes a smaller experiment performed to verify that our conclusions based on the nearest neighbors study are also applicable in the case of pruned and unpruned decision trees for nominal attributes. Section B.3 summarizes experiments investigating repeated cross-validation of a single sample.

B.1 A Nearest-Neighbors Example

In this and following sections, TER denotes a classifier's true error (the rate which would be observed were the classifier tested on the entire population); 1-NN or subscript 1 denotes a nearest neighbor classifier and 3-NN or subscript 3 denotes a 3 nearest neighbors classifier.

Are the error rates of 3-NN really different from those of 1-NN and, if so, which method yields the more accurate classifiers? Under the null hypothesis, the statistic $t = \overline{x}/s(\overline{x})$ is distributed approximately as Student's t, where $\overline{x} = \overline{\text{TER}_3 - \text{TER}_1}$ is the average difference and $s(\overline{x})$ is the estimated standard deviation of $\overline{x}$.

In Table 10 we summarize t-test results of a paired test simulating 20 samples each of several different sizes from populations having different inherent errors (two equally likely classes, each normally distributed on a single attribute, with the same variance but different class means, see [32]). As a rule, the 3-NN classifiers are more accurate. However, this is not the case when the inherent error is very low (0.1%) or the sample size very small ($N = 10$). Examination of the data in Table 10 also suggests that there is no difference at all in the error rates of 1-NN and 3-NN for these populations for a sample size of about 12 or an inherent error near 0.5%.

One explanation for these observations takes into account the data density around the critical region where the classes overlap. The inherent error is the relative density (fraction of the population) in this region and the sample size reflects the overall density. The product of the sample size and inherent error is the expected number of items in this region (the number in any particular sample is random, with a binomial distribution). When this expected number is low, especially when it is less than one, a sample will contain little information from which to infer the placement of the class boundaries. The smoothing effect of 3-NN heavily discounts the apparent information imparted by relatively isolated instances and, in very sparse data, there is little information to spare — while 1-NN overfits the sample, 3-NN underfits when the data are very sparse. These observations on the effects of smoothing in nearest neighbors classifiers are consistent with Schaffer's [45] observations that pruning (smoothing) decision trees may actually be harmful when the data are very sparse relative to the concept to be learned.

A sample size greater than 30 seems necessary for consistent results regarding whether there is a statistically significant advantage for 3-NN over 1-NN (or *vice-versa*) measured over 20 samples. The differences are neither more nor less real for larger or smaller samples, but there is so much variation in the results for smaller samples that we have but little confidence in our measurement of the differences — we would need to average over a larger number of samples (*i.e.,* classifiers) in order to have the same confidence for smaller samples. While 1-NN classifiers appear to be more accurate than 3-NN for very small samples (10 or less), we cannot confidently reject the hypothesis that 1-NN and 3-NN are equally accurate for small samples.

The results in Table 10 would belie any assertion that 3-NN is universally superior to 1-NN. A decision to use 3-NN rather than 1-NN reflects an inferential bias, an *a priori* assumption that 1-NN will overfit the data (and that 3-NN will not underfit). Whether our decision will result in a more accurate classifier depends on how appropriate this bias is to the problem at hand. If the sample size is small or the inherent error very low, 1-NN tends to overfit but 3-NN has a stronger tendency to underfit. Since we have shown [32] that the 1-NN classifier in this case is entirely equivalent to an unpruned CART-style decision tree and that decision tree pruning and nearest neighbors smoothing have similar effects, we expect that these observations are equally applicable to decision tree pruning and other such questions as to differences between inference methods. The choice of one inference algorithm over another is simply a choice (albeit many times a tacit or unawares choice) of one set of assumptions about the data over another set of assumptions. Probably the most crucial step in any statistical inference is matching assumptions to the problem.

In most real-world situations the biases underlying nearest neighbors smoothing and decision tree pruning (namely that the language-intrinsic error is significantly greater than zero, that the data contain mistakes and measurement errors in addition to sampling variation, and that inference methods that are not smoothed or pruned will overfit) are almost certainly more appropriate than the naive counter-assumptions that classes do not overlap and reported data may be relied on as gospel. However, we should be aware that we are relying on these assumptions, and this may have unanticipated consequences, as in the interaction of sample size and inherent error in Table 10.

How important are the nuances of calculating $s(\overline{x})$? Statistical packages and recipe books typically either give only one method for a significance test on means or they give a large variety of methods that may bewilder novice users. For illustrative purposes, Table 11 shows the results of mis-applying the formula for unpaired observations to analyze the data from our paired experiments. All of the t values in Table 11 are lower than their correct counterparts in Table 10, enough so that the wrong conclusion is reached as to the significance of 3-NN *versus* 1-NN in 7 cases out of 35. This is so because we have overestimated $s(\overline{x})$. The assumptions underlying the unpaired t-test formula are inappropriate for the paired experiment at hand, and such errors will likely result if the method of

Table 11: Unpaired Calculations Mis-Applied to Paired Observations

values of t, two-sided test with 19 dof

$|t| > 2.093$ is significant at the 95% level

⊘ indicates that the difference is significant, but does not appear to be so here

Sample	Population Inherent Error Rate						
Size	0.1%	1%	2%	5%	10%	25%	40%
10	.7	⊘ 1.7	.5	− .4	.8	.6	.0
20	1.7	− .9	− .1	− .4	−1.3	⊘ −1.9	⊘ −1.1
30	2.6	.2	− .3	−2.1	−2.3	−5.3	− .7
50	⊘ 1.4	−1.5	⊘ −1.2	−2.8	⊘ −1.4	−3.1	−3.1
100	⊘ 1.8	− .8	−2.7	−3.8	−5.6	−4.0	−2.5

data analysis is not properly matched to the experimental conditions.

The analysis in Table 10 is possible only because we have perfect knowledge of the populations and TER's. In general, this is not the case, and we have to base our analysis on one or another method for estimating error. In Table 12 we show results of the paired t-tests in Table 10 using several estimators (the sets of samples are identical, only the method of estimating error changes; LOO denotes leave-one-out, 10-CV 10-fold cross-validation, 632b Efron's [13] 632 bootstrap, and LOO* Weiss' [50] hybrid estimate — see [32] for a review of these methods).

In 13 of our 40 experiments, the difference between the LOO estimates of 1-NN and 3-NN error rates is not significant at the 95% level, though the difference in TER's is significant. The LOO estimates, though unbiased, are more variable than the TER's, and our t-test consequently less sensitive. The 10-CV results are similar, except that 10-CV is even more variable than LOO, and there are even more cases (15/40) where we cannot reject the null hypothesis at the 95% level, even though the TER's are significantly different.

LOO is more variable than TER because LOO averages the errors of N classifiers, each slightly different from the reference classifier inferred using all of the sample, being inferred from one less instance. Let δ_i be the difference between TER and LOO for the i^{th} sample, $\text{LOO}_i = \text{TER}_i + \delta_i$. LOO's variance is greater than the variance of TER by the mean square of the δ_i. Similar considerations apply to 10-CV, except that the mean square δ is even larger than for LOO, because we average over fewer subsamples, each differing even more from the reference than for LOO (since 10% of the instances are omitted rather than only one), and because of randomness in selecting subsets.

These considerations also apply to all of the resampling estimators, with the important difference that bootstrapping and iterated cross-validation average over a very large number of subsamples, which tends to reduce the variance to a level below that of LOO. For the biased estimators in this family, the difference in the variances of the estimator and TER is no longer simply the mean square δ, and the estimator variance may even be lower than the variance of TER. For instance, the apparent error (the rate observed when a classifier is tested on the same instances used to infer the classifier), which has zero variance for 1-NN.

Note the anomalous results for the 632b estimator. These high t-values and significance levels are not incorrect, but they are a potential pitfall for an unwary user. The 632b error rates of the 1-NN and 3-NN classifiers are different, with a very high degree of confidence, and the 632b rates for 1-NN are always better than those of 3-NN. This is because 632b is biased in both cases and

Table 12: Paired t-test of Estimate Means for 3-NN *vs.* 1-NN

⊘ indicates that the TER's are significantly different, but the estimates do not appear to be
⊙ indicates that the TER's are not significantly different, but the estimates appear to be
‡ indicates that the sign is opposite to the sign of the mean difference in TER's

Sample	Population Inherent Error Rate							
Size	0.1%	1%	2%	5%	10%	25%	40%	50%
	LOO Estimates							
10		⊘		‡			‡	
20	‡		‡	‡		⊘	⊘	
30	⊘			⊘	⊘	⊘		
50	⊘			⊘	⊘		⊘	
100	⊘‡	‡	⊘					
	10-CV Estimates							
10		⊘		‡			‡	
20	‡		‡	‡		⊘	⊘	
30	⊘			⊘	⊘	⊘		⊙
50	⊘		⊘	⊘	⊘		⊘	
100	⊘	‡	⊘			⊘		
	632b Estimates							
10	⊙		⊙	⊙‡	⊙	⊙	⊙	⊙
20		⊙‡	⊙‡	⊙‡	⊙‡	‡	‡	⊙
30		⊙	⊙‡	‡	‡	‡	⊙‡	⊙
50	⊘	⊙‡	‡	‡	‡	‡	‡	⊙
100	⊘	⊙‡	‡	‡	‡	‡	‡	⊙
	LOO* Estimates							
10	⊙		⊙	⊙‡			‡	
20		‡	‡	‡		⊘	⊘	
30			‡	⊘	⊘	⊘		
50	⊘		⊘	⊘	⊘		⊘	
100	⊘‡	‡	⊘					

has a different bias for 1-NN than for 3-NN. Knowing that the 632b estimates differ significantly tells us nothing about whether the TER's are different. This is the great danger inherent in using biased estimators — unless the biases are known to be the same for the cases under study, or unless the biases are known and compensated for in each case, use of these biased estimators can (and almost certainly will) lead to fallacies in inferences about the differences between cases. This is so regardless of any apparent advantage for these estimators in terms of reduced variance.

The LOO* estimator is approximately unbiased for 1-NN and 3-NN, and its t-test behavior is similar to that of LOO or 10-CV, except that LOO* indicates that the advantage of 1-NN over 3-NN for very small samples is significant when the TER, LOO and 10-CV differences are not significant. Based on the results in Table 12, we do not see any advantage for LOO* over LOO or 10-CV.

Table 13: t-test for Stopped *versus* Unpruned Trees

Inherent	Sample	t-statistic					
Error %	Size	TER	APP	LOO	10-CV	632b	LOO*
0.1	24	4.2	4.4	⊘ 1.9	2.4	3.1	2.9
	36	2.3	6.0	4.6	.5	3.9	3.9
	48	3.9	11.2	3.5	6.1	6.1	4.7
	96	1.0	1.5	⊙ 3.0	⊙ 3.1	⊙ 5.1	⊙ 4.8
5	24	2.9	3.9	2.6	2.7	2.9	3.3
	36	1.1	⊙ 7.3	⊙ 3.3	2.0	⊙ 3.1	⊙ 3.3
	48	.9	⊙ 2.8	2.2	1.6	2.1	⊙ 2.8
	96	3.0	2.7	⊙ 2.0	⊘ 1.8	2.4	3.1
10	24	2.2	1.2	.1	.5	.4	.2
	36	1.2	⊙ 2.9	1.6	1.7	⊙ 2.5	1.6
	48	.2	1.8	− .2	.2	.9	.2
	96	1.2	2.2	2.0	1.6	2.0	1.7
25	24	− .3	⊙ 5.3	1.0	.5	1.8	1.0
	36	.2	⊙ 2.9	1.0	.1	2.0	1.0
	48	.6	⊙ 5.9	1.6	1.0	⊙ 7.3	2.3
	96	.1	2.2	.8	1.0	1.2	.9

⊙ TER difference is not significant, but this is
⊘ TER difference is significant, but this is not

B.2 A Similar Decision Tree Experiment

An experiment was conducted to verify that our nearest neighbors results also apply when pruning decision trees derived from nominal rather than continuous attributes. Eight samples each of sizes 24, 36, 48, and 96 were drawn from noisy contact lens populations [32] with inherent error rates of 0.1, 5, 10, and 25%. Both an unpruned and a stopped[20] decision tree were inferred from each sample, and the various error estimates were determined for each of the trees.

In Table 13 we show t-statistics for the paired differences between stopped and unpruned trees, highlighting cases where the test using one of the estimators is misleading as to the significance of the difference in true error. Overall, the stopped trees are less accurate, but the difference is not significant for higher inherent error or larger samples. For each entry in Table 13, there are 7 degrees of freedom, and $|t| > 2.365$ is significant at the 95% level.

The data in Table 13 suggest that 632b and LOO* tend to exaggerate the significance of the difference between stopped and unpruned trees, and that the Type I risk (falsely rejecting the null hypothesis) when using these estimators is markedly higher than 0.05 (the value implied by a nominal 95% test). LOO and 10-CV seem to lead to correct decisions regarding the null hypothesis for inherent error rates of 10% or more, but may sometimes overstate or understate significance for lower error rates. On the whole, 632b and LOO* do not appear to have any advantage over LOO or 10-CV for these data. 10-CV is recommended for these comparisons because it has the lowest computational cost among these four methods and also appears to have the least added Type I or Type II risk. Breiman & Spector [11] have found 10-CV to be more effective than LOO for pruning.

[20] Split and stopped using the hypergeometric probability test [29] (an extension of Fisher's exact test [2]).

The t-tests in Table 13 relate to the average difference over 8 samples. As was the case for nearest neighbors (see Figure 7), the difference for a single sample is highly variable, and not trustworthy even as to the sign of the difference. The SE = $\sqrt{\epsilon(1-\epsilon)/N}$ heuristic for the standard deviation of the paired differences between stopped and unpruned trees is biased. Though similar in shape to the corresponding relationship for 3-NN *vs.* 1-NN (see Figure 6a), the bias is quantitatively different for these discrete attribute trees than for those continuous attribute classifiers.

B.3 Iterating Paired Cross-Validation

In this section we discuss experiments conducted to determine whether iterating the paired cross-validation of a single sample 100 times would markedly improve the reliability of single-sample t tests over a single paired cross-validation.

It is known (see the companion paper [32]) that iterating k-CV can significantly reduce the variance of the estimate when k is small, but the estimates themselves are increasingly biased as k decreases. We note that iterating k-CV does not affect its bias, and has little effect on the variance for $k \geq 10$ (no effect at all for leave-one-out).

This reduced variance, though desirable from the point of view of possibly increasing the power of a t-test, presents special difficulties for the analysis. Firstly, since the estimates are biased, the comparison can be very misleading unless the biases are the same for both classifiers being compared. Secondly, though the variance of the average error over m iterations is lower than the variance of a single iteration, it is not reduced by a factor of $1/m$, as would be the case for m independent observations. The reduction in variance appears to be a function of both the dataset and the inference method, and we know of no analytical expression for predicting the reduction in any particular case. And so, though we can certainly iterate a paired k-fold cross-validation m times and calculate the average, $\overline{x}$, and standard deviation, $S(x)$, of the mk error estimate differences so obtained, we cannot, from a single sample, predict the expected standard deviation of the average, $S(\overline{x})$, except to note that $S(x)/\sqrt{m} \leq S(\overline{x}) \leq S(x)$.

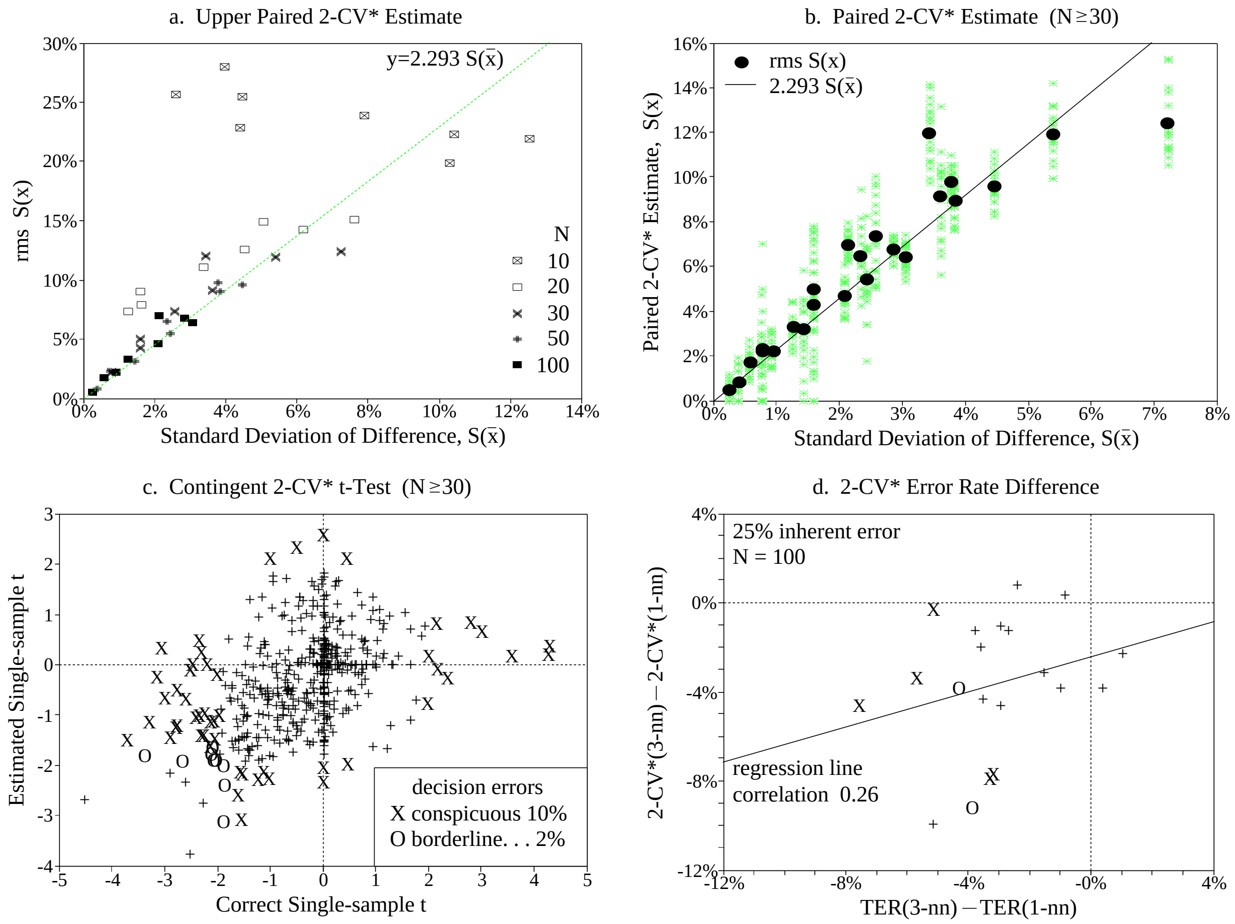
Using the lower bound on this variance, $S(x)/\sqrt{m}$, would grossly exaggerate the significance of the difference $\overline{x}$. Using the upper bound, $S(x)$, would likewise understate the significance. This is a tempting trap, since one could manipulate the value of m to 'prove' significance or non-significance, as desired, using these incorrect expressions for the variance.

Comparing 3-NN and 1-NN for 40 different cases, as in Section B.1, we simulated 100 paired 2-fold cross-validations (2-CV*) for each of 20 samples in each case. Figure 14a shows the rms value of $S(x)$ over the 20 samples *versus* $S(\overline{x})$ for each of the 40 cases. For $N \geq 30$, rms $S(x)$ is roughly proportional to $S(\overline{x})$. We note again that we know of no method for predicting the coefficient (2.293 in this particular experiment) and that it is far from $\sqrt{100}$. For $N \leq 20$, the relationship has a lower slope and a non-zero intercept, and the slope is actually negative for $N = 10$.

The anomalous results for $N \leq 20$ are in part due to the facts that the 2-CV* estimator is biased, more so as the sample size decreases, and that the bias is different for 3-NN than for 1-NN, as shown in Table 15. The bias of 3-NN at $N = 10$ is very large, especially when the inherent error is low. Judging from Figure 14a, the 3-NN rates for $N = 10$ are also highly variable. Weiss [50] also observed this large bias of 2-CV* for nearest neighbors using small samples, and that the bias is different for 3-NN than for 1-NN. In fact, these observations motivated his LOO* estimator.

Obviously, the 2-CV* estimator is not appropriate for comparing 1-NN and 3-NN for $N \leq 20$, and the remainder of our analysis is restricted to $N \geq 30$, where the bias is small for both classifiers

Figure 14: Iterated Paired 2-Fold Cross-Validation



a. Upper Paired 2-CV* Estimate

b. Paired 2-CV* Estimate (N≥30)

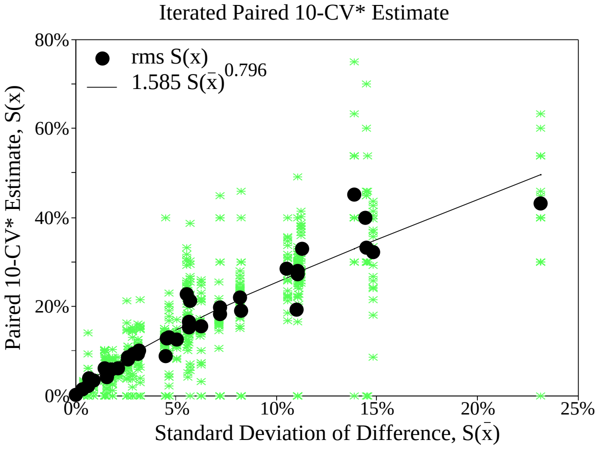
c. Contingent 2-CV* t-Test (N≥30)

d. 2-CV* Error Rate Difference

Table 15: Bias of Iterated Paired 2-Fold Cross-Validation

Sample	Population Inherent Error Rate							
Size	0.1%	1%	2%	5%	10%	25%	40%	50%
				1-NN Bias, %				
10	3.4	3.3	1.0	3.5	6.6	4.6	1.3	−3.5
20	.3	.2	.0	.9	− .9	1.5	1.1	3.0
30	− .1	.3	.1	.0	.2	3.9	.5	− .5
50	− .1	− .5	.5	.9	1.6	− .4	.7	2.4
100	.0	.2	− .2	− .2	1.0	2.3	1.6	.9
				3-NN Bias, %				
10	18.7	17.8	17.0	17.6	11.9	7.4	1.8	−3.6
20	1.6	2.2	3.1	4.8	2.0	4.6	3.1	2.9
30	− .0	.1	.1	.7	.9	2.7	.6	.0
50	.0	.2	− .1	.4	1.3	.6	1.2	.9
100	.1	− .0	− .1	.3	1.4	1.8	1.3	.0

Figure 16: Iterated Paired 10-Fold Cross-Validation



and, thus, the biases very nearly cancel. Figure 14b shows a scatter plot, $S(x)$ *vs.* $S(\overline{x})$, of these data. Compared to the equivalent plot for uniterated 10-CV (Figure 6b in the body of the paper), the relationship here is simpler (proportionality, rather than a power function) and less variable than in the uniterated case.

Assuming that we somehow knew that the $S(x)$ estimates should be reduced by a factor of 2.293, how accurate are the significance decisions that would be made using this $t = 2.293\overline{x}/S(x)$ test? In Figure 14c, we show this estimated t *versus* the t value using the classifier's true error rates (TER), $t = \widehat{x}/\sigma(\widehat{x})$ where $\widehat{x} = \text{TER}_3 - \text{TER}_1$ and $\sigma(\widehat{x})$ is the standard deviation of $\widehat{x}$ for the 20 samples. The correct significance decision would be made in 88% of the cases, but the majority of the wrong decisions are conspicuous (*i.e.,* they assert that the classifier's true error rates differ at the 95% confidence level or greater when they are not different at even the 90% level, or *vice-versa*).

A significant fraction of the data in Figure 14c fall in the upper left and lower right quadrants, indicating that the 2-CV* estimate of the difference between 3-NN and 1-NN has the opposite sign from the difference in the true errors. This is a problem independent of the method for estimating the standard deviation: the estimated differences are highly variable, poorly correlated with the true difference, and not trustworthy even as to the sign of the difference. We illustrate this in Figure 14d for the population (25% inherent error) and sample size ($N = 100$) where the true difference averaged over 20 samples was most highly significant ($t = -6.85$ with 19 degrees of freedom, which is significant at at least the 99.9% level). The sign of the estimated difference is reversed for 4 of the 20 samples, and the correlation between the estimated difference and the true difference is very poor. The single-sample decision accuracy for the 20 samples here is only 65%.

Despite the manifest difficulties of this approach, iterating paired cross-validation for $N \geq 30$ resulted in a simple proportionality and reduced the scatter of $S(x)$ relative to that seen for uniterated 10-CV. Since 10-CV is relatively unbiased and iteration has comparatively little effect on its variance, we speculated that iterating paired 10-CV might possibly give a simple proportionality between $S(x)$ and $S(\overline{x})$, with a coefficient near unity, and also reduce the scatter. Figure 16 shows the results of this experiment (identical to that in Figure 14b except for the number of folds of the cross-validation, and compare Figure 6b in the body of the paper). Unfortunately, the experiment refuted the conjecture on both accounts.