

Classifier Technology and the Illusion of Progress

David J. Hand

Abstract. A great many tools have been developed for supervised classification, ranging from early methods such as linear discriminant analysis through to modern developments such as neural networks and support vector machines. A large number of comparative studies have been conducted in attempts to establish the relative superiority of these methods. This paper argues that these comparisons often fail to take into account important aspects of real problems, so that the apparent superiority of more sophisticated methods may be something of an illusion. In particular, simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.

Key words and phrases: Supervised classification, error rate, misclassification rate, simplicity, principle of parsimony, population drift, selectivity bias, flat maximum effect, problem uncertainty, empirical comparisons.

1. INTRODUCTION

In supervised classification, one seeks to construct a rule which will allow one to assign objects to one of a prespecified set of classes based solely on a vector of measurements taken on those objects. Construction of the rule is based on a “design set” or “training set” of objects with known measurement vectors and for which the true class is also known: one essentially tries to extract from the design set the information which is relevant to distinguishing between the classes in terms of the given measurements. It is because the classes are known for the members of this initial data set that the term “supervised” is used: it is as if a “supervisor” has provided these class labels.

Such problems are ubiquitous and, as a consequence, have been tackled in several different research areas, including statistics, machine learning, pattern recognition, computational learning theory and data mining. As a result, a tremendous variety of algorithms

and models has been developed for the construction of such rules. A partial list includes linear discriminant analysis, quadratic discriminant analysis, regularized discriminant analysis, the naive Bayes method, logistic discriminant analysis, perceptrons, neural networks, radial basis function methods, vector quantization methods, nearest neighbor and kernel nonparametric methods, tree classifiers such as CART and C4.5, support vector machines and rule-based methods. New methods, new variants on existing methods and new algorithms for existing methods are being developed all the time. In addition, different methods for variable selection, handling missing values and other aspects of data preprocessing multiply the number of tools yet further. General theoretical advances have also been made which have resulted in improved performance at predicting the class of new objects. These include ideas such as bagging, boosting and more general ensemble classifiers. Furthermore, apart from the straightforward development of new rules, theory and practice have been developed for performance assessment. A variety of criteria have been investigated, including measures based on the receiver operating characteristic (ROC) and Brier score, as well as the standard measure of misclassification rate. Subtle estimators of these have been

David J. Hand is Professor, Department of Mathematics and Institute for Mathematical Science, Imperial College, Huxley Building, 180 Queen's Gate, London SW7 2AZ, United Kingdom (e-mail: d.j.hand@imperial.ac.uk).

developed, such as jackknife, cross-validation and a variety of bootstrap methods, to overcome the potential optimistic bias which results from simply reclassifying the design set.

An examination of recent conference proceedings and journal articles shows that such developments are continuing. In part this is because of new computational developments that permit the exploration of new ideas, and in part it is because of the emergence of new application domains which present new twists on the standard problem. For example, in bioinformatics there are often relatively few cases but many thousands of variables. In such situations the risk of overfitting is substantial and new classes of tools are required. General references to work on supervised classification include [11, 13, 33, 38, 44].

The situation to date thus appears to be one of very substantial theoretical progress, leading to deep theoretical developments and to increased predictive power in practical applications. While all of these things are true, it is the contention of this paper that the practical impact of the developments has been inflated; that although progress has been made, it may well not be as great as has been suggested. The arguments for this assertion are described in the following sections. They develop ideas introduced by Hand [12, 14, 15, 18, 19] and Jamain and Hand [24]. The essence of the argument is that the improvements attributed to the more advanced and recent developments are small, and that aspects of real practical problems often render such small differences irrelevant, or even unreal, so that the gains reported on theoretical grounds, or on empirical comparisons from simulated or even real data sets, do not translate into real advantages in practice. That is, progress is far less than it appears.

These ideas are described in four steps.

First, model-fitting is a sequential process of progressive refinement, which begins by describing the largest and most striking aspects of the data structure, and then turns to progressively smaller aspects (stopping, one hopes, before the process begins to model idiosyncrasies of the observed sample of data rather than aspects of the true underlying distribution). In Section 2 we show that this means that the large gains in predictive accuracy in classification are won using relatively simple models at the start of the process, leaving potential gains which decrease in size as the modeling process is taken further. All of this means that the extra accuracy of the more sophisticated approaches, beyond that attained by simple models, is

achieved from “minor” aspects of the distributions and classification problems.

Second, in Section 3 we argue that in many, perhaps most, real classification problems the data points in the design set are not, in fact, randomly drawn from the same distribution as the data points to which the classifier will be applied. There are many reasons for this discrepancy, and some are illustrated. It goes without saying that statements about classifier accuracy based on a false assumption about the identity of the design set distribution and the distribution of future points may well be inaccurate.

Third, when constructing classification rules, various other assumptions and choices are often made which may not be appropriate and which may give misleading impressions of future classifier performance. For example, it is typically assumed that the classes are objectively defined, with no arbitrariness or uncertainty about the class labels, but this is sometimes not the case. Likewise, parameters are often estimated by optimizing criteria which are not relevant to the real aim of classification accuracy. Such issues are described in Section 4 and, once again, it is obvious that these introduce doubts about how the claimed classifier performance will generalize to real problems.

The phenomena with which we are concerned in Sections 3 and 4 are related to the phenomenon of *overfitting*. A model overfits when it models the design sample too closely rather than modeling the distribution from which this sample is drawn. In Sections 3 and 4 we are concerned with situations in which the models may accurately reflect the design distributions (so they do not underfit or overfit), but where they fail to recognize that these distributions, and the apparent classification problems described, are in fact merely a single such problem drawn from a notional distribution of problems. The real aim might be to solve a rather different problem. One might thus describe the issue as one of *problem* uncertainty. To take a familiar example, which we do not explore in detail in this paper because it has been explored elsewhere, the relative *costs* of different kinds of misclassification may differ and may be unknown. A very common resolution is to assume equal costs (Jamain and Hand [24] found that most comparative studies of classification rules made this assumption) and to use straightforward error rate as the performance criterion. However, equality is but one choice, and an arbitrary one at that, and one which we suspect is in fact rarely appropriate. In assuming equal costs, one is adopting a particular problem which may not be the one which is really to be solved. Indeed, things are even worse than

this might suggest, because relative misclassification costs may change over time. Provost and Fawcett [36] have described such situations: “Comparison often is difficult in real-world environments because key parameters of the target environment are not known. The optimal cost/benefit tradeoffs and the target class priors seldom are known precisely, and often are subject to change (Zahavi and Levin [47]; Friedman and Wyatt [8]; Klinkenberg and Thorsten [29]). For example, in fraud detection we cannot ignore misclassification costs or the skewed class distribution, nor can we assume that our estimates are precise or static (Fawcett and Provost [6]).”

Moving on, our fourth argument is that classification methods are typically evaluated by reporting their performance on a variety of real data sets. However, such empirical comparisons, while superficially attractive, have major problems which are often not acknowledged. In general, we suggest in Section 5 that no method will be universally superior to other methods: relative superiority will depend on the type of data used in the comparisons, the particular data sets used, the performance criterion and a host of other factors. Moreover, the relative performance will depend on the experience the person making the comparison has in using the methods, and this experience may differ between methods: researcher A may find that his favorite method is best, merely because he knows how to squeeze the best performance from this method.

These various arguments together suggest that an apparent superiority in classification accuracy, obtained in “laboratory conditions,” may not translate to a superiority in real-world conditions and, in particular, the apparent superiority of highly sophisticated methods may be illusory, with simple methods often being equally effective or even superior in classifying new data points.

2. MARGINAL IMPROVEMENTS

This section demonstrates that the extra performance to be achieved by more sophisticated classification rules, beyond that attained by simple methods, is small. It follows that if aspects of the classification problem are not accurately described (e.g., if incorrect distributions have been used, incorrect class definitions have been adopted, inappropriate performance comparison criteria have been applied, etc.), then the reported advantage of the more sophisticated methods may be incorrect. Later sections illustrate how some inaccuracies in the classification problem description can arise.

2.1 A Simple Example

Statistical modeling is a sequential process in which one gradually refines the model to provide a better and better fit to the distributions from which the data were drawn. In general, the earlier stages in this process yield greater improvement in model fit than later stages. Furthermore, if one looks at the historical development of classification methods, then the earlier approaches involve relatively simple structures (e.g., the linear forms of linear or logistic discriminant analysis), while more recent approaches involve more complicated structures (e.g., the decision surfaces of neural networks or support vector machines). It follows that the simple approaches will have led to greater improvement in predictive performance than the later approaches which are necessarily trying to improve on the predictive performance obtained by the simpler earlier methods. Put another way, there is a law of diminishing returns.

Although this paper is concerned with supervised classification problems, it is illuminating to examine a simple regression case. Suppose that we have a single response variable y which is to be predicted from d variables $(x_1, \dots, x_d)^T = \mathbf{x}$. Suppose also that the correlation matrix of $(\mathbf{x}^T, y)^T$ has the form

$$(2.1) \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} (1 - \rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T & \boldsymbol{\tau} \\ \boldsymbol{\tau}^T & 1 \end{bmatrix}$$

with $\Sigma_{11} = (1 - \rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T$, $\Sigma_{12} = \Sigma_{21}^T = \boldsymbol{\tau}$ and $\Sigma_{22} = 1$, where \mathbf{I} is the $d \times d$ identity matrix, $\mathbf{1} = (1, \dots, 1)^T$ of length d and $\boldsymbol{\tau} = (\tau, \dots, \tau)^T$ of length d . That is, the correlation between each pair of predictor variables is ρ , and the correlation between each predictor variable and the response variable is τ . Suppose also that $\rho, \tau \geq 0$. This condition is not necessary for the argument which follows; it merely allows us to avoid some detail.

Let $V(d)$ be the conditional variance of y given the values of d predictor variables \mathbf{x} , as above. Standard results give this conditional variance as

$$(2.2) \quad V(d) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Using the result that

$$(2.3) \quad \begin{aligned} \Sigma_{11}^{-1} &= [(1 - \rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T]^{-1} \\ &= \frac{1}{1 - \rho} \left\{ \mathbf{I} - \frac{\rho\mathbf{1}\mathbf{1}^T}{1 + (d - 1)\rho} \right\} \end{aligned}$$

[with $-(d-1)^{-1} < \rho < 1$, so that Σ_{11} is positive definite], leads to

$$\begin{aligned} V(d) &= 1 - \tau^T \frac{1}{1-\rho} \left\{ \mathbf{I} - \frac{\rho \mathbf{1}\mathbf{1}^T}{1+(d-1)\rho} \right\} \tau \\ (2.4) \quad &= 1 - \frac{d\tau^2}{1-\rho} + \frac{\rho d^2 \tau^2}{(1+(d-1)\rho)(1-\rho)}. \end{aligned}$$

From this it follows that the reduction in conditional variance due to adding an extra predictor variable, x_{d+1} (also correlated ρ with the other predictors and τ with the response variable), is

$$\begin{aligned} X(d+1) &= V(d) - V(d+1) \\ (2.5) \quad &= \frac{\tau^2}{1-\rho} \\ &\quad + \frac{\rho \tau^2}{1-\rho} \left[\frac{d^2}{1+(d-1)\rho} - \frac{(d+1)^2}{1+d\rho} \right]. \end{aligned}$$

Note that the condition $-(d-1)^{-1} < \rho < 1$ must still be satisfied when d is increased.

Now consider two cases:

Case 1. When the predictor variables are uncorrelated, $\rho = 0$. From (2.5), we obtain $X(d+1) = \tau^2$. That is, if the predictor variables are mutually uncorrelated and each has correlation τ with the response variable, then each additional predictor reduces the variance of the conditional variance of y given the predictors by τ^2 . [Of course, by setting $\rho = 0$ in (2.4) we see that this is only possible up to $d = \tau^{-2}$ predictors. With this many predictors the conditional variance of y given \mathbf{x} has been reduced to zero.]

Case 2. $\rho > 0$. Plots of $V(d)$ for $\tau = 0.5$ and for a range of ρ values are shown in Figure 1. When there is reasonably strong mutual correlation between the

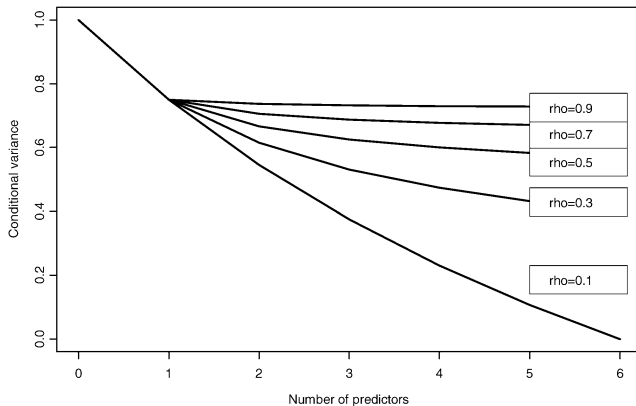


FIG. 1. Conditional variance of response variable as additional predictors are added for $\tau = 0.5$. A range of values of ρ is shown.

predictor variables, the earliest ones contribute substantially more to the reduction in variance remaining unexplained than do the later ones. The case $\rho = 0$ consists of a diagonal straight line running from 1 down to zero. In the case $\rho = 0.9$, almost all of the variance in the response variable is explained by the first chosen predictor.

This example shows that the reduction in conditional variance of the response variable decreases with each additional predictor we add, even though each predictor has an identical correlation with the response variable (provided this correlation is greater than 0). The reason for the reduction is, of course, the mutual correlation between the predictors: much of the predictive power of a new predictor has already been accounted for by the existing predictors.

In real applications, the situation is generally even more pronounced than in this illustration. Usually, in real applications, the predictor variables are not identically correlated with the response, and the predictors are selected sequentially, beginning with those which maximally reduce the conditional variance. In a sense, then, the example above provides a lower bound on the phenomenon: in real applications the proportion of the gains attributable to the early steps is even greater.

2.2 Decreasing Bounds on Possible Improvement

We now return to supervised classification. For illustrative purposes, suppose that misclassification rate is the performance criterion, although similar arguments apply with other criteria. Ignoring issues of overfitting, adding additional predictor variables can only lead to a decrease in misclassification rate. The simplest model is that which uses no predictors, leading, in the two-class case, to a misclassification rate of $m_0 = \pi_0$, where π_0 is the prior probability of the smaller class. Suppose that a predictor variable is now introduced which has the effect of reducing the misclassification rate to $m_1 < m_0$. Then the scope for further improvement is only m_1 , which is less than the original scope m_0 . Furthermore, if $m_1 < m_0 - m_1$, then all future additions necessarily improve things by less than the first predictor variable. In fact, things are even more extreme than this: one cannot further reduce the misclassification rate by more than $m_1 - m_b$, where m_b is the Bayes error rate. To put it another way, at each step the maximum possible increase in predictive power decreases, so it is not surprising that, in general, at each step the additional contribution to predictive power decreases.

2.3 Effectiveness of Simple Classifiers

Although the literature contains examples of artificial data which simple models cannot separate (e.g., intertwined spirals or checkerboard patterns), such data sets are exceedingly rare in real life. Conversely, in the two-class case, although few real data sets have exactly linear decision surfaces, it is common to find that the centroids of the predictor variable distributions of the classes are different, so that a simple linear surface can do surprisingly well as an estimate of the true decision surface. This may not be the same as “can do surprisingly well in classifying the points,” since in many problems the Bayes error rate is high, meaning that no decision surface can separate the distributions of such problems very well. However, it means that the dramatic steps in improvement in classifier accuracy are made in the simple first steps. This is a phenomenon which has been noticed by others (e.g., Rendell and Seshu [37]; Shavlik, Mooney and Towell [41]; Mingers [34]; Weiss, Galen and Tadepalli [45]; Holte [22]). Holte [22], in particular, carried out an investigation of this phenomenon. His “simple classifier” (called $1R$) consists of a partition of a single variable, with each cell of the partition possibly being assigned to a different class: it is a multiple-split single-level tree classifier. A search through the variables is used to find that which yields the best predictive accuracy. Holte compared this simple rule with C4.5, a more sophisticated tree algorithm, finding that “on most of the datasets studied, $1R$ ’s accuracy is about 3 percentage points lower than C4’s.”

We carried out a similar analysis. Perhaps the earliest classification method formally developed is Fisher’s linear discriminant analysis [7]. Table 1 shows misclassification rates for this method and for the best

performing method we could find in a search of the literature (these data were abstracted from the data accumulated by Jamain [23] and Jamain and Hand [24]) for a randomly selected sample of ten data sets. The first numerical column shows the misclassification rate of the best method we found (m_T), the second shows that of linear discriminant analysis (m_L), the third shows the default rule of assigning every point to the majority class (m_0) and the final column shows the proportion of the difference between the default rule and the best rule which is achieved by linear discriminant analysis $[(m_0 - m_L)/(m_0 - m_T)]$. It is likely that the best rules, being the best of rules which many researchers have applied, are producing results near the Bayes error rate.

The striking thing about Table 1 is the large values of the percentages of classification accuracy gained by simple linear discriminant analysis. The lowest percentage is 85% and in most cases over 90% of the achievable improvement in predictive accuracy, over the simple baseline model, is achieved by the simple linear classifier.

I am grateful to Willi Sauerbrei for pointing out that when the error rates of both the best method and the linear method are small, the large proportion in achievable accuracy which can be obtained by the linear method corresponds to the error rate of the linear method being a large multiple of that of the best method. For example, in the most extreme case in Table 1, the results for the segmentation data show that the linear discrimination error rate is nearly six times that of the best method. On the other hand, when the error rates are small, this large difference will correspond to only a small proportion of new data points. Small differences in error rate are susceptible to the issues raised in Sections 3 and 4: they may vanish when problem uncertainties are taken into account.

TABLE 1
Performance of linear discriminant analysis and the best result we found on ten randomly selected data sets

Data set	Best method e.r.	Lindisc e.r.	Default rule	Prop linear
Segmentation	0.0140	0.083	0.760	0.907
Pima	0.1979	0.221	0.350	0.848
House-votes16	0.0270	0.046	0.386	0.948
Vehicle	0.1450	0.216	0.750	0.883
Satimage	0.0850	0.160	0.758	0.889
Heart Cleveland	0.1410	0.141	0.560	1.000
Splice	0.0330	0.057	0.475	0.945
Waveform21	0.0035	0.004	0.667	0.999
Led7	0.2650	0.265	0.900	1.000
Breast Wisconsin	0.0260	0.038	0.345	0.963

2.4 The Flat Maximum Effect

Even within the context of classifiers defined in terms of simple linear combinations of the predictor variables, it has often been observed that the major gains are made by (for example) weighting the variables equally, with only little further gains to be had by careful optimization of the weights. This phenomenon has been termed the *flat maximum* effect [13, 43]: in general, often quite large deviations from the optimal set of weights will yield predictive performance not substantially worse than the optimal weights. An informal argument that shows why this is often the case is as follows.

Let the predictor variables be $(x_1, \dots, x_d)^T = \mathbf{x}$ and, for simplicity, assume that $E(x_i) = 0$ and $V(x_i) = 1$ for $i = 1, \dots, d$. Let $\Sigma = \{r\}_{ij}$ be the correlation matrix between these variables. Now define two weighted sums

$$w = \sum_{i=1}^d w_i x_i \quad \text{and} \quad v = \sum_{i=1}^d v_i x_i,$$

using respective weight vectors (w_1, \dots, w_d) and (v_1, \dots, v_d) . In general, $r(w, v)$, the correlation between w and v , can take extreme values of $+1$ and -1 , but suppose we restrict the weights to be nonnegative, $w_i, v_i \geq 0$ for $i = 1, \dots, d$, and also require $\sum w_i = 1$ and $\sum v_i = 1$. Using these conditions, a little algebra shows that

$$r(v, w) \geq \sum_i \sum_j v_i w_j r(x_i, x_j).$$

Now, with equal weights, $v_i = 1/d, i = 1, \dots, n$, we obtain

$$\begin{aligned} r(v, w) &\geq \frac{1}{d} \sum_i \sum_j w_j r(x_i, x_j) \\ &\geq \frac{1}{d} \sum_i \sum_j w_j r(x_i, x_k), \end{aligned}$$

where $k = \arg \min_j r(x_i, x_j)$.

From this,

$$\begin{aligned} r(v, w) &\geq \frac{1}{d} \sum_i \sum_j w_j r(x_i, x_k) \\ &= \frac{1}{d} \sum_i r(x_i, x_k). \end{aligned}$$

In words, the correlation between an arbitrary weighted sum of the x variables (with weights summing to 1) and the simple combination using equal weights is bounded

below by the smallest row average of the entries in the correlation matrix of the x variables. Hence if the correlations are all high, the simple average will be highly correlated with any other weighted sum: the choice of weights will make little difference to the scores. The gain to be made by the extra effort of optimizing the weights may not be worth the effort.

2.5 An Example

As a simple illustration of how increasing model complexity leads to a decreasing rate of improvement, we fitted models to the sonar data from the University of California, Irvine (UCI) data base. This data set consists of 208 observations, 111 of which belong to the class ‘‘metal’’ and 97 of which belong to the class ‘‘rock.’’ There are 60 predictor variables. The data were randomly divided into two parts, and a succession of neural networks with increasing numbers of hidden nodes was fitted to half of the data, with the other half being used as a test set. The error rates are shown in Figure 2. The left-hand point, corresponding to 0 nodes, is the baseline misclassification rate achieved by assigning everyone in the test set to the larger class. The error bars are 95% confidence intervals calculated from 100 networks in each case. Figure 3 shows a similar plot, but this time for a recursive partitioning tree classifier applied to the same data. The horizontal axis shows increasing numbers of leaf nodes. Standard methods of tree construction were used, in which a large tree is pruned back to the requisite number of nodes. In both of these figures we see the dramatic improvement arising from fitting the first nontrivial model. This far exceeds the subsequent improvement obtained in any later step.

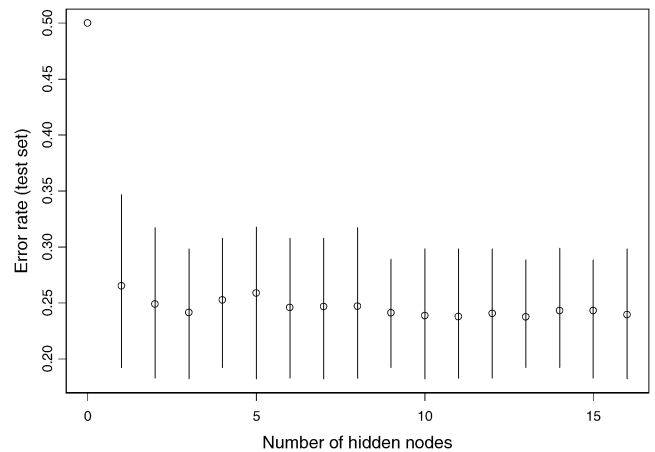


FIG. 2. Effect on misclassification rate of increasing the number of hidden nodes in a neural network to predict the class of the sonar data.

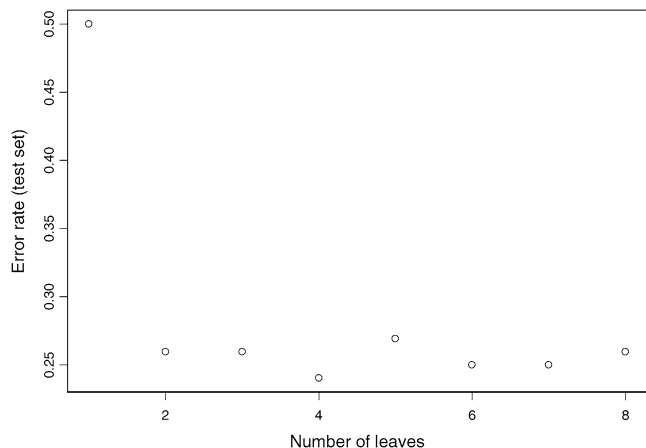


FIG. 3. Effect on misclassification rate of increasing the number of leaves in a tree classifier to predict the class of the sonar data.

3. DESIGN SAMPLE SELECTION

Intrinsic to the classical supervised classification paradigm is the assumption that the data in the design set are randomly drawn from the same distribution as the points to be classified in the future. Sometimes slight variants of the sampling scheme are used, for example, drawing samples separately from each class, but the assumption that future points to be classified are drawn from the same distributions as the design set is always made. Unfortunately, as we illustrate in this section, there are several reasons why this assumption may not be justified. In fact, as with our suggestion that the common choice of equal misclassification costs may be more often inappropriate than appropriate, we suspect that the assumption that the design distribution is representative of the distribution from which future points will be drawn is perhaps more often incorrect than correct.

If the distribution underlying the design data and that underlying future points to be classified do differ, then elaborate optimization of the classifier using the design data may be wasted effort: the performance difference between two classifiers may be irrelevant in the context of the differences arising between the design and future distributions. In particular, we suggest, more sophisticated classifiers, which almost by definition model small idiosyncrasies of the distribution underlying the design set, will be more susceptible to wasting effort in this way: the grosser features of the distributions (modeled by simpler methods) are more likely to persist than the smaller features (modeled by the more elaborate methods).

3.1 Population Drift

A fundamental assumption of the classical paradigm is that the various distributions involved do not change over time. In fact, in many applications this is unrealistic and the population distributions are nonstationary. For example, it is unrealistic in most commercial applications concerned with human behavior: customers will change their behavior with price changes, with changes to products, with changing competition and with changing economic conditions. Hoadley [21] remarked “the test sample is supposed to represent the population to be encountered in the future. But in reality, it is usually a random sample of the current population. High performance on the test sample does not guarantee high performance on future samples, things do change” and “there is always a chance that a variable and its relationships will change in the future. After that, you still want the model to work. So don’t make any variable dominant.” He is cautioning against making the model fit the design distribution too well. The last point about not making any variable dominant is related to the flat maximum effect, described above.

Among the most important reasons for changes to the distribution of applicants are changes in marketing and advertising practices. Changes to the distributions that describe the customers explain why, in the credit scoring and banking industries [16, 20, 39, 42], the classification rules used to predict which applicants are likely to default on loans are updated every few months: their performance degrades, not because the rules themselves change, but because the distributions to which they are being applied change [27].

An example of this is given in Figure 4. The available data consisted of the true classes (“bad” or “good”)

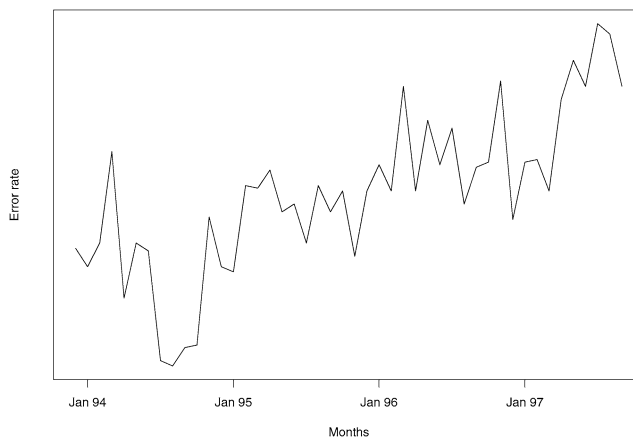


FIG. 4. Evolution of misclassification rate of a classifier built at the start of the period.

and the values of 17 predictor variables for 92,258 customers taking out unsecured personal loans with a 24-month term given by a major UK bank during the period 1 January 1993 to 30 November 1997; 8.86% of the customers belonged to the bad class. The figure shows how the misclassification rate for a classification rule built on data just preceding the start of the displayed period changed over time. Since the coefficients of the classifier were not changing, the deterioration in performance must be due to shifts in the distributions of customers over time.

An illustration of how this “population drift” phenomenon affects different classifiers differentially is given in Figure 5. For the purposes of this illustration we used a linear discriminant analysis (LDA) as a simple classifier and a tree model as a more complicated classifier. For the design set we used customers 1, 3, 5, 7, . . . , 4999. We then applied the classifiers to alternate customers, beginning with the second, up to the 60,000th customer. This meant that different customers were used for designing and testing, even during the initial period, so that there would be no overfitting in the reported results. Figure 5 shows lowess smooths of the misclassification cost [i.e., misclassification rate, with customers from each class weighted so that $c_0/c_1 = \pi_1/\pi_0$, where c_i is the cost of misclassifying a customer from class i and π_i is the prior (class size) of class i]. As can be seen from the figure, the tree classifier (the lower curve) is initially superior (has smaller loss), but after a time its superiority begins to fade. Superficial examination of the figure might suggest that the effect takes a long time to become apparent, not really manifesting itself until around the 40,000th customer, but consider that, in an application such as this, *the data are always retrospective*. In the

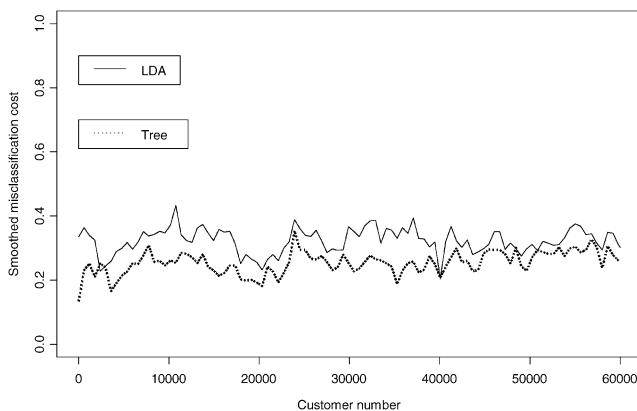


FIG. 5. Lowess smooths of cost-weighted misclassification rate for a tree model and LDA applied to customers 2, 4, 6, . . . , 60,000.

present case, one cannot determine the true class until the entire 24 month loan term has elapsed. [In fact, of course, this is not quite true: if a customer defaults before the end of the term, then their class (bad) is known, but otherwise their true (good or bad) class is not known until the end, so that to obtain an unbiased sample, one has to wait until the end. Survival analysis models can be constructed to allow for this, but that is leading us away from the point.] For our problem, to accumulate an unbiased sample of 5000 customers with known true outcome, one would have to wait until two years after the 5000th customer had been accepted. In terms of the horizontal axis in Figure 5, this means that the model would be built, and would be initially used at around the time that the 40,000th customer was being considered. The figure shows that this is just when the model degrades. The changes in population structure which occurred during the two years which elapsed while we waited for the true classes of the 5000 design set customers to become known have reduced any advantage that the more sophisticated tree model may have.

In summary, the apparent superiority of the more sophisticated tree classifier over the very simple linear discriminant classifier is seen to fade when we take into account the fact that the classifiers must necessarily be applied in the future to distributions which are likely to have changed from those which produced the design set. Since, as demonstrated in Section 2, the simple linear classifier captures most of the separation between the classes, the additional distributional subtleties captured by the tree method become less and less relevant when the distributions drift. Only the major aspects are still likely to hold.

The impact of population drift on supervised classification rules is nicely described by the American philosopher Eric Hoffer, who said, “In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists.”

3.2 Sample Selectivity Bias

The previous subsection considered the impact on classification rules of distributions which changed over time. There is little point in optimizing the rule to the extent that it models aspects of the distributions and decision surface which are likely to have changed by the time the rule is applied. Similar futility applies if a selection process means that the design sample is drawn from a distribution distorted in some way from that to which the classification rule is to be applied.

In fact, I suspect that this may be common. Consider, for example, a classification rule aimed at differential medical diagnosis or medical screening. The rule will have been developed on a sample of cases (including members of each class). Perhaps these cases will be drawn from a particular hospital, clinic or health district. Now all sorts of demographic, social, economic and other factors influence who seeks and is accepted for treatment, how severe the cases being treated are, how old they are and so on. In general, it would be risky to assume that these selection criteria are the same for all hospitals, clinics or health districts. This means that the fine points of the classification rule are unlikely to hold. One might expect its coarser features to be true across different such sets of cases, but the detailed aspects will reflect particular properties of the population from which the design data were drawn. In fact, there are some subtleties here. Suppose that the classification rule follows the diagnostic paradigm [directly modeling $p(c|\mathbf{x})$, the probability of class membership, c , given the descriptor vector \mathbf{x}], rather than the sampling paradigm [which models $p(c|\mathbf{x})$ indirectly from the $p(\mathbf{x}|c)$ using Bayes' theorem]. Then if \mathbf{x} spans the space of all predictors of class membership and if the model form chosen for $p(c|\mathbf{x})$ includes the "true" model, then sampling distortions based on \mathbf{x} alone will not adversely influence the classifier: the classifier built in one clinic will also apply elsewhere. Of course, it would be a brave person who could confidently assert that these two conditions held. Such subtleties aside, what this means, again, is that effort spent on overrefining the classification model is probably wasted effort and, in particular, that fine differences between different classification rules should not be regarded as carrying much weight.

This problem of sample selection and how it might be tackled has been the subject of intensive research, especially by the medical statistics and econometrics communities, but appears not to have been of great concern to researchers on classification methods. Having said that, one area that involves sample selectivity in classification problems which has attracted research interest arises in the retail financial services industry, as in the previous section. Here, as in that section, the aim is to predict, for example, on the basis of application and other background variables, whether or not an applicant is likely to be a good customer. Those expected to be good are accepted, and those expected to be bad are rejected. For those that have been accepted, we subsequently discover their true good or bad class. For the rejected applicants, however, we never know

whether they are good or bad. The consequence is that the resulting sample is distorted as a sample from the population of applicants, which is our real interest for the future. Measuring the performance or attempting to build an improved classification rule using those individuals for which we do know the true class (which is needed for supervised classification) has the potential to be highly misleading for the overall applicant population. In particular, it means that using highly sophisticated methods to squeeze subtle information from the design data is pointless. This problem is so ubiquitous in the personal financial services sector that it has been given its own name—*reject inference* [17].

4. PROBLEM UNCERTAINTY

Section 3 looked at mismatches between the distributions modeled by the classification rule and the distributions to which it was applied. This is an obvious way in which things may go awry, but there are many others, perhaps not so obvious. This section illustrates just three.

4.1 Errors in Class Labels

The classical supervised classification paradigm is based on the assumption that there are no errors in the true class labels. If one expects errors in the class labels, then one can attempt to build models which explicitly allow for this, and there has been work to develop such models. Difficulties arise, however, when one does not expect such errors, but they nevertheless occur.

Suppose that, with two classes, the true posterior class probabilities are $p(1|\mathbf{x})$ and $p(2|\mathbf{x})$, and that a (small) proportion δ of each class is incorrectly believed to come from the other class at each \mathbf{x} . Denoting the apparent posterior probability of class 1 by $p^*(1|\mathbf{x})$, we have

$$p^*(1|\mathbf{x}) = (1 - \delta)p(1|\mathbf{x}) + \delta p(2|\mathbf{x}).$$

It follows that if we let $r(\mathbf{x}) = p(1|\mathbf{x})/p(2|\mathbf{x})$ denote the true odds and let $r^*(\mathbf{x}) = p^*(1|\mathbf{x})/p^*(2|\mathbf{x})$ denote the apparent odds, then

$$(4.1) \quad r^*(\mathbf{x}) = \frac{r(\mathbf{x}) + \varepsilon}{\varepsilon r(\mathbf{x}) + 1}$$

with $\varepsilon = \delta/(1 - \delta)$.

With small ε , (4.1) is monotonic increasing in $r(\mathbf{x})$, so that contours of $r(\mathbf{x})$ map to corresponding contours of $r^*(\mathbf{x})$. In particular, if the true optimal decision surface is $r(\mathbf{x}) = k$ (k is determined by the relative

misclassification costs), then the optimal decision surface when errors are present is given by $r^*(\mathbf{x}) = k^*$, with $k^* = (k + \varepsilon)/(\varepsilon k + 1)$. Unfortunately, if the occurrence of mislabeling is unsuspected, then $r^*(\mathbf{x})$ will be compared with k rather than k^* . In the case of equal misclassification costs, so that $k = 1$, we have $k^* = k = 1$, so that no problems arise from the misclassification. (Indeed, advantages can even arise: see [9].) However, what happens if $k \neq 1$? It is easy to show that $r^*(\mathbf{x}) > r(\mathbf{x})$ whenever $r(\mathbf{x}) < 1$ and that $r^*(\mathbf{x}) < r(\mathbf{x})$ whenever $r(\mathbf{x}) > 1$. That is, the effect of the errors in class labels is to shrink the posterior class odds toward 1, so that comparing $r^*(\mathbf{x})$ with k rather than k^* is likely to lead to worse performance. There is also a secondary issue, that the shrinkage of $r(\mathbf{x})$ will make it less easy to estimate the decision surface accurately because it is a flatter surface: the variance of the estimated decision surface, from sample to sample, will be greater when there is mislabeling of classes. In such circumstances it is better to stick to simpler models, since the higher order terms of the more complicated models will be very inaccurately estimated.

4.2 Arbitrariness in the Class Definition

The classical supervised classification paradigm also takes as fundamental the fact that the classes are well defined. That is, that there is some fixed clear external criterion which is used to produce the class labels. In many situations, however, this is not the case. In particular, when the classes are defined by thresholding a continuous variable, then there is always the possibility that the defining threshold might be changed. Once again, this situation arises in consumer credit, where it is common to define a customer as “defaulting” if they fall three months in arrears with repayments. This definition, however, is not a qualitative one (contrast has a tumor/does not have a tumor) but is very much a quantitative one. It is entirely reasonable that alternative definitions (e.g., four months in arrears) might be more useful if economic conditions were to change. This is a simple example, but in many situations much more complex class definitions based on logical combinations of numerical attributes, split at fairly arbitrary thresholds, are used. For example, student grades are often based on levels of performance in continuous assessment and examinations. In detecting vertebral deformities in studies of osteoporosis, the ranges of the anterior, posterior and mid heights of the vertebra, as well as functions of these, such as ratios, are combined in quite complicated Boolean conditions to provide the definition (e.g., [10]). Definitions formed

in this sort of way are particularly common in situations that involve customer management. For example, Lewis [31] defined a good account in a revolving credit operation (such as a credit card) as someone whose billing account shows (a) on the books for a minimum of 10 months, (b) activity in 6 of the most recent 10 months, (c) purchases of more than \$50 in at least 3 of the past 24 months and (d) not more than once 30 days delinquent in the past 24 months. A bad account is defined as (a) delinquent for 90 days or more at any time with an outstanding undisputed balance of \$50 or more, (b) delinquent three times for 60 days in the past 12 months with an outstanding undisputed balance on each occasion of \$50 or more or (c) bankrupt while the account was open. Li and Hand [32] gave an even more complicated example from retail banking.

Our concern with these complicated definitions is that they are fairly arbitrary: the thresholds used to partition the various continua are not natural thresholds, but are imposed by humans. It is entirely possible that, retrospectively, one might decide that other thresholds would have been better. Ideally, under such circumstances, one would go back to the design data, redefine the classes and recompute the classification rule. However, this requires that the raw data have been retained at the level of the underlying continua used in the definitions. This is often not the case. The term *concept drift* is sometimes used to describe changes to the definitions of the classes. See, for example, the special issue of *Machine Learning* (1998, Vol. 32, No. 2), Widmer and Kubat [46] and Lane and Brodley [30]. The problem of changing class definitions has been examined in [25, 26] and [28].

If the very definitions of the classes may change between designing the classification rule and applying it, then clearly there is little point in developing an overrefined model for the class definition which is no longer appropriate. Such models fail to take into account all sources of uncertainty in the problem. Of course, this does not necessarily imply that simple models will yield better classification results: this will depend on the nature of the difference between the design and application class definitions. However, there are similarities to the overfitting issue. Overfitting arises when a complicated model faithfully reflects aspects of the design data to the extent that idiosyncrasies of that data, rather than merely of the distribution from which the data arose, are included in the model. Then simple models, which fit the design data less well, lead to superior classification. Likewise, in the present context, a model optimized on the design data class definition is reflecting idiosyncrasies of the design data

which may not occur in application data, not because of random variation, but because of the different definitions of the classes. Thus it is possible that models which fit the design data less well will do better in future classification tasks.

The possibility of arbitrariness in the class definition discussed in this section is quite distinct from the possibility of class priors or relative misclassification costs being changed—referred to in the quote from Provost and Fawcett [36] above—but the possibility of these changes, also, casts doubt on the wisdom of modeling the problem too precisely, that is, of using models which are too sophisticated.

4.3 Optimization Criteria and Performance Assessment

When fitting a model to a design set, one optimizes some criterion of goodness of fit (perhaps modified by a penalization term to avoid overfitting) or of classification performance. Many such measures are in use, including likelihood, misclassification rate, cost-weighted misclassification rate, Brier score, log score and area under the ROC curve. Unfortunately, it is not difficult to contrive data sets for which different optimization criteria lead to (e.g.) linear decision surfaces with very different orientations (even to the extent of being orthogonal). Benton [2, Chap. 4] illustrated this for several real data sets. Clearly, then it is important to specify the criterion to be used when building a classification rule. If the use to which the model will be put is well specified to the extent that a measure of performance can be precisely defined, then this measure should determine the criterion of goodness of fit. All too often, however, there is a mismatch between the criterion used to choose the model, the criterion used to evaluate its performance, and the criterion which actually matters in real application. For example, a common approach might be to use likelihood to estimate a model's parameters, use misclassification rate to assess its performance and use some cost-weighted misclassification rate in practice (e.g., some combination of specificity and sensitivity). In circumstances such as these, it would clearly be pointless to refine the model to a high degree of accuracy from a likelihood perspective, when this may be only weakly related to the real performance objective.

Having said that, one must acknowledge that often precise details of how performance is to be measured in the future cannot be given. For example, in most applications it is difficult to give more than general statements about the relative costs of different kinds

of misclassifications. In such cases it might be worthwhile to choose a criterion that is equivalent to averaging over a range of possible costs: likelihood, the area under a receiver operating characteristic curve and the weighted version of the latter described in [1] can all be regarded as attempts to do that.

5. INTERPRETING EMPIRICAL COMPARISONS

There have been a great many empirical comparisons of the performance of different kind of classification rules. Some of these are in the context of a new method having been developed and the effort to gain some understanding of how it performs relative to existing methods. Other comparisons are purely comparative studies, seeking to make disinterested comparative statements about the relative merits of different methods. At first glance, such comparative studies are useful in shedding light on the different methods, on which generally yield superior performance or on which are to be preferred for particular kinds of data or in particular domains. However, on closer examination, such comparisons have major weaknesses and can even be seriously misleading. Various authors have drawn attention to these problems, including Duin [4], Salzberg [40], Hand [13], Hoadley [21] and Efron [5], so we will only briefly mention some of the main points here; in particular, only those points relative to classification accuracy, rather than other aspects of performance. Jamain and Hand [24] also gave a more detailed review of comparative studies of classification rules.

Different categories of users might be expected to obtain different rankings of classification methods in comparative studies. For example, we can contrast an expert user, who will be able to fine-tune methods, with an inexperienced user, perhaps someone who has simply pulled some standard public-domain software from the web. It would probably be surprising if their rankings did not differ. Moreover, experts will tend to have particular expertise with particular classes of method. Someone expert in neural networks may well achieve superior results with those methods than with support vector machines and vice versa. Taken to an extreme, of course, many comparative studies are made to establish the performance and properties of newly invented methods—by their inventors. One might expect substantial bias in favor of the new methods, compared to what others might be able to achieve, in such studies. Duin [4] pointed out the difficulty of comparing, “in a fair and objective way,” classifiers which require substantial input of expertise (so that domain knowledge

can be taken advantage of) and classifiers which can be applied automatically with little external input of expertise. The two extremes (of what is really a continuum, of course) are appropriate in different circumstances.

The principle of comparing methods by applying them to a collection of disparate real data sets is useful, but has its weaknesses. An obvious one is that different studies use different collections of data sets, so making comparisons difficult. Furthermore, the collection will not be representative of real data sets in any formal sense. Moreover, a potential user is not really interested in some “average performance” over distinct types of data, but really wants to know what will be good for his or her problem, and different people have different problems, with data arising from different domains. A given method may be very poor on most kinds of data, but very good for certain problems.

The widespread use of standard collections of data sets (such as the UCI repository [35]) has clear merits: new methods can be compared with earlier ones on a level playing field. However, this also means that there will be some overfitting both to the individual data sets in the collection and to the collection as a whole. That is, some methods will do well on data sets in the collection purely by chance. Indeed, the more successful the collection is in the sense that more and more people use it for comparative assessments, the more serious this problem will become.

Jamain and Hand [24] pointed out the difficulty of saying exactly what a classification “method” is. Is a neural network with a single hidden node to be regarded as from the same family as one with an arbitrary number of hidden nodes? It is clearly not *exactly* the same method. Comparative evaluations using the two models may well yield very different classification results. It is this sort of phenomenon which explains why the comparative performance literature contains many different results for “the same” methods applied to given public data sets. Can one then draw general conclusions about the effectiveness of the method of neural networks? Furthermore, to what extent is preprocessing the data to be regarded as part of the method? Linear discriminant analysis on raw data may yield very different results from the same model applied to data which has been processed to remove skewness. Is, then, linear discriminant analysis good or bad on these data? Likewise, is a data set in which missing values have been replaced by imputed values the same as a data set in which incomplete records have

been dropped? Applying the same method to the two variants of the data is likely to yield different results.

We have already commented that the “accuracy” of a classification rule can be measured in a wide variety of ways, and that different measures are likely to yield different performance rankings of classifiers.

Given all of the above points, it is not surprising that different authors have drawn different conclusions about the relative accuracy of different classifiers. Other commentators have taken things even further. In the discussion that accompanies [3], Efron suggested that new methods always look better than older ones and that complicated methods are harder to criticize than simpler ones. He also noted that it is difficult to make fair comparisons by making the same effort in applying different methods—a point made above. Hoadley, in the same discussion, “coined a phrase called the ‘ping-pong theorem.’ This theorem says that if we revealed to Professor Breiman the performance of our best model and gave him our data, then he could develop an algorithmic model using random forests, which would outperform our model. But if he revealed to us the performance of his model, then we could develop a segmented scorecard, which would outperform his model.”

With so many difficulties in ranking and comparing classifiers, one might naturally have reservations about small differences in performance—of the kind generally asserted for the more complicated and sophisticated methods over the older and simpler models.

6. CONCLUSION

In Section 2 we demonstrated that, when building predictive models of increasing complexity, the marginal gain from complicated models is typically small compared to the predictive power of the simple models. In many cases, the simple models accounted for over 90% of the predictive power that could be achieved by “the best” model we could find. Now, in the idealized classical supervised classification paradigm, certain assumptions are implicit: it is assumed that the distributions from which the design points and the new points are drawn are the same, that the classes are well defined and the definitions will not change, that the costs of different kinds of misclassification are known accurately, and so on. In real applications, however, these additional assumptions will often not hold. This means that apparent small (laboratory) gains in performance might not be realized in practice—they may well be swamped by uncertainties arising from mismatches between the

apparent problem and the real problem. In particular, many of the comparative studies in the literature are based on brief descriptions of data sets, containing no background information at all on such possible additional sources of variation due to breakdown of implicit assumptions of the kind illustrated above. This must cast doubt on the validity of their conclusions. In general, it means that deeper critical assessment of the context of the problem and data should be made if useful practical conclusions are to be drawn. If enough is known about likely additional sources of variability, beyond the classical sources of sampling variability and model uncertainty, then more sophisticated models can be built. However, if insufficient information is known about these additional sources, which we speculate will very often be the case, then the principle of parsimony suggests that it is better to stick to simple models.

We should note, parenthetically, that there are also other reasons to favor simple models. Interpretability, in particular, is often an important requirement of a classification rule. Indeed, sometimes it is even a legal requirement (e.g., in credit scoring). This leads us to the observation that what one regards as “simple” may vary from user to user: some might favor weighted sums of predictor values, others might prefer (small) tree structures and yet others might regard nearest neighbor methods as being simple.

Perhaps it is appropriate to conclude with the comment that, by arguing that simple models are often more appropriate than complex ones and that the claims of superior performance of the more complex models may be misleading, I am not suggesting that no major advances in classification methods will ever be made. Such a claim would be absurd in the face of developments such as the bootstrap and other resampling approaches, which have led to significant advances in classification and other statistical models. All I am saying is that much of the purported advance may well be illusory. Furthermore, although (almost by definition) one cannot predict where the next step-change will come from, one might venture a guess as to its general area. Resampling methods are children of the computer revolution, as indeed are most other recent developments in classifier technology [e.g., classification trees, neural networks, support vector machines, random forests, multivariate adaptive regression splines (MARS) and practical Bayesian methods]. Since progress in computer hardware is continuing, one might reasonably expect that the advances will arise from more powerful data storage and processing ability.

ACKNOWLEDGMENTS

I have given several presentations based on the ideas in this paper. An earlier version of this paper was presented at the 2004 Conference of the International Federation of Classification Societies in Chicago and appeared in the proceedings [18]. I would like to thank all those who commented on the material. In particular, I am grateful to Svante Wolde, Willi Sauerbrey, Foster Provost, Jerome Friedman and Leo Breiman. Of course, merely because they had valuable and interesting things to say about the ideas does not necessarily mean they agree with them.

REFERENCES

- [1] ADAMS, N. M. and HAND, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition* **32** 1139–1147.
- [2] BENTON, T. C. (2002). Theoretical and empirical models. Ph.D. dissertation, Dept. Mathematics, Imperial College London.
- [3] BREIMAN, L. (2001). Statistical modeling: The two cultures (with discussion). *Statist. Sci.* **16** 199–231.
- [4] DUIN, R. P. W. (1996). A note on comparing classifiers. *Pattern Recognition Letters* **17** 529–536.
- [5] EFRON, B. (2001). Comment on “Statistical modeling: The two cultures,” by L. Breiman. *Statist. Sci.* **16** 218–219.
- [6] FAWCETT, T. and PROVOST, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery* **1** 291–316.
- [7] FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.
- [8] FRIEDMAN, C. P. and WYATT, J. C. (1997). *Evaluation Methods in Medical Informatics*. Springer, New York.
- [9] FRIEDMAN, J. H. (1997). On bias, variance, 0/1 loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery* **1** 55–77.
- [10] GALLAGHER, J. C., HEDLUND, L. R., STONER, S. and MEEGER, C. (1988). Vertebral morphometry: Normative data. *Bone and Mineral* **4** 189–196.
- [11] HAND, D. J. (1981). *Discrimination and Classification*. Wiley, Chichester.
- [12] HAND, D. J. (1996). Classification and computers: Shifting the focus. In *COMPSTAT-96: Proceedings in Computational Statistics* (A. Prat, ed.) 77–88. Physica, Berlin.
- [13] HAND, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester.
- [14] HAND, D. J. (1998). Strategy, methods, and solving the right problems. *Comput. Statist.* **13** 5–14.
- [15] HAND, D. J. (1999). Intelligent data analysis and deep understanding. In *Causal Models and Intelligent Data Management* (A. Gammerman, ed.) 67–80. Springer, Berlin.
- [16] HAND, D. J. (2001). Modelling consumer credit risk. *IMA J. Management Mathematics* **12** 139–155.
- [17] HAND, D. J. (2001). Reject inference in credit operations. In *Handbook of Credit Scoring* (E. Mays, ed.) 225–240. Glenlake, Chicago.

- [18] HAND, D. J. (2004). Academic obsessions and classification realities: Ignoring practicalities in supervised classification. In *Classification, Clustering and Data Mining Applications* (D. Banks, L. House, F. R. McMorris, P. Arabie and W. Gaul, eds.) 209–232. Springer, Berlin.
- [19] HAND, D. J. (2005). Supervised classification and tunnel vision. *Applied Stochastic Models in Business and Industry* **21** 97–109.
- [20] HAND, D. J. and HENLEY, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *J. Roy. Statist. Soc. Ser. A* **160** 523–541.
- [21] HOADLEY, B. (2001). Comment on “Statistical modeling: The two cultures,” by L. Breiman. *Statist. Sci.* **16** 220–224.
- [22] HOLTE, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning* **11** 63–90.
- [23] JAMAIN, A. (2004). Meta-analysis of classification methods. Ph.D. dissertation, Dept. Mathematics, Imperial College London.
- [24] JAMAIN, A. and HAND, D. J. (2005). Mining supervised classification performance studies: A meta-analytic investigation. Technical report, Dept. Mathematics, Imperial College London.
- [25] KELLY, M. G. and HAND, D. J. (1999). Credit scoring with uncertain class definitions. *IMA J. Mathematics Management* **10** 331–345.
- [26] KELLY, M. G., HAND, D. J. and ADAMS, N. M. (1998). Defining the goals to optimise data mining performance. In *Proc. Fourth International Conference on Knowledge Discovery and Data Mining* (R. Agrawal, P. Stolorz and G. Piatetsky-Shapiro, eds.) 234–238. AAAI Press, Menlo Park, CA.
- [27] KELLY, M. G., HAND, D. J. and ADAMS, N. M. (1999). The impact of changing populations on classifier performance. In *Proc. Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (S. Chaudhuri and D. Madigan, eds.) 367–371. ACM, New York.
- [28] KELLY, M. G., HAND, D. J. and ADAMS, N. M. (1999). Supervised classification problems: How to be both judge and jury. In *Advances in Intelligent Data Analysis. Lecture Notes in Comput. Sci.* **1642** 235–244. Springer, Berlin.
- [29] KLINKENBERG, R. and THORSTEN, J. (2000). Detecting concept drift with support vector machines. In *Proc. 17th International Conference on Machine Learning* (P. Langley, ed.) 487–494. Morgan Kaufmann, San Francisco.
- [30] LANE, T. and BRODLEY, C. E. (1998). Approaches to online learning and concept drift for user identification in computer security. In *Proc. Fourth International Conference on Knowledge Discovery and Data Mining* (R. Agrawal, P. Stolorz and G. Piatetsky-Shapiro, eds.) 259–263. AAAI Press, Menlo Park, CA.
- [31] LEWIS, E. M. (1990). *An Introduction to Credit Scoring*. Athena, San Rafael, CA.
- [32] LI, H. G. and HAND, D. J. (2002). Direct versus indirect credit scoring classifications. *J. Operational Research Society* **53** 647–654.
- [33] MCLACHLAN, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.
- [34] MINGERS, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning* **4** 227–243.
- [35] NEWMAN, D. J., HETTICH, S., BLAKE, C. L. and MERZ, C. J. (1998). UCI repository of machine learning databases. Dept. Information and Computer Sciences, Univ. California, Irvine. Available at www.ics.uci.edu/~mllearn/MLRepository.html.
- [36] PROVOST, F. and FAWCETT, T. (2001). Robust classification for imprecise environments. *Machine Learning* **42** 203–231.
- [37] RENDELL, A. L. and SESHU, R. (1990). Learning hard concepts through constructive induction: Framework and rationale. *Computational Intelligence* **6** 247–270.
- [38] RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press.
- [39] ROSENBERG, E. and GLEIT, A. (1994). Quantitative methods in credit management: A survey. *Oper. Res.* **42** 589–613.
- [40] SALZBERG, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery* **1** 317–328.
- [41] SHAVLIK, J., MOONEY, R. J. and TOWELL, G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning* **6** 111–143.
- [42] THOMAS, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International J. Forecasting* **16** 149–172.
- [43] VON WINTERFELDT, D. and EDWARDS, W. (1982). Costs and payoffs in perceptual research. *Psychological Bulletin* **91** 609–622.
- [44] WEBB, A. R. (2002). *Statistical Pattern Recognition*, 2nd ed. Wiley, Chichester.
- [45] WEISS, S. M., GALEN, R. S. and TADEPALLI, P. V. (1990). Maximizing the predictive value of production rules. *Artificial Intelligence* **45** 47–71.
- [46] WIDMER, G. and KUBAT, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning* **23** 69–101.
- [47] ZAHAVI, J. and LEVIN, N. (1997). Issues and problems in applying neural computing to target marketing. *J. Direct Marketing* **11**(4) 63–75.

Comment: Classifier Technology and the Illusion of Progress

Jerome H. Friedman

This paper provides a valuable service by asking us to reflect on recent developments in classification methodology to ascertain how far we have progressed and what remains to be done. The suggestion in the paper is that the field has advanced very little over the past ten or so years in spite of all of the excitement to the contrary.

It is of course natural to become overenthusiastic about new methods. Academic disciplines are as susceptible to fads as any other endeavor. Statistics and machine learning are not exempt from this phenomenon. Often a new method is heavily championed by its developer(s) as the “magic bullet” that renders past methodology obsolete. Sometimes these arguments are accompanied by nontechnical metaphors such as brain biology, natural selection and human reasoning. The developers become gurus of a movement that eventually attracts disciples who in turn spread the word that a new dawn has emerged. All of this enthusiasm is infectious and the new method is adopted by practitioners who often uncritically assume that they are realizing benefits not afforded by previous methodology. Eventually realism sets in as the limitations of the newer methods emerge and they are placed in proper perspective.

Such realism is often not immediately welcomed. Suggesting that an exciting new method may not bring as great an improvement as initially envisioned or that it may simply be a variation of existing methodology expressed in new vocabulary often elicits a strong reaction. Thus, the messengers who bring this news tend to be, at least initially, unpopular among their colleagues in the field. It therefore takes courage to provide this type of service, and Professor Hand is to be congratulated for this thoughtful article.

Of course, simply because new methodologies are often overhyped does not necessarily imply that they do not, at least sometimes, represent important progress.

Jerome H. Friedman is Professor, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, California 94305, USA (e-mail: jhf@stanford.edu).

In the case of classification, I believe that there have been major developments over the past ten years that have substantially advanced the field, both in terms of theory and practice. Although I find myself in agreement with most of the premises of this article, I do not see how they lead to the implication that such advances are “largely illusionary.”

There appear to be three main premises presented in the article. First, the improvements realized by the newer methods over the previous ones are less than those achieved by the previous ones over their predecessors, presumably no methodology at all. Second, the evidence often presented (at least initially) in favor of the superiority of the newer methods is often suspect. Finally, the newer methods do not solve all of the outstanding important problems that remain in the field of classification. In my view these observations are correct and underappreciated in the field. The article does an important service by illustrating them so forcefully. However, the truth of these assertions does not imply lack of important progress; only that low-lying fruit is often easier to gather, we should be more thorough concerning validation when initially presenting new procedures and there is still important work to be done.

One of the main assertions in the paper is that, in many applications, older methods often yield error rates comparable to the more modern ones. This is of course true and is intrinsic to the classification problem, especially when the metric used to measure performance is based on error rate. First, there is the irreducible error caused by the fact that the predictor variables \mathbf{x} often do not contain enough information to specify a unique value for the outcome variable y . At best, they specify a probability distribution of possible values $\Pr(y|\mathbf{x})$ which is hopefully different for differing values of \mathbf{x} , indicating some predictive power. This phenomenon afflicts all prediction problems. A second phenomenon is peculiar to classification; it is not necessary to accurately estimate $\Pr(y|\mathbf{x})$ to achieve minimal error rate. All that is required of the estimates $\hat{\Pr}(y|\mathbf{x})$ is

$$(1) \quad \arg \max_y \hat{\Pr}(y|\mathbf{x}) = \arg \max_y \Pr(y|\mathbf{x}).$$

The actual values of the estimates for differing values of y need not be close to their respective underlying true values. The estimates for the nonmaximizing probabilities need not even be in the correct order. Thus, more flexible (modern) procedures that are better able to estimate more complex probability structures need not produce dramatically lower error rates in many applications. This also accounts for the “flat minimum” effect discussed in the paper.

As pointed out in the paper, classification procedures are often used in contexts where error rate is not the relevant quantity; functionals of $\Pr(y|\mathbf{x})$ other than (1) are of interest. For example, in many two-class classification problems $y \in \{-1, 1\}$, the important quantity is the rank order of $\{\Pr(y = 1|\mathbf{x}_i)\}_{i \in T}$, where T is a set of observations with unknown outcome. In other applications, interest is in the actual probabilities themselves. In such settings it is likely that more accurate estimates of $\Pr(y|\mathbf{x})$ afforded by more flexible modern techniques will yield distinctly superior results to the older less flexible methods, even though their respective error rates are not dramatically different. The paper properly criticizes the classification literature for presenting comparisons mostly in terms of error rate, even though this is the criterion used for nearly all of the classification comparisons presented in the paper.

The primary evidence intended to suggest lack of progress is the comparisons presented in Table 1. Here the error rate of an older method, linear discriminant analysis (LDA), is compared with that of the current best method for each of a selected set of problems. In spite of the general insensitivity of error rate as differentiating criterion (as discussed above), LDA seems to produce distinctly inferior results in many of these problems. In more than half of the examples, its error rate is at least 45% greater; in one example, it is nearly six times as great. Of course there is a selection bias of unknown magnitude in choosing the best method, but it is difficult to conclude from the evidence presented that LDA is competitive with the best current methods, even in terms of error rate. The paper suggests that large ratios in small error rates “will correspond to only a small proportion of new data points.” This is true but not relevant. If a zip code classifier makes twice as many errors, it costs the post office twice as much to handle the misdirected mail. I have yet to see a problem where costs are proportional to the Prop linear statistic shown in the last column of Table 1.

The paper presents a regression example (Section 2.1) to illustrate that including additional predictor variables that are highly correlated with those that are

already part of the analysis produces little gain in performance. This is true of all methods, old and new, and no evidence is provided to suggest that older methods are better able to incorporate additional information from such variables.

A second principal premise of the paper is that the evidence for the superiority of new methods is generally based on empirical comparisons which are susceptible to major weaknesses that place their validity in question. I could not be in more agreement with this point. Section 5 of the paper should be required reading for all practitioners and researchers in the field. In my data mining course, I have a lecture called “comparison caution” that addresses many of the same issues. Empirical comparisons should be viewed with skepticism, especially when the authors’ new method is one of the competitors. Even when this is not the case, the authors performing the study often have a favorite technique which usually emerges as the top performer. When interpreting such studies, I tend to ignore the apparent top performer and look at the relative rankings of the other methods, presuming that the authors have less expertise and vested interest in them. Even when a comparison is free of all of the biases discussed in Section 5, its results should not be extrapolated beyond the specifics of the problem represented by the data set being used. All methods have particular problems for which they are especially well suited and others for which they are not. Sometimes only a minor change in the problem setup can produce substantial changes in performance rankings. Results of empirical comparisons can be useful, especially when aggregated over time, but the natural tendency to overinterpret individual studies should be avoided. Of course, the same caution should be applied to the empirical comparisons presented in this paper.

Simply because the initial evidence for the superiority of a method can be questioned does not necessarily imply that is not useful or that it does not represent progress. Practitioners try various methods and, as time evolves, some emerge as being more useful than others. Many of the “new” proposals of the distant past have not survived the test of time and are now long forgotten. Those that have emerged as being generally useful, such as logistic regression, LDA and decision trees, have survived to see common use. No one is claiming that all of the new techniques proposed in the literature over past ten years represent major advances. However, I believe that a body of evidence is emerging that suggests that some of them, such as the ensemble methods (bagging and boosting) and support vector machines, offer substantial advantages over the earlier methods

in enough situations to be regarded as major advances. This is especially the case in scientific and engineering applications, where decision boundaries are often complex and far from being linear.

Another major premise of the paper is that there are important issues that affect classification performance that are not addressed by most modern methodology. These include population drift, sample selectivity bias, errors in class labels and arbitrariness in class definitions. Again I could not agree more. Issues of non-representative training data tend to be overlooked by the academic community, although they are probably well known to most practitioners. (See [3]. I spend several lectures in my data mining course covering these topics.) Obtaining high-quality representative training data is generally more important to success than choice of a particular classifier, although given such data, choosing the best classifier can often provide considerable additional benefit. In many data mining applications, the data were collected for a different purpose than solving the current problem and one does not have influence over its quality or value. The analyst is forced to do the best that can be done with the data at hand.

The problem of training data being different from future data to be predicted is common to all prediction, not just classification. The fundamental issue is similar whether the differences arise through random sampling from a static population or are caused by one of the more deterministic mechanisms cited in the paper. As noted in the paper, the antidote is to limit reliance on the training data by not fitting it too closely. This is the basic principal underlying regularization. The paper argues that older methods are “simpler,” thereby inducing more regularization, which in turn causes them to be more resistant to these types of problems. This need not be the case.

Almost all of the modern procedures incorporate a regularization parameter that controls the degree to which they are allowed to fit the training data. By adjusting the value of this parameter, one can produce a sequence of models of increasing complexity from the very simplest that makes the same prediction everywhere to highly complex functions that capture the fine details of the predictive relationship as reflected in the training data. Highly regularized versions of different procedures may capture somewhat different aspects of the gross features of the probability distribution, but in the absence of knowledge concerning the nature of the population drift, there is no a priori reason to suspect that one is better than the other. An important consequence of the presence of population drift

and related problems is that model selection based on traditional techniques such as bootstrapping or cross-validation becomes overly optimistic; they will tend to produce insufficient regularization. Thus, care must be taken to regularize more heavily than suggested by these model selection techniques when such problems are suspected.

Most older classification methods limit the degree to which one can control the amount of regularization. It is not clear that the amount arbitrarily applied by these procedures is necessarily appropriate in any particular problem. In fact there are many situations in which older methods provide insufficient regularization. This is especially the case in modern analytical chemistry and bioinformatics applications, where there are many more predictor variables than training observations, and simple logistic regression and LDA completely fail. There has been considerable recent research that has led to modern classification methods that allow the application of more regularization than the older traditional methods. These, in my view, also represent major progress.

Errors in class labels is a classic robustness issue. Estimation in the presence of badly measured outcomes has been extensively studied in the regression literature, but less so in classification. As in regression, the solution is to employ loss criteria that are less sensitive to individual extreme measurements. It has been suggested that logistic likelihood and the support vector machine hinge loss are more robust to misspecification of class labels than squared-error loss or, especially, the exponential loss associated with AdaBoost, since they weight realized outcomes of low estimated probability less heavily. Even more robust (nonconvex) loss criteria have been proposed for classification (see [1, 2]). Some older methods such as logistic regression should be fairly robust to mislabeling, but others like LDA are likely to exhibit poor robustness properties; estimates of the pooled covariance matrix can be highly distorted by only a few mislabeled observations, especially at the extremes of the data distribution.

The problem of arbitrariness of class labels is often caused by trying to make the problem conform to the method rather than the other way around. If an outcome variable realizes continuous numeric values, then it should be treated as such and regression rather than classification technology would be more appropriate. There have been recent important advances in regression methodology that parallel those in classification. If thresholding numeric variables to create a classification problem happens to be appropriate and the class

labels have changed, then, as the paper suggests, one can simply retrain the classifier with the new definitions. This requires that the original raw data be saved. Given the very low cost of storage media, this should always be encouraged for a wide variety of reasons.

Recent research has not solved all of the outstanding problems in the field of classification, especially those associated with nonrepresentative training data. All procedures are vulnerable to these effects and, as discussed above, it is not clear that the older methods enjoy more immunity than the more recent ones. Also, these problems are more prevalent in the commercial sector involving financial and consumer behavior applications than in scientific and engineering fields where the laws governing the systems under study tend to be more stable. Nevertheless, solutions to these problems would also represent major advances. The paper does an important service by directing our attention to them, but this does not imply that there has not been substantial progress in other important aspects of the classification problem in the recent past.

Whether or not a new method represents important progress is, at least initially, a value judgement upon which people can agree to disagree. Initial hype can be misleading and only with the passage of time can such

controversies be resolved. It may well be too soon to draw conclusions concerning the precise value of recent developments, but to conclude that they represent very little progress is at best premature and, in my view, contrary to present evidence.

I thank Professor Hand for this thoughtfully provocative article. It gives all of us an opportunity to look past our enthusiasm and take a deeper look at the remaining central issues. I look forward to research that produces solutions to these outstanding problems and to future discussions as to whether they represent major progress. Finally, I would like to add another relevant quote to that of Eric Hoffer mentioned in the article. This one is attributed to Yogi Berra: "Prediction is difficult, especially when it's about the future."

REFERENCES

- [1] FREUND, Y. (2001). An adaptive version of the boost by majority algorithm. *Machine Learning* **43** 293–318.
- [2] FRIEDMAN, J. H. and POPESCU, B. E. (2004). Gradient directed regularization. Technical report, Dept. Statistics, Stanford Univ.
- [3] TWO CROWS (1999). *Introduction to Data Mining and Knowledge Discovery*, 3rd ed. Available at www.twocrows.com/booklet.htm.

Comment: Classifier Technology and the Illusion of Progress—Credit Scoring

Ross W. Gayler

These comments support Hand's argument for the lack of practical progress in classifier technology by pursuing them a little deeper in the specific context of credit scoring. Academic development of modeling techniques tends to ignore the role of the practitioner and the impact of business objectives. In credit scoring it can be seen that the nature of the task forces practitioners to adopt modeling strategies that positively favor simple techniques or, at least, limit the possible advantage of sophisticated techniques. The strategies adopted by credit scorers can be viewed as a heuristic approach to inference of the unobserved (and unobservable) distribution of possible data sets. The technical progress examined by Hand has been aimed toward better goodness of fit. However, technical progress toward a more principled basis for inferring the distribution of future problem data would be more likely to be adopted in practice.

1. CREDIT SCORING

I am approaching this commentary as a domain-specific consumer of statistical technology. My concern is credit scoring (the use of predictive statistical models to control operational decision-making in consumer finance). Classical credit scoring is applied at the point of application for a loan to predict the risk of default (nonpayment) and to make the decision whether to approve that application for credit. The total value of the loans made under the control of credit scoring is immense, and the value added to the economy by better decision-making because of credit scoring is correspondingly large. Thus, credit scoring is a domain where improved decision-making due to better predictive modeling would be valuable and technical progress would be expected.

Ross W. Gayler is Honorary Associate, School of Communication, Arts and Critical Enquiry, La Trobe University, Melbourne, Australia and Senior Research and Development Consultant, Baycorp Advantage, Melbourne, Australia. Mailing address: 102 Through Road, Camberwell VIC 3124, Australia.

Somewhat surprisingly, the statistical techniques currently used in credit scoring seem rather old-fashioned (often being simple regression models). This is not for lack of attempts to change the state of the art. New modeling techniques are regularly proposed for credit scoring (typically by academic researchers), but they are rarely adopted in practice. This lack of uptake cannot be blamed entirely on conservatism in the credit scoring community. The rewards of improvement are sufficiently high that once any lender adopts a better technique, there will be high competitive pressure for other lenders to do likewise. Rather, the continued use of simple predictive modeling techniques suggests that they have a practical advantage over more sophisticated techniques in credit scoring. Understanding the reasons for this advantage would be useful for the practice of applied predictive modeling in credit scoring and, more generally, might suggest productive avenues for the development of predictive modeling techniques to be applied in practical domains.

Professor Hand has worked extensively in credit scoring and it is likely that his experience in that domain motivated the writing of his paper, although his thesis, as stated, is not restricted to credit scoring. As a practitioner of credit scoring, I agree with the points he has raised. My aim here is to examine Hand's points a little further in the specific context of credit scoring, looking at the interaction of the technicalities of modeling with the demands imposed by the nature of the business task.

A brief description of the classical credit scoring problem is as follows. When credit is granted to consumers, some of the borrowers will default on their loans. The lender typically takes a loss on a defaulted loan. Ideally, a lender would predict which applicants would default and decline their applications for credit, thus avoiding the loss. The lender uses data available at the time of application to make that prediction and decision. The data may come from an application form, a credit bureau and the lender's own records if the applicant is an existing customer.

The potential predictors available at the time of application are not causally related to the outcome of

default. Consequently, credit scoring models are correlative rather than causal. The outcome of default is not just dependent on the characteristics of the borrower, but also on external factors such as subsequent lender management actions and the state of the economy. Furthermore, the data are processed by the operational systems of lenders. These systems are constructed with the primary objective of carrying out the operational actions. Data collection and data quality issues that are relevant to statistical modeling are often an afterthought in system design (if they are considered at all). Consequently, the data quality is often not what would be desired, and data quality problems can be quite dynamic, because changes are made to the systems to accommodate short term operational needs. The data are noisy, and the quality of the noise is subject to drifts and jumps.

2. REGRESSION RATHER THAN CLASSIFICATION

Given that the occurrence of default is a binary outcome, it seems natural to treat credit scoring as a classification problem, and many academic papers have done so. Assuming a classification framework comes close to assuming that there is some ideal predictor space in which the outcome classes are perfectly separated. Even if such a predictor space does actually exist, it is not available to the credit scoring practitioner. The available predictors are not causally related to the outcome and some predictors (e.g., account management actions and changes in the economy) are not available at the time of the application because they occur subsequently. For problems such as this, as Hand notes more generally, “the Bayes error rate is high: meaning that no decision surface can separate the distributions of such problems very well” (Section 2.3). Given that the outcome classes cannot be separated, it may be better to adopt a regression framework for modeling and predict the probability of default conditional on the predictors.

However, in credit scoring there is an even more important consideration than the match between the theoretical form of the model and the true state of affairs. Lenders need to be able to control the rate at which loan applications are declined. This allows them to adjust workloads and to control the trade-off of profit against volume of business. A classification model yields predictions of “default” or “repay” which are mapped to decisions to “decline” or “accept” the loan application. Consequently, the decline rate is fixed by the predictions and the lender has no direct control of the decline rate from a classification model. This illustrates the

point that credit scoring practitioners need to be mindful of the operational requirements of lending over and above goodness of fit and the theoretical form of models.

Hand’s paper is written in terms of classifiers, but his arguments apply just as well to regression models used as classifiers. A regression model may be trivially converted to a classifier by having the predicted outcome be the probability of class membership and comparing it to a threshold. In fact, this is the standard form of credit scoring models. Conversely, some classification models can be converted to adequate regression models, but this is not generally true. A decision tree with two leaves will never make a good regression model. Consequently, even though classification models are not well suited to credit scoring, Hand’s arguments do apply to credit scoring as it is practiced.

3. EQUIVALENCE OF MODELS AND DEGREES OF FREEDOM IN THE MODELER

Hand observed that “a tremendous variety of algorithms and models has been developed for the construction of such [classification] rules” (Section 1). Different algorithms have different representational biases and a different bias/variance trade-off. For a fixed set of predictors we would expect different algorithms to generate different approximations to the outcome. However, in credit scoring the set of predictors is not fixed. The model developer is free to generate new derived variables in the data set and will generally do so to accommodate the particular representational bias of the modeling technique used. For example, decision tree induction and projection pursuit regression are able to automatically model interactions in the data, whereas regression works only with the predictors it is given and does not create interactive combinations. The credit scoring modeler using regression would construct interaction predictors if they were thought necessary.

The objective of every modeling technique is to approximate the data. Thus, in the limit (and the hands of a skilled modeler), every modeling technique should end up in agreement because they are all approximating the same data. However, the effort required to achieve that degree of approximation may vary greatly between techniques. Even for techniques that require the same effort to achieve a given accuracy of approximation, the models may differ in other properties that are operationally important to the lender.

It is also worth recalling Hand’s comment about the high Bayes error rate (Section 2.3). When the ratio of

variance accounted for by the response surface is low compared to the error about the response surface (as it is in credit scoring), it becomes harder to distinguish between different representational biases. Thus we would not expect the differences between different modeling techniques to be readily observable.

The impact of the skilled modeler warrants some further investigation. Effectively, the modeler supplies extra degrees of freedom in addition to those supplied by the modeling technique. The natural consequence of this is to reduce the difference between techniques in terms of goodness of fit. Rather than compare modeling techniques in terms of predictive power, it would be more useful to look at the effort required of the modeler to achieve a given goodness of fit and other properties of the models that are of operational relevance to the lender.

4. MAIN EFFECTS AND INTERACTIONS

In Section 2.3, Hand mentions “examples of artificial data which simple models cannot separate (e.g., intertwined spirals or checkerboard patterns),” noting that “such data sets are exceedingly rare in real life [and] it is common to find that the centroids of the predictor variable distributions of the classes are different.” This is a claim that problems which can be modeled only as interactions of the variables (with no observable main effects) are rare. This may well be true in general because of the improbability of interactions exactly canceling out to leave no main effects. However, in credit scoring it is also true for domain-specific reasons. The inclusion of each predictor in a decision-making system has to be justified (operationally and legally). It is much easier to argue for the inclusion of a predictor if the argument can be made for that predictor in isolation. Conversely, it is harder to argue for the inclusion of a predictor if it can be shown to add value only in the context of other predictors.

Furthermore, credit scoring practitioners are very concerned with the stability over time of their models. Some credit scoring models are used for years before being replaced. Therefore, it is important to ensure that the predictive relationships on which the model is based are stable over time. Credit scoring practitioners tend to believe that main effects are more stable than interactions (all other things being equal). When interactions are included as predictors, it is generally because the modeler has a prior belief that the interaction reflects some stable mechanism in the world. An otherwise unmotivated interaction that is discovered by an

automated search procedure is unlikely to be included in a predictive model or, if it is included, to have its influence intentionally limited relative to the main effects. The effect of these selection biases is to ensure that credit scorers prefer simpler models based on main effects.

5. SENSITIVITY TO ARBITRARY MODELING DECISIONS

Hand notes that when constructing classification rules, “various . . . assumptions and choices are often made which may not be appropriate” (Section 1) and even when they are entirely appropriate, the choices may be somewhat arbitrary. He gives the example of typically defining “a customer as ‘defaulting’ if they fall three months in arrears with repayments . . . [while] [i]t is entirely reasonable that alternative definitions (e.g., four months in arrears) might be more useful if economic conditions were to change” (Section 4.2). Credit scoring necessarily involves many detailed decisions concerning the modeling process. Many of these decisions involve compromises and trade-offs, with no obviously correct answer. While the experienced credit scorer would have arguments for the specific decisions made, it would be a bold modeler who would argue that the decisions taken were uniquely and obviously correct. Thus, there is an element of arbitrariness in the modeling process.

It is possible to conceive of a space of feasible modeling decisions. Similar sets of decisions are nearby in that space. A small change in the modeling decisions would generally lead to a small change in the models. However, the possibility exists that a small change in modeling decisions may lead to a large change in the models that arise from them. This would be very unsatisfactory in credit scoring because the results of the modeling would be strongly dependent on arbitrary modeling choices. Therefore, credit scorers tend to restrict their attention to regions of the modeling decision space where the gradient of models with respect to modeling decisions is low. In these regions, all the models generated as a result of the different modeling choices would yield similar results. If a new modeling technique yielded markedly different results, it would be unlikely to be favored by credit scorers unless it was surrounded by a region of other models yielding similar results. It would be more difficult for the modeler to argue for the correctness of the unique results given that the choice of modeling technique might be regarded as arbitrary.

6. DEVELOPMENT DATA NOT REPRESENTATIVE OF OPERATION

Hand points out “that in many . . . real classification problems the data points in the design set are not . . . randomly drawn from the same distribution as the data points to which the classifier will be applied” (Section 1). Furthermore, any design set represents “merely a single . . . problem drawn from a notional distribution of problems” (Section 1). Later he notes that “a fundamental assumption of the classical paradigm is that the various distributions involved do not change over time . . . [although this assumption] is unrealistic in most commercial applications, concerned with human behaviour” (Section 3.1). His concern here is with population drift. This would not be a problem if the predictive model were the “true” model, but as Hand states “it would be a brave person who could confidently assert that [this] held” (Section 3.2).

Population drift is a particular concern in credit scoring. Loans which default do so over an extended period after the loan has been granted. Consequently, an extended outcome period (typically at least one year) is required to allow a reasonable proportion of loans to default. To this must be added time to accumulate enough applications to provide a reasonable number of observations for modeling and to allow for seasonal variation in the applicant population. Allowing time for data preparation, data modeling and implementation of the models into the operational system, it is common for the oldest data on which a model is based to be three years old when the model is first switched on. Then the model may be in use for some while (three years is common, and more than five years not unknown). Even if the applicant population distribution is stationary, the data collecting process is subject to random jumps, because lenders may change their systems and procedures at any time. Thus, a large part of the value added by credit scoring practitioners comes from anticipating possible future shifts in the data distribution and designing the models to be relatively insensitive to such shifts. This can be seen as another aspect of attempting to reduce the sensitivity of the models to arbitrary features of the specific design set (in this case, characteristics of the data that just happen to hold at the time the data are collected).

The expertise of the credit scoring modeler can be thought of as applying a bias to the modeling techniques to move the models toward the notional distribution of problems. For example, Hand discusses the application of a tree model and linear discriminant

analysis (as competing techniques) to consumer credit data, and points out that because the design set is always retrospective, the population may have drifted by the time the model is built and “reduced any advantage that the more sophisticated tree model may have” (Section 3.1). A tree model fits better than linear discriminant analysis, but degrades more rapidly. There is the possibility that the tree model may actually become worse than the linear discriminant model with the passage of time. Rather than view the techniques as competing, a credit scorer might model the data with linear discriminant analysis and then build a tree model of the residuals. This hybrid model puts a bound on deterioration by predicting the majority of the outcome variance using the more stable modeling technique.

7. FREEDOM VIA THE FLAT MAXIMUM EFFECT

Hand mentions the flat maximum effect in the context of explaining that a reasonable fraction of the maximum attainable predictive power can be obtained from an equally weighted combination of predictors (Section 2.4). The existence of the flat maximum effect is a great advantage in credit scoring. It implies that there may be many alternative models with similar goodness of fit. This provides the credit scoring modeler the opportunity to choose between those models on some basis other than goodness of fit (e.g., susceptibility to population drift or ability to finely control the decline rate). The freedom this confers is so valuable that credit scoring modelers prefer to choose predictors that make the flat maximum effect more likely to exist. This is the case where there is a conditional monotone relationship between each of the predictors and the outcome (which also happens to be the circumstances under which a simple linear combination is likely to perform well).

8. VALUE ADD AND MODELING TECHNIQUES

In credit scoring, much of the value added by modelers is not via goodness of fit to the development sample, but by anticipation of possible changes in the operational systems and data. This can be viewed as a problem of trying to infer the unobserved distribution of possible development data sets. Credit scorers attempt to achieve this by biasing their models toward simple models and techniques. These models are not only more likely to generalize across potential data sets, but also, as Hand points out, to yield most of the predictive power of more complex models. More com-

plex models of the current data set are unlikely to be attractive to credit scorers. However, techniques that

provide a more principled basis for generalizing to the distribution of possible data sets would be welcome.

Elaboration on Two Points Raised in “Classifier Technology and the Illusion of Progress”

Robert C. Holte

1. INTRODUCTION

This short note elaborates two points raised in David Hand’s target article. First, I provide additional evidence that simple classification rules should be given serious consideration in any application and that there are often diminishing returns in considering increasingly complex classifiers. Second, I refine Hand’s basic argument that small improvements in performance are irrelevant because of the uncertainty about many aspects of the situation in which the classifier will be deployed. In particular, I briefly describe a recently developed method for analyzing and comparing classifier performance when the class ratios and misclassification costs are unknown. This does not refute his general argument, but it does provide an important exception to it.

2. SIMPLICITY-FIRST METHODOLOGY AND DIMINISHING RETURNS

Hand (Section 2.3) cites my 1993 study [4] in which the accuracy of one-level decision trees, which classify examples based on the value of a single feature, was compared to the accuracy of the decision trees learned by C4.5 [8], a state-of-the-art decision tree learning algorithm. The article caused quite a stir, because nobody at the time suspected that most of C4.5’s classification accuracy could be achieved, on many of the standard test data sets, by building just the first level of the decision tree. The overall conclusion of my 1993 article is the same as Hand’s—not that the more complex decision rules should be cast aside, but that the simple decision rules should not be dismissed out of hand. One can never tell, a priori, how much of the structure in a domain can be captured by a very simple decision rule, and since simplicity is advantageous for both theoretical and practical reasons,

it is incumbent on a responsible experimentalist or practitioner to begin with the simplest decision rules. Only if they prove unacceptable should more complex decision rules be considered. I coined the term “simplicity-first methodology” to describe this systematic approach of proceeding from simple to more complex decision rules.

In a follow-up paper [1], Maass and Auer developed an efficient algorithm for constructing a decision tree of fixed depth d , with the minimal error rate on the training data, and we proved theoretical bounds on the generalization error rate of this decision tree. This empirical study showed that the performance advantage of C4.5 over one-level trees in my original study [4] greatly diminishes when depth is increased to two, with the two-level trees actually being superior to C4.5’s trees on 4 of the 15 data sets in the study.

Table 1 herein compares the accuracies achieved when $d = 0$, $d = 1$ and $d = 2$. These accuracies are averages of nine repetitions of 25-fold cross-validation on each data set. The $\Delta(1-0)$ column gives the accuracy improvement achieved by moving from a zero-level tree, which classifies all examples according to the majority class, to a one-level tree, and the $\Delta(2-1)$ column gives the accuracy improvement achieved by moving from a one-level tree to a two-level tree. Comparing these two columns, we see clear confirmation of Hand’s observation that increasing complexity produces diminishing returns on accuracy improvement in many domains.

There have been other studies that showed that simple classifiers perform well on standard test data sets. Domingos and Pazzani [2] showed that a naive Bayesian classification algorithm significantly outperformed state-of-the-art systems for decision tree learning, decision rule learning and instance-based learning in a substantial number of the 28 data sets in their study. Kohavi [5] showed that wrapper-based feature selection, combined with a majority classifier, can produce simple classifiers that are as accurate as C4.5’s trees in many cases. Linear discriminants (perceptrons)

Robert C. Holte is Professor, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, T6G 2E8 (e-mail: holte@cs.ualberta.ca).

TABLE 1
*Diminishing returns with additional complexity**

Data set	Zero-level	One-level	Two-level	$\Delta(1-0)$	$\Delta(2-1)$
BC	70.3	67.2	66.3	-3.1	-0.9
HE	79.4	79.2	78.6	-0.2	-0.6
AP	80.2	80.0	88.6	-0.2	8.6
SE	90.7	95.0	97.3	4.3	2.3
LA	64.9	71.6	86.6	6.7	15.0
PI	65.1	73.6	74.8	8.5	1.2
SP (3)	51.9	63.2	79.4	11.3	16.2
CH	52.2	66.1	86.9	13.9	20.8
IO	64.1	78.3	86.1	14.2	7.8
PR	50.0	66.3	69.3	16.3	3.0
HD	54.5	70.9	67.1	16.4	-3.8
G2	53.4	76.2	79.7	22.8	3.5
CR	55.5	85.5	84.2	30.0	-1.3
SO (4)	36.2	85.3	91.1	49.1	5.8
IR (3)	33.3	91.9	95.7	58.6	3.8

*The first column gives the acronym for the data set as in [1], with the number of classes shown in parentheses if it is different from two. The next three columns give the accuracy of the majority classifier (zero-level decision tree), one-level decision tree and two-level decision tree, respectively. The $\Delta(1-0)$ column gives the difference in accuracy between the one-level and zero-level trees, and the final column gives the difference in accuracy between the two-level and one-level trees. The rows are sorted according to $\Delta(1-0)$.

have also been seen to perform surprisingly well [6, 9].

3. EMPIRICAL COMPARISONS OF CLASSIFIERS IN UNKNOWN CIRCUMSTANCES

The fundamental argument put forward by David Hand has two parts: (1) that often only small performance gains arise from using complex classifiers and (2) that the small gains seen in the idealized laboratory setting will be swamped, in practical applications, by unpredictable and changing conditions that have a substantial effect on performance. I agree with both of these statements, in general, but I would like to point out, with regard to the latter, that we do possess methods for coping perfectly well with certain important kinds of unpredictable and changing circumstances.

Among the most important examples Hand gives of unpredictable and changing factors that affect a classifier's usefulness in practice are the costs of the different types of misclassification and the distribution of data to which the classifier will be applied. I agree entirely that in many practical settings these factors cannot be determined at the time classifiers are being evaluated and compared, and that these factors often change with time.

Drummond and I have developed a method, called cost curves, for analyzing and comparing two-class classifier performance when the misclassification costs and the relative frequency of the two classes are unknown [3]. The key idea is to plot performance (expected cost, normalized to be between 0 and 1) as a function of these unknowns. It turns out that, for the case of expected cost, these unknowns can be combined into a single aggregate unknown that also varies between 0 and 1. Cost curves therefore are a two-dimensional plot, with performance (normalized expected cost) as the y-axis and the aggregate unknown, which we call $PC(+)$, as the x-axis.

The cost curve for a given classifier is a straight line that depicts its performance across all possible combinations of misclassification costs and class ratios. Empirical confidence intervals can be computed for cost curves and for differences between cost curves, allowing one to answer the all-important question, "Under what circumstances does classifier *A* significantly outperform classifier *B*?" A software tool that fully supports cost curve analysis is available upon request.

Figure 1 herein shows the cost curves for two classifiers on the Japanese credit screening data from the UCI repository [7]. The solid line is the cost curve for C4.5's decision tree on this data set and the dashed line is the cost curve for the one-level decision tree produced by my 1R system [4]. We can see that these two

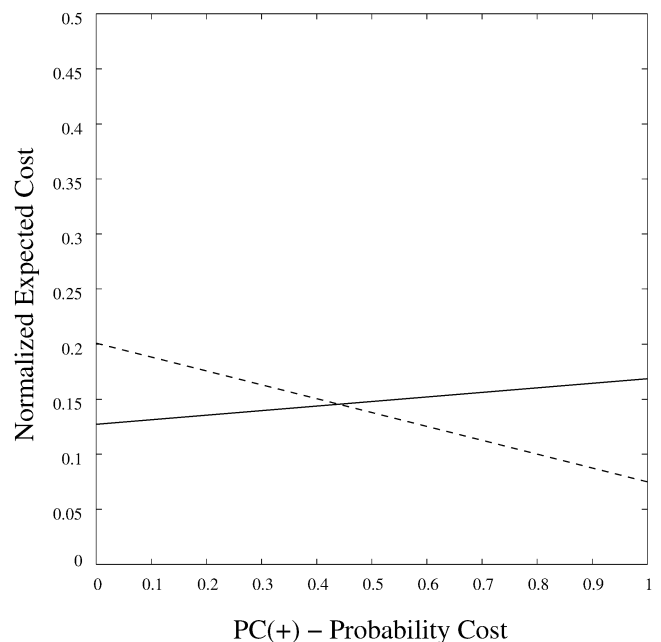


FIG. 1. Cost curves for C4.5 (solid line) and 1R (dashed line) on the Japanese credit screening data set.

classifiers have identical performance when $PC(+)$ has a value of roughly 0.45, that the one-level tree has a lower expected cost than C4.5's decision tree for larger values of $PC(+)$ and that C4.5's tree outperforms the one-level tree for smaller values.

My aim here is not to give a tutorial on cost curves, but to point out that there are sound, practical ways to cope with some of the factors that Hand correctly identifies as often being unknown, or subject to change, at the time of classifier evaluation. Cost curves provide a concrete example of how we can do classifier evaluation and comparison perfectly well without any knowledge about misclassification costs or the class ratios. By considering all possible combinations of the unknown factors, exact analysis and comparison is possible, and small performance differences can be significant. However, this does not refute Hand's general point. There are other factors and kinds of changes, such as shifting distributions within a class [10], that we do not yet know how to cope with—a challenge for future research.

ACKNOWLEDGMENTS

I would like to thank the Natural Sciences and Engineering Research Council of Canada for financial support, and Alberta Ingenuity for funding the Alberta Ingenuity Centre for Machine Learning. All my work on cost curves is joint with Chris Drummond of the Institute for Information Technology (Ottawa) of the Canadian National Research Council.

REFERENCES

- [1] AUER, P., HOLTE, R. C. and MAASS, W. (1995). Theory and applications of agnostic PAC-learning with small decision trees. In *Proc. Twelfth International Conference on Machine Learning* 21–29. Morgan Kaufmann, San Francisco.
- [2] DOMINGOS, P. and PAZZANI, M. (1997). On the optimality of the simple Bayesian classifier under zero–one loss. *Machine Learning* **29** 103–130.
- [3] DRUMMOND, C. and HOLTE, R. C. (2000). Explicitly representing expected cost: An alternative to ROC representation. In *Proc. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 198–207. ACM Press, New York.
- [4] HOLTE, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning* **11** 63–90.
- [5] KOHAVI, R. (1995). The power of decision tables. In *Proc. Eighth European Conference on Machine Learning. Lecture Notes in Artificial Intelligence* **912** 174–189. Springer, Berlin.
- [6] MICHIE, D., SPIEGELHALTER, D. J. and TAYLOR, C. C., eds. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.
- [7] NEWMAN, D. J., HETTICH, S., BLAKE, C. L. and MERZ, C. J. (1998). UCI repository of machine learning databases. Dept. Information and Computer Sciences, Univ. California, Irvine. Available at www.ics.uci.edu/~mllearn/MLRepository.html.
- [8] QUINLAN, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
- [9] SHAVLIK, J., MOONEY, R. J. and TOWELL, G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning* **6** 111–143.
- [10] WEBB, G. and TING, K. M. (2005). On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning* **58** 25–32.

Comment: Classifier Technology and the Illusion of Progress

Robert A. Stine

It is my pleasure to contribute to the discussion of this paper. David Hand has the credibility one needs to write such an article and not have it dismissed out of Hand. Along with publishing numerous papers and books on classification and data mining, he “works in the trenches” with real data. His contributions to credit modeling are particularly well known and respected, and his knowledge of that domain reaches far deeper into the substance than the casual illustration often chosen to show off a new methodology. He is a fascinating lecturer and I have learned a great deal by listening carefully to his ideas. When he writes that claims of the superiority of neural networks and support vector machines “fail to take account of important aspects of real problems,” I have to stop and think about my own research and experiences.

The thrust of Hand’s paper is the argument that most recent developments in classification, say anything since Fisher’s linear discriminant function, offer little benefit in practice. The mismatch between theory and practice dwarfs incremental claims for superiority established in theorems. For instance, theory that shows that a support vector machine classifies better than a simple linear model is an “illusion,” bordering on sophistry.

I have a great deal of sympathy for this point of view, but I doubt that many statisticians will change what they do after reading this paper. I agree with many of his criticisms, but I am already in the choir. I suspect that it will take quite a bit more to convince others, particularly along the lines of proposals for what ought to be done. Consider the impact of Tukey’s “The future of data analysis” (Tukey, 1962). After chastising the field for its preoccupation with “optimization in terms of a precise, similarly inadequate criterion,” Tukey proposed alternatives, including exploratory data analysis and robust methods. Forty years later, Hand’s criticisms echo his concerns.

Robert A. Stine is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: stine@wharton.upenn.edu).

Hand presents a range of criticisms of modern classifiers. I find it useful to organize my discussion by grouping them into two clusters:

- Creeping incrementalism
- Square pegs in round holes.

Let me start with the first of these.

Creeping incrementalism. Hand argues that concerns for optimality emphasize tiny improvements that are dwarfed by other issues in real applications. He argues that the first predictor or the most simple of models finds most of the structure. Adding bells and whistles contributes little more than complex window dressing, and the advantages are illusions that disappear during the application. The argument is analogous to saying that linear Taylor series make pretty good approximations to most functions; generally, you do not need those messy, higher order terms. I certainly agree that simple models—or at least simple methodologies—take you a long way. Dean Foster and I wrote a paper to make just this point when mining financial data: with a few adjustments, stepwise linear regression can predict bankruptcy as well as elaborate trees (Foster and Stine, 2004).

A convincing argument for preferring simpler models requires careful discussions of applications. Given the depth of his experience, I had expected Hand to offer a rich portfolio of examples that demonstrate the failures of complex models. Instead, he relies more on an idealized example (one of equally correlated predictors) and a summary of fitted models to selected data sets from the repository at UC Irvine. One has to be careful basing arguments on made-up examples, because it is too easy to turn the examples around. With equally correlated predictors, the first one or two predictors capture most of the signal, with diminishing benefits left to the others. Although I have had similar experiences modeling real data, it is all too easy to make up normal models in which later variables appear to explain the most variation. For example, define

$$(1) \quad \begin{aligned} X_1 &= \tau Y + \varepsilon_1 + \varepsilon_2, \\ X_2 &= \tau Y + \varepsilon_1 - \varepsilon_2, \\ X_3 &= \tau Y - \varepsilon_1 + \varepsilon_3, \\ X_4 &= \tau Y - \varepsilon_1 - \varepsilon_3, \end{aligned} \quad \text{where } Y, \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, 1).$$

Each predictor X_j has equal correlation τ with Y and the predictors have a block structure. In this setting, what happens as we greedily expand the regression model is shown in Table 1. With $\tau = 0.25$, we have superadditive growth in the fit of the model: the addition of a subsequent predictor adds more to the model than any predecessor. I am not claiming that this example is more natural than the one in the paper. That is not the point. The point is that, separated from a real application, it is easy to construct examples that support any argument. What matters is what is useful in practice, and we need to see more evidence from real applications to appreciate the flaws of complex models.

I think that one needs to “go easy” when it comes to criticism of the use of statistical inference to judge improvements in a model. Inferential statistics concerns the separation of even a little signal from noise. This perspective is ideally suited to applications in traditional science. Discovery of statistically significant anomalies from the standard theory *is* important. A statistically significant anomaly, even a small one, cannot be dismissed as random variation and leads to revisions of the current theory. However, there needs to be a current theory in the first place. Without an established point of reference, the yardstick used to gauge improvements should be different. Most real applications lack such a benchmark and resemble an entirely new domain. When I was first learning about the connection between statistics and information theory, I was interested in the use of statistical models for data compression. (Think of tools used to compress the files on your computer disk.) Early on, improvements to algorithms for data compression regularly brought reductions of 20 or 30% in the amount of disk space required to store a data file. As the area matured, the gains got smaller and issues of statistical significance became relevant. Statistical significance in this context amounts to resolving whether you can save two or three more bits!

TABLE 1

Number of predictors	Explained variation		
	$\tau = 0.25$	$\tau = 0.5$	
0	0	0	
1	$\frac{1}{1+2/\tau^2} =$	0.03	0.11
2	$\frac{1}{1+1/(2\tau^2)} =$	0.11	0.33
3	$\frac{1}{1+1/(5\tau^2)} =$	0.24	0.55
4	1	1	

It is also important to establish what it means for a model to be better than another. Statistical significance offers one scale, but it may be poorly suited to the task. Finding an acceptable alternative can be particularly hard (e.g., in the social sciences), but is often easy in business. In business, improvements generally get measured in dollars, and statistical significance seldom guarantees much in the way of economic benefits. This point needs to be stressed as prominently and concretely as possible. Hand discusses the choice of the loss function used to judge classifiers and rightfully criticizes the casual use of error rates. Unfortunately, the survey of fitted models summarized in his Table 1, however, compares error rates. Who is to say that a small improvement in predictive accuracy is not valuable? Consider the data set “Segmentation” in the first row of his Table 1. Perhaps the reduction in the error rate from 0.083 to 0.014 is worth quite a lot of money. Without deeper insights into these applications, I cannot judge whether the improvements are impressive or unimportant. I doubt that enough is known about these applications to set costs, but perhaps Hand could offer other examples from his own experience in which the costs are known.

Square peg in the round hole. Statistics has rightly been criticized for often devoting too much energy to unrealistic problems. As Tukey pointed out, “Better to have an approximate answer to the right question than the exact answer to the wrong question.” Knowing the right question, however, often means knowing more about the application than most of us get from clients. In working with banks on credit modeling, the proprietary nature of their business makes it nearly impossible for them to be able to disclose enough for me to think that I am answering the right question. That does not mean that I have stopped trying, but it gets painful to jam your foot in the door over and over. It can be a lot more satisfying to prove a theorem or write code for a new algorithm.

Another reason for solving the wrong problem is that by the time one has the data and builds a model, the problem has changed. I would push to the front of the line to agree with Hand that changes in the underlying population pose a serious problem. This problem is particularly acute in business because of its competitive environment. If a company builds a model that produces a change in its behavior (such as a better way to evaluate the risk of loans that it makes), you can be sure that the competition will react and change as well.

I recently had a first-hand experience with this type of problem. The task was to help a company improve

the methods that it uses to evaluate prospective employees. Based on attributes known at the time of an application, we developed a classifier that was able to identify those most likely to succeed. The usual sorts of validation exercises showed that the effects we found were real, at least for the population represented by our data. As pointed out by Hand in Section 3, it takes a long time to get the data needed for this type of modeling. In our case, we had to wait and see which employees succeeded before we got the response. The delay was two years. By the time that the company tried to use the model, the economy had changed and the nature of the people applying for jobs had shifted. In fact, because we identified certain factors as important, the company changed the way that it collected these factors, rearranging the application form to emphasize the presentation of the key questions. I have little doubt that the revised questions measure different things than those used to build our model. Our model was a disappointment, but then I doubt that any model would have handled these disruptions.

I owe a favorite example of how the use of a model changes the population to Professor Hand. Suppose we are building a model to score the credit-worthiness of our customers. We discover that customers who, like me, drive white cars are poor risks. As a result, we stop offering loans to those driving white cars. Now think about what happens in several years when it is time to refresh the scoring model. By this time, none of our customers drives a white car, so this characteristic no longer appears to be a risk factor. Our successor will have to learn this all over again—that is, if these drivers have not changed their color preference. In the utopian world of repeated sampling from the population, these things do not happen. The population does not change because you start to use a model.

What next? Einstein once remarked, “Everything should be made as simple as possible, but not simpler.” Given a preference for simple models, I would very much like for Hand to offer some guidance *suited to applications* on how one is supposed to decide whether it is useful to look for more structure. If not by the ruler

given by statistical inference, then how? In my toy example, the sum $X_1 + X_2 + X_3 + X_4$ predicts Y perfectly. What should we do, however, when we have a wide data set with relatively few cases and 1000 predictors? How would we know to try the sum of them all as a predictor? Stepwise methods that build up models are good at finding subadditive models, but superadditive structures are difficult to identify. Similarly, we have methods that capture nonlinear features in data, but how are we to know whether to try them? If we only look for simple models, then we will always find simple models. To find nonlinearities requires that we entertain models that allow them. For example, our regression model for predicting bankruptcy uses interactions that, in effect, segment the population. Without them, the predictions were much less able to predict bankruptcies and left a lot of money on the table (Foster and Stine, 2004).

Professor Hand has had more experience with the challenges of dealing with real applications than most statisticians. I would be very interested in his approach to deciding when additions to a simple model *are* worthwhile. Similarly, what are his thoughts on methods to assess population drift? Certainly, statisticians have been concerned about population drift for a long time. For example, consider the article by Brown, Durbin and Evans (1975) on detecting changes in a linear model, Kalman filters that explicitly model an evolving state variable or models for evolutionary time series dating back to Priestley (1965). Do these fail in practice?

REFERENCES

- BROWN, R. L., DURBIN, J. and EVANS, J. M. (1975). Techniques for testing the constancy of regression relationships over time (with discussion). *J. Roy. Statist. Soc. Ser. B* **37** 149–192.
- FOSTER, D. P. and STINE, R. A. (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *J. Amer. Statist. Assoc.* **99** 303–313.
- PRIESTLEY, M. B. (1965). Evolutionary spectra and nonstationary processes. *J. Roy. Statist. Soc. Ser. B* **27** 204–237.
- TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.

Rejoinder: Classifier Technology and the Illusion of Progress

David J. Hand

I would like to thank the discussants for some very stimulating comments. Being only human, I am naturally pleased when others produce evidence or arguments in support of my contentions, but being a scientist, I am also pleased when others produce evidence or arguments against my proposals (although I may have to take a deep breath first), since this represents the scientific process in action.

I should first make one thing clear: I agree with Professor Friedman that substantial advances have been made in recent years. Indeed, in my paper I remarked that “developments such as the bootstrap and other resampling approaches ... have led to significant advances in classification and other statistical models.” However, what I question is whether the advances, when taken in the context of real practical problems, are as great as is often claimed—the recognition of the limitations of the new methods to which Professor Friedman refers.

Professor Friedman agrees with my three points that the improvements of newer methods over older ones are less than those of the older ones over still older ones, that the evidence favoring the superiority of new methods is often suspect and that the new methods fail to tackle important problems. I draw the conclusion from these points that progress is not as great as is imagined. Professor Friedman draws the conclusion that low lying fruit is easier to gather, that initial validation of new methods should be more rigorous and that much work remains to be done. Perhaps, then, we are really broadly in agreement—only perhaps I am describing a half empty glass (the new classification tools are not as wonderful as they are claimed), while Professor Friedman is describing a half full glass (some classification tools represent advances over the older ones).

I admit that I did criticize error rate as a performance measure and then used it in the examples. Since most performance comparisons of classifiers use error rate, this seemed justifiable, and I believe that my conclusions will generalize to other performance measures. For example, I agree that in some two-class problems it is the rank order of the estimated class 1 membership

probabilities which matters and that modern methods may well be able to estimate this more accurately than older methods. However, surely my points about population drift, class definition uncertainty and so on still apply and, of course, my point that people often use one criterion to fit a model and another to evaluate it applies even more strongly.

In fact, this point about people using different criteria manifests itself at a higher level when Professor Friedman and I examine my Table 1. I see the proportion of reduction of error rate achieved by the best method which can be achieved by discriminant analysis, whereas Professor Friedman sees the ratio of the error rates. I see a large initial improvement so that subsequent improvements are relatively small; he sees a large reduction in the proportion remaining. Back to the half full/half empty glasses again. We are both right, of course, although perhaps the different perspectives are valuable for different uses. For example, I agree with Professor Friedman’s example of the zip code classifier—and here the ratio of error rates might be a sensible measure—but (I would imagine) this is a problem in which the distributions are fairly static. In other problems, the distributions will change rapidly and I can imagine many contexts when I would not want to place too much trust in a reduction of error rate by a factor even as large as 10, if it corresponded to a change from a starting point as small as 0.001 to an even smaller one of 0.0001. A slight shift in the shapes of the distributions might induce sufficiently large changes in error rate so as to make this change irrelevant.

My regression example in Section 2.1 was merely intended as an additional illustration of the fact that the sequential nature of modeling means that typically later improvements are smaller than early ones. I am suggesting that the first, relatively crude, models will generally yield greater marginal improvements in predictive power than the later models. This is the low hanging fruit phenomenon—although, as noted below and as Professor Stine illustrates, there are exceptions.

I am glad Professor Friedman agrees so strongly with Section 5 of the paper, on the difficulties of obtaining

generally valid empirical comparisons. I think this is one of the most important parts of the paper. Professor Friedman's suggestion that the top performers in comparison studies should be ignored and attention should be focussed on the relative rankings of the others is very valuable. I also recommend looking at those methods which generally perform well, even if they seldom perform best, since they will have some sort of robustness. I think these sorts of issues, which represent aspects of the *art* of statistics, are fundamental to good statistical practice. They are the sorts of things which are not taught in standard statistics texts.

Professor Friedman comments that certain methods (he uses ensemble methods and support vector machines as examples) "offer substantial advantages over the earlier methods in enough situations to be regarded as major advances." I agree that such methods do represent significant theoretical and practical advances. My point is the milder one that "the practical impact of the developments has been inflated; that although progress has been made, it may well not be as great as has been suggested." Again referring to population drift as an example, a better fit to data drawn from a given distribution is not so wonderful if the distribution has changed. In fact, of course, it is likely that Professor Friedman and I have slightly different experience in terms of application domains. He cites "scientific and engineering applications" and I cite examples such as credit scoring and fraud detection: he draws attention to the differences between domains toward the end of his contribution; it is possible that population drift is more apparent in the latter than the former.

I entirely endorse Professor Friedman's comment that "obtaining high quality representative training data is generally more important to success than choice of a particular classifier." We are agreed on this, but in part my paper aims to point out that obtaining "representative training data" may be harder than is often imagined. Incidentally, I often go one step further and suggest that the best way to dramatically improve classifier performance is to add suitably chosen extra discriminating variables—that this is likely to exceed the performance improvement attained by juggling with classification rules, but, of course, this does depend on the specifics of the application.

Professor Friedman points out that almost all modern procedures incorporate a regularization parameter that controls the goodness of fit to the training data, and that one way to overcome problems such as population drift or uncertainty in the class definitions is to regularize more heavily than one would if such problems

were not suspected. I agree, and I also agree that there is no reason to suppose that the arbitrary amount of extra regularization implied by simpler older methods is the right amount. Indeed, of course one can always find examples where it is not, such as the large d small n cases of bioinformatics. However, if one is unable to get a handle on the amount of regularization which is needed, then there is no reason to suppose that the more heavily regularized modern method will be any better than the implicitly regularized older method.

Professor Friedman provides a useful discussion of tools for handling errors in class labels. These are fine if one suspects that one has such errors. However, I was concerned with the question of robustness to such errors if one is using a more standard method, unaware of the possibility.

I am sorry to have disappointed Professor Stine by not giving "a rich portfolio of examples that demonstrate the failures of complex models." To some extent I am caught in a Catch-22 situation here. For example, had I demonstrated the superiority of a simple linear classifier over a complex support vector machine in a real example involving dramatic population drift, then an obvious response would have been to build a more elaborate dynamic classifier or apply a modern model with heavier than standard regularization, as suggested by Professor Friedman. For this particular situation, the "even more elaborate model" would then win—and this will always be the case for any particular example. However, across examples, when one does not have specific reasons to expect such departures from the classic "fixed underlying distributions, precise class definitions" and so on of the standard problem, then one will not use a tool specially matched to the problem, so there is a risk that one will miss important features of the problem. Perhaps all I am really saying is that every problem has unique features, and that ideally one would carefully model and allow for those features, but if one is unaware of them (implicit in the use of standard tools), then simple is better.

My reason for using the idealized example of equally correlated predictors in Section 2.1 was merely to make the mathematics particularly transparent. Indeed, I pointed out that in real applications, the phenomenon I demonstrated was likely to be even more pronounced. However, I take Professor Stine's point that artificial examples can be used to support any argument (the intertwined spirals example being a case in point!), but, in spite of the ingenuity of his superadditive growth example, I believe that empirical evidence shows that

decreasing marginal improvement as extra terms are added to a model is the norm.

I am not arguing that there are *no* contexts in which a small improvement in performance is valuable. Professor Stine's example of data compression is a nice one. Another, of course, would be a small improvement in classification accuracy in a medical screening context—correctly diagnosing people in time to be treated, for example. My argument relates this apparent small improvement to other sources of uncertainty in the problem. If the distributions of characteristics of people with the disease differ from the distributions used to construct the classification rule, then the apparent improvement may be illusory. Statistical significance does not affect this argument. If the distributions are not the right ones, it does not matter how statistically significant the apparent improvements are.

Like Professor Friedman, Professor Stine takes me to task for criticizing the use of inappropriate performance criteria (which we agree is wrong) but then using error rates in my example in Table 1. I agree, of course, and in an ideal world I would have used performance criteria better matched to the particular problems and objectives. To do this I would have had to use my own examples, for which I knew the relevant performance criteria, and then compared linear discriminant analysis with the best performance I could achieve using neural networks, support vector machines, random forests and the whole panoply of other methods. However, if I then tried to argue, as I did in the paper, that these sophisticated tools were not that much better than linear discriminant analysis, I would immediately be vulnerable to the criticism that this was simply because I was not very adept at using the other methods. I thought it would be more compelling to use the results of other, expert, analysts. This meant I was forced to use error rate in my comparisons, simply because this is the most widely used criterion.

Professor Stine's comment about the difficulty of extracting the full story from commercial clients, so that one is confident that one is answering the right question, struck a chord. Even worse, all too often the client is *incapable* of formulating a precise question. This is not intended as a criticism: often the intrinsic uncertainties of the world (especially the commercial world) make precise formulation impossible. This, of course, was one of the issues which stimulated my writing of the paper.

Professor Stine's example of population drift in a personnel selection problem is very nice. It involves the key issue of drift due to natural background changes

(the economy), but also, presumably, the employees on which the model was built were not a random selection from previous applicants, but had been chosen because someone thought they were likely to be successful employees. This means, of course, that the classifier would have been modeling inappropriate distributions, unless some effort was made to represent this prior selection process. This is the same problem as that in the example of drivers of white cars which Professor Stine cites, although to a less extreme extent. I suspect that Professor Stine is right when he doubts that any model would have been very successful on this problem. Personnel selection problems are notoriously difficult. My point is merely that there are aspects of this problem which are not considered in the classical supervised classification paradigm, which consists of trying to model underlying distributions from a sample of data drawn from those distributions.

Toward the end of his contribution, Professor Stine asks for my suggestions on how to decide whether it is useful to look for extra structure. I think one should always look for this, but there are different kinds of structure. There is the structure represented by shape of the distribution from which the design data were drawn, and there is structure in the overall problem (e.g., population drift). I am suggesting that we are now pretty good at modeling the former, but that often the extra features of the distributions that our clever modern methods pick up are relatively unimportant compared with the potential impact of taking into account the latter kind of structure. So my answer to Professor Stine's question about my approach to deciding when additions to a simple model are worthwhile is that I think it is a matter of priorities. It is one thing to be able to add another hidden node to a neural network and hence reduce the misclassification rate (on those distributions) by 0.5%. It is another (and often a more useful) thing to be able to say that one is really interested in cost weighted error rate and is uncertain about the costs, so that the Gini coefficient is a more appropriate measure of performance, or that one believes the design data do not properly represent the distributions of new cases and so on.

As far as population drift is concerned, I think Professor Stine's final paragraph hits the nail on the head: statisticians now have a powerful armory of methods to tackle this, but how often does one see them integrated into the design of a classification rule?

It is in this vein that Professor Holte rightly points out that there are methods for dealing with some of the factors that I identify as being unknown at the time of

classifier design or subject to change after that time. In fact, I would be surprised if methods do not exist for *all* such factors: the Kalman filters Professor Stine refers to for population drift, Heckman models for sample selectivity, the cost curves of Professor Holte and the weighted Gini coefficients of Adams and Hand for unknown relative misclassification costs, for example. In addition, if tools for coping with a particular kind of uncertainty in the problem indeed do not exist, then it is, as Professor Holte says, a challenge for future research. Even if such tools exist, how often are they applied? Once again I wonder if, perhaps, it is just that it is easier to refine an existing form of classification model (the extra nodes of the neural network, the more sophisticated metric in nearest neighbor methods, . . .) than to model the sample distortion or adopt a more complicated performance criterion. Perhaps many of us academic researchers are still guilty of focusing too much on Tukey's exact answer to the wrong problem. I hope I may be forgiven for making that comment, since I, too, am an academic researcher and I, too, know the pleasure of developing a classification tool which appears to have a slight edge over its competitors.

Professor Gayler's comments were interesting, not least because they were from precisely the perspective which had stimulated many of my observations—the “nonclassical” problems which arise when applying supervised classifiers in the context of modeling human behavior, specifically credit scoring.

Professor Gayler points out the great financial gains which would result from a small increase in predictive accuracy in this application domain, so that one might have expected a premium to be placed on such performance, making the fact that relatively simple old-fashioned approaches are still used rather surprising. He also points out that the new methods are regularly investigated by the credit scoring community, but rarely make the transition to everyday practice, suggesting that the simpler older methods have some kind of advantage. Professor Gayler and I agree that this advantage arises from the kinds of issues described in my paper.

Professor Gayler mentions yet other kinds of complications. For example, he refers to account management changes (which will occur after the accept/reject classification has been made). This is a special case of a more general class of problem. Often we want to predict into what class an object (often a person) will fall if we take some action. However, if our prediction suggests that they will fall into some undesirable class,

then we take some other action. This, of course, invalidates the prediction. It is a generalization of the reject inference problem and leads to particular sample selectivity issues.

Professor Gayler is right to point out that in many problems the value of the threshold (to be compared with the estimated probability of belonging to class 1, e.g.) above which objects are assigned to class 1 depends on operational decisions, and these will be determined by all sorts of external factors.

I was particularly struck by Professor Gayler's observation that “in the limit (and the hands of a skilled modeler), every modeling technique should end up in agreement because they are all approximating the same data.” I am reminded of Hoadley's ping-pong theorem, which presumably represents alternate steps toward this limit! I was also taken by his suggestion that it might be more useful “to look at the effort required of the modeler to achieve a given goodness of fit and other properties of the models that are of operational relevance to the lender.” I endorse this. Of what good (at least in the credit scoring context) is a tool so highly sophisticated that it can be used effectively only after years of practice and experience? Operational relevance is a key factor.

In fact, my comment about “intertwined spirals or checkerboard patterns” refers to more than problems which can be modeled only as interactions between the variables. I meant it also to refer to those problems which have an extremely complicated (or perhaps contrived) decision surface. Such problems appear to be extremely rare in the real world, so demonstration of the power of new methods by showing that they can tackle such problems is rarely relevant to real problems. I conjecture that such problems are rare because in real problems the predictor variables will generally have been chosen because they are thought to have some discriminatory power, and predicting that the classes would be separated in such a complex way by a combination of variables would be an extraordinary intellectual feat. It is much easier to identify variables on which the members of one class have a tendency toward higher values than the members of the other class.

I like Professor Gayler's observation that it would be undesirable (in a credit scoring context) for a small change in decisions made when modeling to lead to a large change in the models. This is true and is a nice example of the pressures that favor simple modeling strategies. In such an environment, the organizations need to be confident of their modeling strategy and that

it will be reliable in the hands of other, perhaps less experienced staff. This is a phenomenon similar to, but at a level different from, the flat maximum effect. There the users of the models want to be confident that slight changes in the model (and indeed, the modeling conditions) will not lead to sudden dramatic deterioration in performance: as Gayler says, the flat maximum effect is a great advantage in credit scoring.

I still have a suspicion that there is too much emphasis on trying to squeeze the last drops of performance out of classifiers matched to a particular data set when these distributions might not be the right ones, when the performance criterion being used is inappropriate, when the class definitions might be incorrect or subject to change and so on, with all the mismatches illustrated in the paper and others. Instead, I believe that more effort should be spent on trying to identify and model aspects of the problem which deviate from the classical supervised classification paradigm, and which may have a substantial impact on performance. For example, if you suspect the populations will change (perhaps not in Professor Friedman's scientific and engineering problems, but certainly in the personnel and social applications of Professor Stine, Professor Gayler

and myself), then either model this or regularize more heavily to allow for it; if you suspect that the sample has not been randomly drawn but has been purposively selected (as in Professor Stine's employee selection example), use a model which adjusts for the hypothesized selectivity or more heavily regularize to avoid overfitting a suspected inaccurate distribution; if you know you are concerned with maximizing profit, then use profit as a performance criterion, and not misclassification rate or likelihood, or else regularize more heavily to allow for the fact that there is a mismatch between the criterion being used and the one of real interest; and so on.

I am extremely grateful to the discussants for their thoughtful comments on the paper. It is apparent that they spent a considerable amount of time and effort carefully considering my points, and marshalling coherent and instructive responses. Their comments covered a wide range of issues and approached things from different perspectives. It is very clear that, whatever the merits of the paper itself, the discussion contributions have substantial intrinsic value, and I have certainly learnt a great deal from them.