Statistica Sinica **16**(2006), 635-657

COMPARING LEARNING METHODS FOR CLASSIFICATION

Yuhong Yang

University of Minnesota

Abstract: We address the consistency property of cross validation (CV) for classification. Sufficient conditions are obtained on the data splitting ratio to ensure that the better classifier between two candidates will be favored by CV with probability approaching 1. Interestingly, it turns out that for comparing two general learning methods, the ratio of the training sample size and the evaluation size does not have to approach 0 for consistency in selection, as is required for comparing parametric regression models (Shao (1993)). In fact, the ratio may be allowed to converge to infinity or any positive constant, depending on the situation. In addition, we also discuss confidence intervals and sequential instability in selection for comparing classifiers.

Key words and phrases: Classification, comparing learning methods, consistency in selection, cross validation paradox, sequential instability.

1. Introduction

The beginning of the information age has prompted new challenges to statistical learning. For example, in gene expression data analysis for medical diagnose of patients, one is faced with the difficulty of building a classifier with thousands or more variables as input but only a small number of training cases. With a high input dimension and/or a relatively small sample size, statistical behavior of a learning method becomes complicated and often difficult to characterize.

Obviously, understanding the relative performance of the various choices of learning methods is important. Besides theoretical investigations of their risk properties (such as rate of convergence), numerical comparisons have been reported in the literature. For an empirical comparison of classifiers, the observations are often split into a training set and a test (or evaluation) set, and the process is usually replicated in a certain way to reduce variability in data splitting. Then usually the candidate with the lowest test error rate is favored. How reliable is such a comparison? Is it consistent in selection in the sense that the better or best classifier will be selected with probability going to 1? How does the data splitting ratio influence the consistency property?

In this work, we intend to address these and some related questions on comparing classifiers. We set up the framework as follows.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. observations having the same distribution as (X, Y), where X is the explanatory variable taking values in $\mathcal{X}$ and Y is the response variable taking one of two values in $\{0, 1\}$. Let $f(x)$ be the conditional probability function: $f(x) = P(Y = 1|X = x)$. For convenience, let Z^n denote $(X_i, Y_i)_{i=1}^n$.

A classifier $\delta(x) = \delta(x; Z^n)$ is a rule to declare membership status of Y based on the observed data Z^n and the new X value. Mathematically speaking, δ is a measurable mapping from $\mathcal{X} \times [\mathcal{X} \times \{0, 1\}]^n$ to $\{0, 1\}$. We are interested in comparing the performances of different classifiers.

As is well known, the Bayes rule $\delta^*(x) = I_{\{f(x) \geq 1/2\}}$ minimizes the error probability $P(\delta(X) \neq Y)$ over all δ mapping from $\mathcal{X}$ to $\{0, 1\}$. Since f is unknown, the Bayes rule is not a formal classifier and one has to rely on the data to estimate δ^* one way or the other. Let $PE(\delta; n) = P(\delta(X; Z^n) \neq Y)$ be the probability of error of a classifier δ at sample size n. Since the Bayes rule minimizes the error probability at each x, it is proper to assess the performance of a classifier δ relative to it. Thus we consider the probability error regret $PER(\delta; n) = PE(\delta; n) - PE^*$, where PE^* denotes the error probability of the Bayes rule.

In this work, we are mainly interested in the situation where the candidate classifiers are general (parametric or nonparametric) and are not necessarily closely related to each (as in e.g., the situation of the classifiers being based on nested parametric models, or from empirical risk minimization over classes of sets with increasing VC dimensions).

Of course, the topic of comparing classifiers is not new. In the literature, results on classifier selection and classification error rate estimation have been obtained. In the theoretical direction, Devroye (1988) derived error probability bounds for classifier selection based on data splitting, and obtained interesting consistency (in terms of error probability convergence) and an asymptotic optimality property. Various methods have been proposed for error rate estimation for a classifier, including parametric methods, bootstrap, cross validation and jackknife. See Efron (1983, 1986), McLachlan (1992, Chap. 10) and Devroye, Györfi and Lugosi (1996, Chap. 8) for results and references. More recently, Efron and Tibshirani (1997) proposed the .632+ bootstrap method to improve on cross validation for error rate estimation. Other results related to classifier comparison using cross validation or related methods include, e.g., Kohavi (1995) and Dietterich (1998).

When classifiers are compared, consistency in selection is a desirable property and one which, to our knowledge, has not been obtained in generality. Although good estimates of error rates of individual classifiers can be used to judge whether

two classifiers are different in accuracy, this may be a sub-optimal practice for comparing classifiers.

The organization of the rest of the paper is as follows. In Section 2, we investigate the consistency (in selection) property of cross validation methods and give sufficient conditions on the data splitting ratio. In Section 3, we consider confidence intervals for the difference of the conditional error probabilities. Section 4 addresses the issue of sequential instability in selection. The proofs of the theoretical results are given in Section 5.

2. Consistency in Selection

Popular classifiers include parametric and nonparametric methods, such as LDA by Fisher, logistic regression, classification trees, nearest neighbor, support vector machines, neural networks, empirical risk minimization, as well as plug-in methods based on estimation of the probability function $f(x)$ (see Devroye, Györfi and Lugosi (1996) and Hastie, Tibshirani and Friedman (2001)). These methods perform well under different conditions and they may have different rates of convergence in terms of PER. Under strong assumptions on the behavior of f when it takes values close to 1/2, rates of convergence faster than $1/\sqrt{n}$ are possible. For example, Tsybakov (2004) showed that the minimax rate of PER, under a condition on f and a metric entropy assumption on the class containing f, is $n^{-(1+\gamma)/[2(1+\gamma)+\rho\gamma-\gamma]}$, where ρ is an index of the order of the metric entropy and $\gamma > 0$ is a margin parameter. Note that, for $0 < \rho < 1$, the rate is always faster than $n^{-1/2}$ (but slower than $1/n$).

With the many methods available, one naturally wants to find the best classifier for the current data. Assuming that one classifier is asymptotically the best, how can we identify it with probability approaching one?

We focus on the data splitting approach. It is well-known that for this approach, due to data splitting, the training size is necessarily smaller than the actual sample size, and thus the comparison of the classifiers is in fact comparing the classifiers at a reduced sample size, which may introduce a bias. However, this approach is still valuable for several reasons. One is that it is possible to assess whether the comparison of the classifiers at a reduced sample size is reasonably close to the targeted sample size via a certain sequential instability assessment (see Section 4). Another is that the data splitting approach is basically distribution-assumption free (except that the observations are i.i.d.) and thus is more reliable when there is no compelling evidence to justify parametric models. Criteria for comparing learning methods based on the whole sample without data splitting typically are derived with heavy use of distributional information. For example, in AIC (Akaike (1973)) or BIC (Schwarz (1978)), one needs the likelihood function, which is not available for non-model-based classifiers (e.g.,

nearest neighbor rules). In addition, one can readily have performance bounds for this approach.

Consider two candidate classifiers δ_1 and δ_2. We split the data into two parts with n_1 and n_2 observations respectively: $Z_1 = (X_i, Y_i)_{i=1}^{n_1}$ and $Z_2 = (X_i, Y_i)_{i=n_1+1}^{n}$. Apply δ_1 and δ_2 on Z_1 to get $\delta_1(x; Z_1)$ and $\delta_2(x; Z_1)$, respectively. Let $\widehat{Y}_{1,i}$ and $\widehat{Y}_{2,i}$ $(n_{1+1} \le i \le n)$ be the predictions by δ_1 and δ_2 at $X_{n_1+1}, \ldots, X_n$, respectively.

Let $TE(\delta_j) = \sum_{i=n_1+1}^{n} I_{\{Y_i \neq \widehat{Y}_{j,i}\}}$ be the test error for each classifier δ_j. We select the one that minimizes $TE(\delta_j), j = 1, 2$. When there is a tie, any tie-breaking method can be used.

The issue of consistency in selection for comparing general procedures was considered in regression in Yang (2005b). It turns out that for classification, there are two drastic differences due to aspects of classification that are not present in usual regression with a continuous response. One is that the requirement on the splitting ratio can be much more stringent compared to the regression case to handle fast rates of convergence of PER mentioned earlier, and the other is that the disagreement rate of the two competing classifiers (not just the error rates) plays a role.

2.1. When does consistency hold?

For defining consistency in selection, the candidate classifiers need to be orderable in accuracy. For a classifier δ, let $CPER(\delta; n) = P(\delta(X; Z^n) \neq Y | Z^n) - PE^*$ be the conditional probability error regret. Obviously, $PER(\delta; n) = E(CPER(\delta; n))$.

Definition 1. δ_1 is said to be asymptotically better than δ_2 if for every $0 < \epsilon < 1$, there exists a constant $c_\epsilon > 0$ such that when n is large enough, we have $P(CPER(\delta_2; n)/CPER(\delta_1; n) \ge 1 + c_\epsilon) \ge 1 - \epsilon$.

The definition basically says that the loss of δ_2 (i.e., $CPER(\delta_2; n)$) is larger with high probability, possibly by a tiny bit, than that of δ_1. The loss of the asymptotically better one does not have to converge at a faster rate than that of the other classifier. For a toy example, suppose that the true probability function is $f(x) \equiv 1/3$, that δ_1 randomly assigns label 0 with probability $1 - 1/n$, and δ_2 randomly assigns label 0 with probability $1 - 2/n$. Then δ_1 is asymptotically better than δ_2 by definition, but their convergence rates are the same. Let r_n be a sequence of non-increasing positive numbers.

Definition 2. δ is said to converge exactly at rate r_n in probability if $CPER(\delta; n) = O_p(r_n)$ and for every $0 < \epsilon < 1$, there exists a constant $d_\epsilon > 0$ such that when n is large enough, we have $P(CPER(\delta; n) \ge d_\epsilon r_n) \ge 1 - \epsilon$.

The word "exact" in the definition emphasizes that CPER does not converge faster on a set with probability bounded away from zero.

Assume that $CPER(\delta_1; n_1)$, $CPER(\delta_2; n_1)$, and $P(\widehat{Y}_{2,n_1+1} \neq \widehat{Y}_{1,n_1+1}|Z_1)$ converge exactly at rates p_{n_1}, q_{n_1} and s_{n_1} respectively. We allow s_n to not converge to zero.

Theorem 1. *Under the condition that one of δ_1 and δ_2 is asymptotically better than the other, we have that the classifier selection rule that minimizes $TE(\delta_j)$ is consistent as long as $n_1 \to \infty$ and $n_2 \max(p_{n_1}^2, q_{n_1}^2)/s_{n_1} \to \infty$.*

Remarks.

1. In the context of regression, Yang (2005b) obtained a similar result for comparing regression estimators. There are substantial differences. One is that the quantity s_{n_1} did not appear in the regression case. This is an important term, because it indicates the potentially large difference between the uncertainty in estimating error rates of the individual classifiers and the uncertainty in estimating the difference of the error rates (see Section 3). Another is that for regression, the rate of convergence is usually not faster than $n^{-1/2}$ and thus the most stringent requirement on the largeness of n_2 is that $n_2/n_1 \to \infty$. For classification, however, the error rate can be between $n^{-1/2}$ and n^{-1}, as shown in Mammen and Tsybakov (1999) and Shen et al. (2003). As a result, we may need to choose n_2 so that $n_2/n_1^2 \to \infty$.
2. The property of consistency in selection is different from achieving the best possible performance in classification accuracy. The core issue is that, due to uncertainty in selecting the best classifier, the risk of the selected classifier from a consistent selection rule is not necessarily the best in rate. Yang (2005a) showed that the goals of consistency in selection and optimal regression estimation (in a minimax sense) cannot be achieved simultaneously in a regression context. See Section 2.5 for more discussions.
3. When there are k candidate classifiers ($k \geq 3$), assuming there is one that is asymptotically best, a sufficient condition for consistency in selection is that $n_1 \to \infty$ and that $n_2 \max(p_{n_1}^2, q_{n_1}^2)/s_{n_1} \to \infty$ holds for comparing the best candidate with each of the other candidate classifiers.

When the input dimension is high, the classification problem usually becomes more difficult due to the curse of dimensionality, and the rate of convergence can be very slow. Suppose that $\max(p_n, q_n)$ is of order $n^{-\beta}$ for some $0 < \beta < 1$, and suppose that s_n is of order $n^{-\eta}$ for some $0 \leq \eta \leq \beta$. Then the requirement on data splitting ratio is $n_2/n_1^{2\beta-\eta} \to \infty$ (and of course $n_1 \to \infty$). Clearly when $2\beta - \eta < 1$, it suffices to have $n_1 = O(n_2)$ (e.g., half-half splitting). Actually, when the rates of convergence of δ_1 and δ_2 are very different, it may even be enough to distinguish the classifiers with $n_2 = o(n_1)$. This is in a dramatic contrast

with Shao's result (1993) for linear regression, where the requirement is always $n_2/n_1 \to \infty$.

When it is hard to assess $P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)$, an obvious upper bound is 1.

Corollary 1. *Under the same conditions as in Theorem* 1*, if p_n and q_n go to zero no faster than $1/n$, then a sufficient condition for ensuring consistency in selection is $n_2/n_1^2 \to \infty$.*

The stringent requirement on the splitting ratio in Corollary 1 can be substantially weakened when s_{n_1} converges to zero. It is even possible that s_{n_1} and $\max(p_{n_1}, q_{n_1})$ are of the same order, for which case the sufficient condition on the splitting ratio in Theorem 1 becomes $n_2 \max(p_{n_1}, q_{n_1}) \to \infty$. Under the conditions in the theorem, the requirement of $n_2 \max(p_{n_1}^2, q_{n_1}^2)/s_{n_1} \to \infty$ is equivalent to $n_2 \max(p_{n_1}, q_{n_1}) R_n \to \infty$ in probability, where

$$R_n = \frac{\left| P(\widehat{Y}_{2,n_1+1} \neq Y_i|Z_1) - P(\widehat{Y}_{1,n_1+1} \neq Y_i|Z_1) \right|}{P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)}.$$

Note that R_n is always between 0 and 1. We call it the essential error probability difference. For classification, it is possible that two learning methods both have very small PER, yet they disagree with each other often. For an extreme example, take $f(x) = 1/2$, and δ_1 and δ_2 are just independent random assignments of the labels with equal probability. Then both have PER equal zero, yet they disagree with each other with probability $1/2$, and $R_n = 0$. From the theorem and above, R_n plays an important role in the requirement of data splitting ratio for consistency.

Note that in the supervised learning literature, when data splitting is used to compare procedures empirically, a popular guideline is to have $1/4$ or so observations for evaluation (see Hastie et al. (2001)). Does this provide enough power to differentiate the classifiers? Based on our result, the answer is that it depends. When the classifiers are parametrically accurate (i.e., with PER of order $n^{-1/2}$) or better, this choice would not work even asymptotically. In applications, particularly challenging high-dimensional cases, classifiers typically have rates slower than $n^{-1/2}$. Then n_1 and n_2 of the same order would be sufficient (which includes the choice of $1/4$) for evaluation. When the sample size is not very large and the competing classifiers are quite accurate, the larger choice of $1/3$ or even $1/2$ can perform better.

2.2. Selection based on CV

To utilize the data in a balanced way in applications, one usually takes a cross-validation method instead of a single data splitting (see, e.g., Lachenbruch

and Mickey (1968), Allen (1974), Stone (1974) and Geisser (1975)). There are several versions of CV in the literature, including a sample of all possible splittings (this is called repeated learning-testing, see, e.g., Burman (1989) or Zhang (1993)), or dividing the data into r sub-groups and making predictions for one group at a time based on estimation using the rest of the sub-groups (this is called r-fold CV, see, Breiman, Friedman, Olshen and Stone (1984)). Compared to a single data splitting, these CV methods typically reduce the variability in selection.

Theorem 1 can be extended to these CV methods. We consider the repeated learning-testing version. Let M be an integer. We randomly permute the data M times. After each permutation, we use the first n_1 observations for training the classifiers and use the last n_2 for evaluation. Let $\pi_1, \ldots, \pi_M$ denote the permutations and let $\tau_{\pi_j} = 1$ if δ_1 is selected based on the j-th permutation, $\tau_{\pi_j} = 0$ otherwise. Then the final selection by voting is: select δ_1 if and only if $\sum_{j=1}^{M} \tau_{\pi_j} \geq M/2$.

Corollary 2. *Under the same conditions as in Theorem* 1*, the CV selection method is consistent.*

Note that for Corollary 2, it does not matter how many permutations are done for voting. Of course, in practice, a reasonable number of permutations is helpful to reduce the chance of accidental selection of a classifier simply due to the randomness of data splitting. The conclusion also applies to multi-fold cross validation and other versions of CV methods.

2.3. Examples

Here we consider three examples. Two toy examples are used to illustrate the influence of the essential error probability difference, a feature of classification.

Example 1. Suppose $f(x) \equiv 1/2$. Then the silly rule that always classifies a case as class 1 (or 0) is a Bayes rule. Consider a classifier which randomly assigns a label with probability 1/2 and another classifier which randomly assigns a case as class 1 with probability $1/2 - \epsilon_n$ for some small ϵ_n. Then, because $P(\widehat{Y}_{2,i} \neq \widehat{Y}_{1,i} | Z_1)$ is of order 1, the essential error probability difference R_n is of order $|\epsilon_n|$. Then we need $n_2 \epsilon_{n_1}^2 \to \infty$ for consistency. When one estimates the parameter $P(Y = 1)$ by $(1/n) \sum_{i=1}^{n} Y_i$, then ϵ_n is of order $O_p(n^{-1/2})$. Consequently, for this situation, we need to choose $n_2/n_1 \to \infty$ to ensure consistency in selection.

Example 2. For two classifiers δ_1 and δ_2, let $B_{in} = \{x : \delta_i(x; Z^n) = 1\}$. Suppose that δ_2 is more conservative than δ_1 in assigning label 1 in the sense that $B_{2n} \subset B_{1n}$. Suppose that $f(x) \geq 1/2$ on $A_n = B_{1n} \backslash B_{2n}$. Then $CPER(\delta_2; n) - CPER(\delta_1; n) = P(A_n | Z_1)$. For this case, the essential error probability difference is of order 1. Consequently, for consistency in selection, it suffices to have

$n_2 P(A_{n_1}|Z_1) \to \infty$. For a case with $P(A_{n_1}|Z_1)$ of order $n_1^{-1/2}$, we need only that $n_2^2/n_1 \to \infty$, which is much less stringent than required in Example 1.

Example 3. Let $\mathcal{X}$ be $[0,1]^d$, with a moderate or large d. Consider the logistic regression model on the conditional probability function

$$f(x) = f(x;\theta) = \frac{\exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d)}{1 + \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d)}.$$

Suppose that X has Lebesgue density $p(x) > 0$ on $[0,1]^d$. Let $\widehat{\theta}$ be the MLE of θ and let the estimator of f be $f(x;\widehat{\theta})$. The resulting plug-in classifier is $\delta_1(x) = I_{\{f(x;\widehat{\theta}) \geq 1/2\}}$. Under this model, if $\theta_i \neq 0$ for at least one $1 \leq i \leq d$, and if f is not always above or below $1/2$ on $[0,1]^d$ (to avoid triviality), then the classifier δ_1 has PER of order $O(1/n)$. To protect from model mis-specification, we consider a classifier based on the more relaxed assumption that f is in a Sobolev class with unknown order of interaction and smoothness, as follows.

For $r \geq 1$, $\mathbf{l} = (l_1, \ldots, l_r)$ with nonnegative integer components l_i, define $|\mathbf{l}| = \sum_{i=1}^r l_i$. Let $\mathbf{z}_r = (z_1, \ldots, z_r) \in [0,1]^r$. Let $D^{\mathbf{l}}$ denote the differentiation operator $D^{\mathbf{l}} = \partial^{|\mathbf{l}|}/\partial z_1^{l_1} \cdots \partial z_r^{l_r}$. For an integer α, define the Sobolev norm to be $\| g \|_{W_2^{\alpha,r}} = \| g \|_2 + \sum_{|\mathbf{l}|=\alpha} \int_{[0,1]^r} |D^{\mathbf{l}} g|^2 d\mathbf{z}_r$. Let $W_2^{\alpha,r}(C)$ denote the set of all functions g on $[0,1]^r$ with $\| g \|_{W_2^{\alpha,r}} \leq C$. Then consider the following function classes on $[0,1]^d$ of different interaction orders and smoothness:

$$\begin{aligned}
S_1(\alpha;C) &= \{\textstyle\sum_{i=1}^d g_i(x_i) : g_i \in W_2^{\alpha,1}(C), 1 \leq i \leq d\},\\
S_2(\alpha;C) &= \{\textstyle\sum_{1\leq i<j\leq d} g_{i,j}(x_i,x_j) : g_{i,j} \in W_2^{\alpha,2}(C), 1 \leq i < j \leq d\},\\
&\vdots\\
S_d(\alpha;C) &= W_2^{\alpha,d}(C),
\end{aligned}$$

with $\alpha \geq 1$ and $C > 0$. From Yang (1999a), the minimax rate of convergence of PER over $S_r(\alpha;C)$ is $n^{-\alpha/(2\alpha+r)}$, which does not depend on the input dimension d, but rather on the true interaction order as in Stone (1985).

Since r and α are unknown, Yang (1999b) considered tensor-product spline models of different interaction orders and used a penalized maximum likelihood criterion to choose the spline order, the number of knots and the interaction order jointly. The resulting plug-in classifier was shown to adaptively achieve the minimax rate $n^{-\alpha/(2\alpha+r)}$ whenever f is in $S_r(\alpha;C)$ over $1 \leq r \leq d$ and $\alpha \geq 1$. Note that the minimax rates for the Sobolev classes are all slower than $n^{-1/2}$. Indeed, when f is not parametrically simple around $f = 1/2$ (as is the case for logistic regression, or expressed by a margin assumption by Mammen and Tsybakov (1999)), one cannot expect a faster rate of convergence than $n^{-1/2}$.

From above, when the logistic regression model holds, δ_1 is better. When it does not, but f is in one of the Sobolev classes and is not a monotone function in any linear combination of the input variables, δ_1 does not converge at all in PER. For this example, $p_n = n^{-1}$ and $q_n = n^{-\alpha/(2\alpha+r)}$ when the logistic regression model holds, and $p_n = 1$ and $q_n = n^{-\alpha/(2\alpha+r)}$ otherwise. In both cases, there is not good control over s_{n_1}. Thus for consistently selecting the better classifier, by Theorem 1, we need the data splitting ratio to satisfy $n_2 n_1^{-2\alpha/(2\alpha+r)} \to \infty$ and $n_1 \to \infty$. Since α and r are unknown, it suffices to take n_2 and n_1 of the same order.

2.4. Cross validation paradox

Suppose a statistician's original data splitting scheme works for consistency in selection. Now suppose that the same amount of (or more) independent and identically distributed data is given to the statistician. Obviously with more data, he can make the estimation accuracy better for each candidate procedure, and can also make the evaluation component more reliable. Thus he decides to add half of the new data to the estimation part and the remaining half to the evaluation part. He naturally thinks that with improvement in both the training and evaluation components, the comparison of the candidate classification procedures becomes more reliable.

But this may not be the case at all! With the original data splitting ratio, the performance difference of the two learning methods is large enough relative to the evaluation size. But when the estimation size is increased, e.g., by half of the original sample size, since the estimation accuracy is improved for both of the classifiers, their difference may no longer be distinguishable with the same order of evaluation size (albeit increased). This is quite clear from the previous subsection.

The surprising requirement of the evaluation part in CV to be dominating in size (i.e., $n_2/n_1 \to \infty$) for differentiating nested parametric models was discovered by Shao (1993) in the context of linear regression.

A simulation study

We present a simulation result to demonstrate the cross validation paradox. We compare two different uses of Fisher's linear discrimination analysis (LDA) method in R with library MASS (by Venables and Ripley).

At the sample size $n = 100$, for 40 observations with $Y = 1$, we generate three independent random variables X_1, X_2, X_3, all standard normal; for the remaining 60 observations with $Y = 0$, we generate the three predictors also independent, but with $N(0.4, 1)$, $N(0.3, 1)$ and $N(0, 1)$ distributions, respectively. Then X_3 is not useful for classifying Y. We compare LDA based on only X_1 and X_2 with

LDA based on all of the three predictors. Obviously, the first one is expected to give a better classifier.

We split the 100 observations in ratio 30/70 (70 for evaluation) for comparing the two classifiers by CV. With 100 such random splittings of the data, the first classifier is declared winner if it performs no worse than the other on the evaluation set, on average. One thousand replications of this are used to approximate the probability that the first classifier is preferred by the CV method. Then suppose that we have two hundred additional observations with 80 at $Y = 1$ and 120 at $Y = 0$, and the predictors are generated in the same way as above. We randomly select 100 of the additional observations and add them to the estimation set, add the remaining 100 to the evaluation set, do estimation and prediction, repeat this 100 times to reduce the effect of splitting bias, and approximate the probability of selecting the first classifier again by Monte Carlo. We continue doing this until the total sample size is 900. The Monte Carlo approximations of the true probabilities that the better use of LDA is the winner in the CV comparison are all based on 1,000 independent replications. The results are in Table 1. The ratios in the parentheses of the first row are the corresponding splitting ratios for the full data.

Table 1. More observations can harm CV selection of the better classifier.

	$n=100$ (30/70)	300 (130/170)	500 (230/270)	700 (330/370)	900 (430/470)
Sel. Prob.	0.835	0.825	0.803	0.768	0.772

Clearly, with more observations added to both the original estimation and evaluation sets, the ability to detect the better classifier by CV is actually decreased. The reason, again, is that the equal splitting of the additional observations for adding to the estimation and evaluation sets makes the decreased difference (in accuracy) between the two classifiers trained on the estimation set less distinguishable with not enough increase of the evaluation size. In contrast, if we maintain the ratio of 30/70, the probabilities of selecting the better classifier are significantly improved over that from the equal splitting of the additional observations, as shown in Table 2. Furthermore, if we increase the proportion of the evaluation set as the sample size increases, the CV comparison of the two classifiers does an even better job when we have more observations, as seen in Table 3. Note that the fractions of the evaluation size at the five sample sizes are 70%, 75%, 80%, 85% and 90%, respectively. The probability of selecting the better classifier reaches 97.5%.

Table 2. Probability of selecting the better classifier: constant splitting ratio.

	$n = 100$ (30/70)	300 (90/210)	500 (150/350)	700 (210/490)	900 (270/630)
Sel. Prob.	0.835	0.892	0.868	0.882	0.880

Table 3. Probability of selecting the better classifier: increasing fraction for evaluation.

	$n = 100$ (30/70)	300 (75/225)	500 (100/400)	700 (105/595)	900 (90/810)
Sel. Prob.	0.835	0.912	0.922	0.936	0.976

In summary, more is not necessarily better for cross validation comparison of learning methods!

2.5. Risk of the selected classifier

It is useful to emphasize that a distinction should be made between different uses of CV. One is for choosing a tuning parameter, where the concern is mostly on the final classifier. Another is for comparing classifiers, where one is mainly interested in finding the best classifier. For the former, deleting a small proportion of cases is not necessarily inappropriate, and even delete-one can be sufficient (see, e.g., Li (1987) and Shao (1997) in a regression context). For the latter, however, under relatively few situations, we can have n_2 of a smaller order than n_1.

For a better understanding of the difference between selecting the better classifier and pursuing accuracy in classification with selection, we give a simple risk bound below. For simplicity, consider a single data splitting as in Theorem 1.

Theorem 2. *Let $\widehat{\delta}$ be the selected classifier. Then*

$$PER(\widehat{\delta}; n) \leq \min_{j=1,2} PER(\delta_j; n_1) + \frac{4 \log n_2 + 3}{3n_2} + \sqrt{\frac{2 \log n_2}{n_2}} \sqrt{P\left(\hat{Y}_{1,n_1+1} \neq \hat{Y}_{2,n_1+1}\right)}.$$

First note that for a typical classification problem, $PER(\delta; n)$ converges at the same rate as $PER(\delta; n_1)$ as long as n_1 is of the same order as n. Consider three scenarios: $\min_{j=1,2} PER(\delta_j; n)$ converges at the parametric rate $n^{-1/2}$ (S1); $\min_{j=1,2} PER(\delta_j; n)$ converges no faster than $n^{-1/2}(\log n)^{1/2}$ (S2); $\min_{j=1,2} PER(\delta_j; n)$ converges faster than $n^{-1/2}$ (S3). Note that fast rate (S3) scenarios have been given in the literature (see, Mammen and Tsybakov (1999), Shen et al. (2003), Tsybakov (2004)), they are obtained under margin assumptions; the slower rate of (S2) is a typical minimax rate without the margin assumption (see Yang (1999a)).

If we have consistency in selection with $n_2 \max(p_{n_1}^2, q_{n_1}^2)/s_{n_1} \to \infty$, then $\max(p_{n_1}^2, q_{n_1}^2)$ is of larger order than s_{n_1}/n_2. Ignoring a possible logarithmic term, the two additional terms in the risk bound in Theorem 1 may or may not affect the rate of convergence. In the best situation, the risk of the selected classifier converges as fast as $\min_{j=1,2} PER(\delta_j; n_1)$. When n_1 is forced to be of a smaller order than n for consistency in selection (as is possibly needed for S1 and S3), this rate is sub-optimal (compared to $\min_{j=1,2} PER(\delta_j; n)$).

Now consider a proper splitting for optimal risk rate. From the risk bound, if the classifiers converge at the parametric rate or more slowly (S1 and S2), for the final selected classifier to converge optimally or near optimally (i.e., ignoring a logarithmic factor), it is sufficient to take n_1 and n_2 of the same order. (The risk bound is sometimes also optimal for S3 in order, and can even be as small as $\log n/n$ if $P(\hat{Y}_{1,n_1+1} \neq \hat{Y}_{2,n_1+1}) = O(n^{-1})$, which occurs e.g., when δ_1 and δ_2 are both based on correct parametric models but one with extra parameters.) In contrast, for consistency in selection, from Theorem 1, we must require that one of the two classifiers be asymptotically better than the other, and taking n_2 of the same order as n is not sufficient for consistency in selection for S1 and S3.

In real applications, once a classifier is selected, one typically re-trains the classifier using the full data, though theoretical properties are hard to obtain. Also, the difference between a single splitting and multiple splittings (CV) can show up in terms of the risk property of the selected procedure. For example, in regression, it is known that delete-1 CV shares an asymptotic efficiency property of AIC in nonparametric estimation with linear approximation models, while one cannot expect this property to hold with a single $(n-1):1$ splitting of the data for training and evaluation. In contrast, for consistency in selection, there does not seem to be a major difference between a single splitting and multiple splittings. Corollary 2 shows that when a single splitting works, CV also works. None of the results in the literature seem to provide evidence to suggest that for finding the better classifier, multiple splitting can rescue an inconsistent single splitting based method.

3. Confidence Interval for Comparing Classifiers

How much confidence do we have in the observed error rate difference of two classifiers through a data splitting approach? In this section, via a central limit theorem, we give an asymptotic confidence interval for the error probability difference. Normal approximation based confidence intervals for error probability have been considered in the literature. However, when a confidence interval (CI) is sought for the difference of the error probabilities, the issue becomes more complicated. In fact, the normal approximation may be invalid if the splitting ratio is not appropriate.

3.1. A single data splitting

For $n_1 + 1 \leq i \leq n$, let

$$W_i = \begin{cases} 0, & \text{if } \widehat{Y}_{1,i} = \widehat{Y}_{2,i}, \\ -1, & \text{if } \widehat{Y}_{1,i} = Y_i \neq \widehat{Y}_{2,i}, \\ 1, & \text{if } \widehat{Y}_{2,i} = Y_i \neq \widehat{Y}_{1,i}. \end{cases}$$

Obviously, given Z_1, $W_{n_1+1}, \ldots, W_n$ are i.i.d. with mean $-\Delta_{n_1}$, where $\Delta_{n_1} = P(\widehat{Y}_{1,n_1+1} = Y_{n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1) - P(\widehat{Y}_{2,n_1+1} = Y_{n_1+1} \neq \widehat{Y}_{1,n_1+1}|Z_1)$. One would then apply the Central Limit Theorem for $\overline{W} = 1/n_2 \sum_{i=n_1+1}^{n} W_i$, which estimates the conditional error probability difference $-\Delta_{n_1}$. However, the normal approximation can be misleading because $-\Delta_{n_1}$ is not a fixed quantity but often converges to zero as $n \to \infty$. The issue becomes one of conditions under which we can use the normal approximation to build a CI. Let v be the variance of W_{n_1+1} conditional on Z_1.

Theorem 3. *Suppose* $n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1) \to \infty$ *in probability, then*

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{\sum_{i=n_1+1}^{n} (W_i - EW_i)}{\sqrt{n_2 v}} \leq x \right) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy \right| \to 0 \text{ in probability.}$$

Remarks.

1. For a confidence interval for the error probability of a given classifier, as long as the Bayes error probability is not zero, a central limit theorem typically does apply, and thus the normal approximation based CI is fine (as long as n_2 is large enough).
2. It is possible that the CI based on the normal approximation of $\overline{W}$ does not contain zero, yet $n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)$ is small. For such a case, one may erroneously claim that one classifier is better than the other when one actually does not have the declared confidence.

Theorem 3 enables the construction of an asymptotic confidence interval for the conditional error probability difference. Let $sd_W = ((n_2-1)^{-1} \sum_{i=n_1+1}^{n} (W_i - \overline{W})^2)^{1/2}$ be the standard deviation of the W_i's. Under the condition that $n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1) \to \infty$ in probability, it is easy to show that sd_W is a consistent estimator of $\sqrt{v}$ in the sense that $sd_W/\sqrt{v} \to 1$ in probability. Thus, given confidence level $1-\alpha$, an asymptotic confidence interval for the difference of the conditional error probability between the two classifiers, $CPE(\delta_1; n_1) - CPE(\delta_2; n_1)$, is $\overline{W} \pm z_{\alpha/2} sd_W/\sqrt{n_2}$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The condition that $n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)$ is large enough for the normal approximation to be accurate can be assessed by examining $D = \sum_{i=n_1+1}^{n} I_{\{\widehat{Y}_{1,i} \neq \widehat{Y}_{2,i}\}}$, which is an unbiased estimator of $n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)$ given Z_1. Based on a simple analysis using the Berry-Essen bound (see the proof of Theorem 3), to guarantee the normal approximation error to be less than 25%, D needs to be as large as 48^2. Obviously this is based on an upper bound and thus is conservative. In addition, multiple data splittings may significantly help

reduce the variability in classifier comparison. Nonetheless, claiming one classifier is better than another via cross validation with D as small as 4 (even if the permutation standard error is very small, see Section 3.3), as in some empirical results with small sample sizes in the literature, seems highly questionable. A larger threshold, say 10, would be more reasonable.

3.2. Comparing classifiers based on CIs for the individual error rates or a CI for the error probability difference?

In the empirical comparisons of classifiers in statistical applications, typically the confidence intervals (or standard errors) of the error rates of the candidate classifiers are given. Although these intervals do provide information on comparing classifiers, when the sample size is not very large, their use is a suboptimal practice due to the loss of power in differentiating the classifiers in terms of accuracy. Indeed, as is expected, working with the differences W_i can be much more reliable.

For $n_1 + 1 \le i \le n$, let $G_i = I_{\{Y_i \neq \widehat{Y}_{1,i}\}}$ and $H_i = I_{\{Y_i \neq \widehat{Y}_{2,i}\}}$. Then a $1-\alpha$ asymptotic confidence interval for the error probability $CPE(\delta; n_1) = P(\delta(X; Z_1) \neq Y | Z_1)$ is

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} G_i \pm z_{\alpha/2}\frac{\sqrt{\widehat{v}_1}}{\sqrt{n_2}}$$

and for δ_2 is

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} H_i \pm z_{\alpha/2}\frac{\sqrt{\widehat{v}_2}}{\sqrt{n_2}},$$

where $\widehat{v}_1$ and $\widehat{v}_2$ are the sample variances of G_i and H_i, respectively. Without enough knowledge on the relationship between G_i and H_i, by the Bonferroni method, observing that ${n_2}^{-1}\sum_{i=n_1+1}^{n} G_i - n_2^{-1}\sum_{i=n_1+1}^{n} H_i = \overline{W}$, we declare the classifiers to be different in accuracy when $|\overline{W}| > z_{\alpha/4}(\sqrt{\widehat{v}_1} + \sqrt{\widehat{v}_2})/\sqrt{n_2}$. This has an asymptotic type I error at most α. In contrast, one may use the difference-based CI $\overline{W} \pm z_{\alpha/2}\sqrt{\widehat{v}}/\sqrt{n_2}$, where $\widehat{v} = sd_w^2$. When $|\overline{W}| > z_{\alpha/2}\sqrt{\widehat{v}/n_2}$, we declare the classifiers to be different. This has an asymptotic type I error α. The comparison of this method and the earlier one then amounts to the comparison of $z_{\alpha/4}(\sqrt{\widehat{v}_1} + \sqrt{\widehat{v}_2})$ and $z_{\alpha/2}\sqrt{\widehat{v}}$.

It is easily shown that $(\sqrt{\widehat{v}_1} + \sqrt{\widehat{v}_2})/\sqrt{\widehat{v}}$ can approach ∞ in probability. As an example, suppose Y takes values 0 or 1 with roughly equal probability. If $G_i = H_i$ for almost all i, but $\overline{G}$ is not close to zero or 1, then the aforementioned variance estimate ratio is very large (and can be arbitrarily large), in which case the use of the two CI is much worse than the single CI method. More generally, $\widehat{v}_1$ and $\widehat{v}_2$ typically do not converge to zero in probability. Thus if

$P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1) \to 0$ in probability, then the variance ratio converges to ∞ in probability. Note also that the ratio $(\sqrt{\widehat{v}_1} + \sqrt{\widehat{v}_2})/\sqrt{\widehat{v}}$ can easily be shown to be always lower bounded by 1. This confirms the simple fact that the two CI method should not be used for comparing classifiers due to its low power. This is particularly relevant, for example, for classification problems (e.g., cancer classification) based on gene expression data, where the sample size is typically small (sometimes even less than 50).

Obviously, we are not criticizing the construction of CI's for the error probabilities of the candidate classifiers, which is useful in its own right.

3.3. CI based on cross validation

In Section 3.1, the construction of the confidence interval is quite simple (although the validity of normal approximation should not be taken for granted), but the result depends on the outcome of data splitting. When multiple splittings are used in CV, the theoretical issues involved in constructing a rigorous confidence interval become complicated and, to the best of our knowledge, little theoretical advancement has been made.

From a practical perspective, a natural thing to try is the following. One modifies the CI in Section 3.1 in terms of the center and the variance estimate: the modified CI for the error rate difference is $\overline{\overline{W}} \pm z_{\alpha/2}\sqrt{\widetilde{v}}/\sqrt{n_2}$, where for the center, one replaces $\overline{W}$ by the average over the multiple splittings, denoted by $\overline{\overline{W}}$, and $\widetilde{v}$ is a modified variance estimate. The previous CI in Section 3.1 is for the conditional error probability difference given Z_1. Due to multiple splittings and averaging, it seems that $\overline{\overline{W}}$ might be more appropriate for the unconditional error probability difference, though obviously the overall expectation is unchanged by the averaging, i.e., $E\overline{W} = E\overline{\overline{W}}$. This averaging, however, makes the variance of $\overline{\overline{W}}$ very hard to analyze. There are other ways one might consider replacing $sd_W/\sqrt{n_2}$.

One way is the following. After each splitting of the data, one obtains the difference of the error rates of the competing classifiers. Let $\overline{W}^j$, $1 \leq j \leq N$, denote the difference based on the jth splitting of the data, where N is the total number of data splitting. Then one finds the standard error of $\overline{\overline{W}}$:

$$se_{split,\overline{W}} = \sqrt{\frac{\frac{1}{N-1}\sum_{j=1}^{N}\left(\overline{W}^j - \overline{\overline{W}}\right)^2}{N}}.$$

A similar formula for estimating the error rate of a classifier is often used in the literature as the standard error of the estimate. This would be correct if the $\overline{W}^j$ were independent for different j, which of course they are not. Nonetheless, the

formula is partially meaningful. It captures the part of the uncertainty of the average error rate $\overline{\overline{W}}$ due to randomness in splitting of the data. Adding error bars from these standard errors in a graph of error rates (or difference) of competing learning methods may seem to provide useful information. However, this is not appropriate. The splitting standard error, whether done by random splitting or by a mutli-fold CV fashion, conveys only the reliability of using a subset of all possible data splittings with the same ratio for training and evaluation. Note that the splitting standard error gets smaller as the number of splitting increases. This error bar diminishes when computation over all possible splittings is feasible (in which case the splitting standard error is theoretically zero). With $se_{split,\overline{W}}$ small for a large number of data splitting, any difference, no matter how small it is, would become significant. In general, it seems there is little relationship between the splitting standard error $se_{split,\overline{W}}$ and the actual standard error of the estimate $\overline{\overline{W}}$. Despite explanations and warnings given in the literature (e.g., Efron and Tibshirani (1997) and Dietterich (1998)), the splitting standard error has still been mistakenly interpreted as the real standard error in statistical applications.

With a single splitting, let $sd_{W,1}$ be the standard deviation of $W_{n_1+1}, \ldots, W_n$. It estimates the conditional standard deviation of W_{n_1+1} given the estimation part of the data. The standard error of $\overline{W}$ as an estimate of the conditional mean of W_{n_1+1} (again given Z_1) is $se_W = sd_{W,1}/\sqrt{n_2}$. One may average this over the N splittings of the data to get $\overline{se}_W$.

A simulation study

We conduct a simple simulation for numerical understanding. We compare two learning methods: Fisher's linear discrimination analysis (LDA) and the support vector machine (SVM).

Consider three independent predictors, all standard normal. The conditional probability function is

$$f(x) = \frac{\exp(1 + 0.2x_1 + 0.2x_2 + 3x_3)}{1 + \exp(1 + 0.2x_1 + 0.2x_2 + 3x_3)},$$

with the probability of $Y = 1$ being roughly 0.6. We took the sample size to be 100. The simulation was conduced in R with libraries MASS (by Venables and Ripley) and e1071 (by Dimitriadou et al.). We chose the default settings of the controlling parameters for both methods (note that our interest here was not on optimizing the tuning parameters). Four values of n_2 were considered: 75, 50, 25 and 10. The number of random data splitting was 200, and 200 replications of the whole process were done to simulate the theoretical means and standard

deviations of interest. The error rate difference refers to the error rate of the LDA minus that of SVM.

The results are in Table 4. For the last two columns, the numbers in the parentheses are the corresponding standard deviations. Note that if one uses all 100 observations to train the two classifiers, based on additional simulations the mean of $\overline{W}$ is -0.0155 and the standard deviation of $\overline{W}$ is 0.011 (of course, these (simulated) theoretical values are not available in applications).

Table 4. Comparing standard error estimates.

	$E\overline{W}$	$E\overline{\overline{W}}$	$sd\left(\overline{W}\right)$	$sd\left(\overline{\overline{W}}\right)$	$Esd_{split,\overline{W}}$	$Ese_{split,\overline{W}}$	Ese_W	$E\overline{se}_W$
$n_2 = 75$	-0.042	-0.044	0.050	0.011	0.053	0.0037	0.039 (0.003)	0.039 (0.010)
$n_2 = 50$	-0.023	-0.025	0.052	0.014	0.044	0.0030	0.038 (0.014)	0.040 (0.004)
$n_2 = 25$	-0.019	-0.019	0.053	0.017	0.052	0.0030	0.047 (0.028)	0.048 (0.010)
$n_2 = 10$	-0.018	-0.017	0.076	0.021	0.079	0.0056	0.054 (0.060)	0.057 (0.017)

From the table, not surprisingly, given n_2, the simulated values of $E\overline{W}$ and $E\overline{\overline{W}}$ are very close (they should be the same) but their standard deviations are very different (as expected), with the single splitting standard deviation about two times larger. This clearly supports the common practice of doing multiple data splitting. Regarding the choice of n_2, observe that $E\overline{\overline{W}}$ decreases as n_2 decreases, and in the meantime, $sd(\overline{\overline{W}})$ increases, which strongly suggests that for this example, for the comparison of the two learning methods, the choice of n_2 large (50 or 75) is better than small.

The table also shows that the splitting standard deviation and standard error ($sd_{split,\overline{W}}$ or $se_{split,\overline{W}}$) are inappropriate as uncertainty measures of $\overline{\overline{W}}$. The other estimates, $\overline{se}_W$ and se_W, are quite similar to each other. They are still much larger than the actual standard deviation of $\overline{\overline{W}}$, but are better than the splitting standard deviation or the splitting standard error.

To summarize, this example demonstrates:

1. Multiple data splittings and averaging help to improve the accuracy of the estimates of the classification error rates and their difference.
2. The splitting standard deviation or splitting standard error are definitely not suitable for describing the uncertainty in comparing classifiers.
3. Although conservative, $\overline{se}_W$ (or se_W) at least yields a valid confidence interval for comparing classifiers.

In general, without additional assumptions, CV is probably one of the most reliable methods for comparing classifiers, although it may reduce the effective

sample size. For the case of bootstrap-based error rate estimation, Efron and Tibshirani (1997) derived standard error formulas that were demonstrated to perform well.

4. Instability of CV selection in splitting ratio

Recall that for consistency in selection, with a single splitting or CV, n_2 needs to be suitably large. A major concern for this approach is whether the accuracy comparison at a significantly reduced sample size can tell the truth at the full sample size. To that end, we can investigate the agreement of CV at different splitting ratios. If the comparisons at various choices of splitting ratios (in a proper range) actually tell the same story, then we have more confidence on the relative performance of the classifiers. In contrast, if the comparison is sensitive to the choice of the splitting ratio, it indicates that the relative performance of the classifiers is perhaps in a transition zone, and thus one should be careful about the outcome of the comparisons.

Consider the following *sequential instability in selection* for assessing the tendency of selecting a different classifier due to sample size reduction. For each choice of n_1, let λ_{n_1} be the fraction of times δ_1 is selected over the different data splittings in cross validation. Then we plot (or table) λ_{n_1} versus n_1 (or n_1/n) to gain a graphical understanding of the effect of n_1. If the λ_{n_1} values are stable over a range of small n_1, then the data reduction in CV does not seem to be a serious problem. In contrast, if λ_{n_1} changes quickly around small n_1, it indicates that we may be in an unstable sample size zone in terms of the relative performance of the classifiers and thus should not be overly confident about our comparison result. Note that this approach provides additional information that is not available with a fixed choice of n_1.

Example 4. Follow the same set-up as in Section 3.3. We randomly generated a data set of 100 observations. We obtained λ_{n_1} for 6 choices of n_1 based on 500 random splittings of the data. The results are in Table 5.

Table 5. Sequential instability in selection.

	$n_1 = 50$	60	70	80	90	95
λ_{n_1}	83.4%	82.6%	81.6%	80.0%	82.4%	92.2%

For this data set, there is little sequential instability in selection. Clearly LDA is strongly preferred, and there should be little concern on sample size reduction in CV.

5. Proofs

Proof of Theorem 1. Without loss of generality, assume that δ_1 is asymptotically better than δ_2. Let $\Delta = -E(W_{n_1+1}|Z_1)$ be the conditional expectation of $-W_{n_1+1}$ given the first part of the data. It is the difference of the conditional error probability of the two classifiers. Indeed,

$$\begin{aligned}
-\Delta &= P\left(\widehat{Y}_{2,n_1+1} = Y_{n_1+1} \neq \widehat{Y}_{1,n_1+1}|Z_1\right) - P\left(\widehat{Y}_{1,n_1+1} = Y_{n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1\right) \\
&= P\left(\widehat{Y}_{2,n_1+1} = Y_{n_1+1}|Z_1\right) - P\left(\widehat{Y}_{2,n_1+1} = Y_{n_1+1} = \widehat{Y}_{1,n_1+1}|Z_1\right) \\
&\quad - \left(P\left(\widehat{Y}_{1,n_1+1} = Y_{n_1+1}|Z_1\right) - P\left(\widehat{Y}_{1,n_1+1} = Y_{n_1+1} = \widehat{Y}_{2,n_1+1}|Z_1\right)\right) \\
&= P\left(\widehat{Y}_{2,n_1+1} = Y_{n_1+1}|Z_1\right) - P\left(\widehat{Y}_{1,n_1+1} = Y_{n_1+1}|Z_1\right) \\
&= P\left(\widehat{Y}_{1,n_1+1} \neq Y_{n_1+1}|Z_1\right) - P\left(\widehat{Y}_{2,n_1+1} \neq Y_{n_1+1}|Z_1\right).
\end{aligned}$$

Under the condition that δ_1 is asymptotically better than δ_2, we know that for an arbitrary $\epsilon > 0$, there exists n_0 such that when $n_1 \geq n_0$, with probability at least $1-\epsilon$, $CPER(\delta_2; n_1) - CPER(\delta_1; n_1) \geq c_\epsilon CPER(\delta_1; n_1) \geq 0$. Let A be the exceptional event. Then, conditional on the first part of the data, on A^c the mis-selection probability satisfies

$$\begin{aligned}
P\left(TE(\delta_1) > TE(\delta_2)|Z_1\right) &= P\Big(\sum_{i=n_1+1}^{n} I_{\{Y_i \neq \widehat{Y}_{1,i}\}} > \sum_{i=n_1+1}^{n} I_{\{Y_i \neq \widehat{Y}_{2,i}\}}|Z_1\Big) \\
&= P\Big(\sum_{i=n_1+1}^{n} W_i > 0|Z_1\Big) \\
&= P\Big(\sum_{i=n_1+1}^{n} (W_i - EW_i) > n_2\Delta|Z_1\Big) \\
&\leq \exp\Big(-\frac{n_2\Delta^2}{2V + \frac{4}{3}\Delta}\Big),
\end{aligned}$$

where the inequality follows from the Bernstein's inequality (see, e.g., Pollard (1984)), and V is the conditional variance of W_{n_1+1} given Z_1. Note that $V \leq E(W_{n_1+1}^2|Z_1) = P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1})$. Consequently, on A^c, we have

$$P\left(TE(\delta_1) > TE(\delta_2)|Z_1\right) \leq \exp\Big(-\frac{n_2\Delta^2}{2P\left(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1\right) + \frac{4}{3}\Delta}\Big).$$

Since the upper bound is no larger than 1, a sufficient condition for $P(TE(\delta_1) > TE(\delta_2)) \to 0$ is that $P(A) \to 0$ and the exponent in the right hand side of the above inequality converges to $-\infty$ in probability. That $P(A) \to 0$ follows

from the assumption in the theorem if $n_1 \to \infty$. The second condition is equivalent to $n_2\Delta \to \infty$ in probability and $n_2\Delta R_n \to \infty$ in probability. Since the essential error probability difference $R_n = \Delta/P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)$ is always between 0 and 1, the last two conditions reduce to $n_2\Delta R_n \to \infty$ in probability. Under the assumption that δ_1 is asymptotically better than δ_2, and that $CPER(\delta_1; n_1)$, $CPER(\delta_2; n_1)$, and $P(\widehat{Y}_{2,n_1+1} \neq \widehat{Y}_{1,n_1+1}|Z_1)$ converge exactly at rates $p_{n_1}, q_{n_1}, s_{n_1}$ respectively. This is equivalent to $n_2 q_{n_1}^2/s_{n_1} \to \infty$ and completes the proof of Theorem 1.

Proof of Corollary 2. Without loss of generality, assume that δ_1 is asymptotically better than δ_2. Since the observations are i.i.d., the random variables $\tau_{\pi_0}, \tau_{\pi_j}$ $(1 \le j \le M)$ are identically distributed, where π_0 denotes the original order of observations. Then $E(M^{-1}\sum_{j=1}^M \tau_{\pi_j}) = E\tau_{\pi_0} = P(TE(\delta_1) \le TE(\delta_2))$. From Theorem 1, we know $P(TE(\delta_1) \le TE(\delta_2)) \to 1$, and thus $E(M^{-1}\sum_{j=1}^M \tau_{\pi_j}) \to 1$. Together with the fact that $M^{-1}\sum_{j=1}^M \tau_{\pi_j}$ is between zero and 1, we must have $M^{-1}\sum_{j=1}^M \tau_{\pi_j} \to 1$ in probability. Consequently, $P(\sum_{j=1}^M \tau_{\pi_j} \ge M/2) \to 1$. This completes the proof of Corollary 2.

Proof of Theorem 2. From the proof of Theorem 1, when $\Delta > 0$, we have

$$P\left(TE(\delta_1) - TE(\delta_2) \ge 0|Z_1\right) \le \exp\Big(-\frac{n_2\Delta^2}{2V + \frac{4}{3}\Delta}\Big).$$

Let $\Delta^2/(V+2\Delta/3) = t_n$. Taking the positive root, we get $\Delta = t_n/3 + \sqrt{t_n^2/9 + Vt_n}$ $\le 2t_n/3 + \sqrt{Vt_n}$. We take $n_2 t_n/2 = \log n_2$, i.e., $t_n = 2\log n_2/n_2$. Then when $\Delta \ge 2t_n/3 + \sqrt{Vt_n}$,

$$\begin{aligned}
&P\Big(TE(\delta_1) - TE(\delta_2) \ge 0|Z_1\Big)\\
&\le P\Big(\sum_{i=n_1+1}^{n}(W_i - EW_i) \ge \Big(\frac{2n_2t_n}{3} + n_2\sqrt{Vt_n}\Big)|Z_1\Big)\\
&\le P\Big(\sum_{i=n_1+1}^{n}(W_i - EW_i) \ge \Big(\frac{n_2t_n}{3} + n_2\sqrt{\frac{t_n^2}{9} + Vt_n}\Big)|Z_1\Big)\\
&\le \exp\Big(-\frac{n_2t_n}{2}\Big) = n_2^{-1}.
\end{aligned}$$

Let δ_* be the classifier (trained on Z_1) that minimizes the error probability at sample size n_1 over the two candidate classifiers. Let S denote the event that $\Delta \ge 4\log n_2/(3n_2) + \sqrt{Vt_n}$, where Δ is now the conditional error probability difference between the other classifier and δ_* given Z_1. Let δ_B denote a Bayes

classifier. Then

$$
\begin{aligned}
&PER(\widehat{\delta};n)\\
&=E\left(P\left(\widehat{\delta}(X)\neq Y|Z^n\right)\right)-PE^*\\
&=E\left(P\left(\delta_*(X)\neq Y,\delta_*=\widehat{\delta},S|Z^n\right)\right)+E\left(P\left(\widehat{\delta}(X)\neq Y,\delta_*\neq\widehat{\delta},S|Z^n\right)\right)\\
&\quad+E\left(P\left(\delta_*(X)\neq Y,\delta_*=\widehat{\delta},S^c|Z^n\right)\right)+E\left(P\left(\widehat{\delta}(X)\neq Y,\delta_*\neq\widehat{\delta},S^c|Z^n\right)\right)-PE^*\\
&\le E\left(P\left(\delta_*(X)\neq Y|Z^n\right)I_{\{\delta_*=\widehat{\delta}\}}\right)+n_2^{-1}+E\left(P\left(\widehat{\delta}(X)\neq Y|Z^n\right)I_{\{\delta_*\neq\widehat{\delta}\}\cap S^c}\right)-PE^*\\
&\le E\left(\left[P\left(\delta_*(X)\neq Y|Z^n\right)-P(\delta_B(X)\neq Y)\right]I_{\{\delta_*=\widehat{\delta}\}}\right)+n_2^{-1}\\
&\quad+E\left(\left[P\left(\widehat{\delta}(X)\neq Y|Z^n\right)-P(\delta_B(X)\neq Y)\right]I_{\{\delta_*\neq\widehat{\delta}\}\cap S^c}\right)\\
&\le E\left(\left[P\left(\delta_*(X)\neq Y|Z^n\right)-P(\delta_B(X)\neq Y)\right]I_{\{\delta_*=\widehat{\delta}\}}\right)+n_2^{-1}\\
&\quad+E\left(\left[P\left(\delta_*(X)\neq Y|Z^n\right)-P(\delta_B(X)\neq Y)\right]I_{\{\delta_*\neq\widehat{\delta}\}\cap S^c}\right)+E\left(\frac{4\log n_2}{3n_2}+\sqrt{Vt_n}\right)\\
&\le PER(\delta_*;n_1)+n_2^{-1}+\frac{4\log n_2}{3n_2}+E\sqrt{\frac{2\log n_2}{n_2}V}\\
&\le \min_{j=1,2}PER(\delta_j;n_1)+\frac{4\log n_2+3}{3n_2}+\sqrt{\frac{2\log n_2}{n_2}}\sqrt{EV}.
\end{aligned}
$$

The conclusion follows. This completes the proof of Theorem 2.

Proof of Theorem 3. We apply Berry Esseen Theorem (see, e.g., Stroock (1993)). For our case of comparing two classifiers, conditional on Z_1, $W_{n_1+1},\ldots,$ W_n are i.i.d., and obviously since $|W_i-EW_i|\le 2$, we have the upper bound

$$
\begin{aligned}
&\sup_{-\infty<x<\infty}\left|P\left(\frac{\sum_{i=n_1+1}^{n}(W_i-EW_i)}{\sqrt{n_2v}}\le x\right)-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{y^2}{2}}dy\right|\\
&\le\frac{c}{\sqrt{n_2}}\frac{E|W_i-EW_i|^3}{\left(E\left(W_i-EW_i\right)^2\right)^{\frac{3}{2}}}.
\end{aligned}
$$

Let $P(W_i=1)=p$, $P(W_i=-1)=q$, and $P(W_i=0)=1-p-q$. Then

$$
\begin{aligned}
E|W_i-EW_i|^3&=p|1-(p-q)|^3+q|-1-(p-q)|^3+(1-p-q)\,|(p-q)|^3\\
&=p\,(1-(p-q))^3+q\,(1+(p-q))^3+(1-p-q)\,|(p-q)|^3\\
&=(p+q)-3(1-p-q)(p-q)^2-(p-q)^4+(1-p-q)\,|(p-q)|^3\\
&\le(p+q)+|(p-q)|^3\\
&\le 2(p+q),
\end{aligned}
$$

and

$$\begin{aligned} E|W_i - EW_i|^2 &= p|1-(p-q)|^2 + q| - 1 - (p-q)|^2 + (1-p-q)\,|(p-q)|^2 \\ &= (p+q) - (p-q)^2 \\ &\geq (p+q)\,(1-|p-q|)\,. \end{aligned}$$

Therefore

$$\frac{E|W_i - EW_i|^3}{\left(E\,(W_i - EW_i)^2\right)^{\frac{3}{2}}} \leq \frac{2(p+q)}{(p+q)^{\frac{3}{2}}\,(1-|p-q|)^{\frac{3}{2}}} = \frac{2}{(p+q)^{\frac{1}{2}}\,(1-|p-q|)^{\frac{3}{2}}}.$$

Consequently, we need $\sqrt{n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)} \to \infty$ in probability for the bound in (1) to converge in probability to zero.

For the bound in (1) to be smaller than 25%, based on the above calculation, we might ask that $6 \cdot 2/\sqrt{n_2 P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)} \leq 0.25$. With $P(\widehat{Y}_{1,n_1+1} \neq \widehat{Y}_{2,n_1+1}|Z_1)$ estimated by the number of disagreements between the two classifiers on the test data (denoted by D), this becomes $D \geq 48^2$ and completes the proof of Theorem 3.

Acknowledgement

This work was supported by US NSF CAREER grant # 0094323. The author wishes to thank a referee and the Editor for helpful comments.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory* (Edited by B. N. Petrov and F. Csaki), 267-281. Akademia Kiado, Budapest.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125-127.

Breiman, L. Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

Burman, P. (1989). A comparative study of ordinary cross-validation, ν-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503-514.

Devroye, L. (1988). Automatic pattern recognition: a study of the probability of error. *IEEE Trans. Pattern Anal. Mach. Intell.* **10**, 530-543.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* **10**, 1895-1924.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92**, 548-560.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70**, 320-328.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the* 14*th International Joint Conference on Artificial Intelligence* (Edited by C. S. Mellish), 1137-1143. Morgan Kaufmann Publishers, Inc.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1-11.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 958-975.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808-1829.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley, New York.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statistics* **6**, 461-464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221-242.

Shen, X., Tseng, G. C., Zhang, X. and Wong, W.H. (2003). On ψ-Learning. *J. Amer. Statist. Assoc.* **98**, 724-734.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36**, 111-147.

Stroock, D. W. (1993). *Probability Theory: An Analytic View*. Cambridge University Press. Cambridge, UK.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 135-166.

Yang, Y. (1999a). Minimax nonparametric classification − Part I: rates of convergence. *IEEE Trans. Inform. Theory* **45**, 2271-2284.

Yang, Y. (1999b). Minimax nonparametric classification—Part II: model selection for adaptation. *IEEE Trans. Inform. Theory* **45**, 2285-2292.

Yang, Y. (2005a). Can the strengths of AIC and BIC be shared?–A conflict between model identification and regression estimation. *Biometrika* **92** 937-950.

Yang, Y. (2005b). Consistency of cross validation for comparing regression procedures. Submitted.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21**, 299-313.

School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, MN 55455, U.S.A.

E-mail: yyang@stat.umn.edu

(Received June 2005; accepted September 2005)