Assessing and improving classification rules

David J. Hand, Keming Yu, and Niall Adams

Department of Statistics
The Open University
MK7 6AA
UK

d.j.hand@open.ac.uk
n.adams@open.ac.uk
k.yu@open.ac.uk

1. Introduction

The last few years have witnessed a resurgence of research effort aimed at developing improved techniques for supervised classification problems. In a large part this resurgence of interest has been stimulated by the novelty of multi-layer feedforward neural networks (Hertz *et al*, 1991; Ripley, 1996) and similar complex and flexible models such as MARS (Friedman, 1991), projection pursuit regression (Friedman and Stuetzle, 1981), and additive models in general (Hastie and Tibshirani, 1990)). The flexibility of these models is in striking contrast to the simplicity of models such as simple linear discriminant analysis, perceptrons, and logistic discriminant analysis, which assume highly restricted forms of decision surface.

The merit of the flexibility of neural networks is countered by the dangers that they will *overfit* the design data. This relationship between model flexibility and the danger of overfitting has long been understood within the statistical community. For example, in the 1960s the optimistic bias of resubstitution error rate became widely recognised and it was replaced by the leave-one-out method as the method of choice. (Later, in the 1980s, an apparently large variance of the latter led to its being abandoned in favour of bootstrap methods, in particular the 632 bootstrap.) Early work on neural networks also fell into this trap, producing inflated claims of the performance of such models derived from optimistic performance measures based on overfitting the design set. In recent years the risk has been recognised, and some sophisticated proposals have been made for overcoming the problem. They are based on ideas such as penalising the goodness of fit measure (by combining it with a measure of model complexity), restricting the form of the model (to few nodes in a network, for example), shrinking an overfitted model (by *weight decay*, for example), or even by adding randomly perturbed replicates to the design set. The problem with all such methods is *how* to strike the optimum compromise between modelling the design data and overfitting.

Greater theoretical insight into the problem can be obtained by close examination of the criteria which are optimised to fit the model. Two strands of work can be identified in these investigations, viewing the problem from opposite perspectives. The first (e.g. Hand, 1994, 1995, 1997 (chapter 6)) has focused on criteria for *assessing* the performance of classification rules. As outlined below, this strand considers different performance criteria and the

decomposition of these criteria into distinct components. Such decompositions mean that the weaknesses of a given classification rule can be identified and steps taken to improve it. It was largely motivated by the realisation that *performance* of a classification rule has many different aspects, and that what is important varies from problem to problem. The second strand of work (e.g. Breiman, 1995, 1996; Kohavi and Wolpert, 1996; Tibshirani, 1996), has typically focused on criteria for fitting and selecting classification rules and has considered the direct improvement of components of those measures (such as bias or variance - see below). Clearly the two strands are opposite sides of the same coin.

An interesting aspect of the bulk of the work on supervised classification is that different criteria have often been adopted for parameter estimation and for performance assessment. In particular, error rate is by far the most popular measure of performance. (This despite the fact that crude error rate is seldom of primary interest when a rule is applied). However, error rate, being a discontinuous function, is not amenable to conventional mathematical optimisation methods and so has generally not been used for parameter estimation. An exception here is early work on the perceptron, with its well-known adaptive estimation algorithm based on minimising error rate. In contrast, most of the recent work on multi-layer neural networks uses the least squares criterion or the cross entropy. We say more about effect of using different criteria for these two roles below.

2. Aspects of performance

Performance assessment is motivated by the basic question: which classifier should one use? This question does not permit a simple answer. As is described in detail in Hand (1997), the answer depends critically on the precise objectives of the classification exercise. However, certain measures have gained currency as the most popular. Of these, by far the most widely used is straightforward error rate. Unfortunately, this is not because it is necessarily the most appropriate criterion. In fact, we would argue, it is seldom ideally matched to the problem. In particular, error rate implicitly assumes that the costs of different types of misclassification (from class 1 to class 2 and vice versa, for example) are equal. Needless to say, this is unlikely to be the case in most real applications. Similarly, error rate says nothing about the relative severity of different misclassifications to the same class: an object misclassified because it is just barely on the wrong side of a threshold will be accorded the same weight as one misclassified because it is well away from the threshold, on the wrong side. In some situations this is the appropriate thing to do, but in others a measure of relative severity would provide useful information. Error rate also tells us nothing about the accuracy of the probability estimates implicit in the classifier. If the threshold is not determined *a priori* for all future applications (if, for example, priors or costs might vary with time) then one will want these probability estimates to be accurate over a range of potential thresholds, and not simply at one point.

These apparent demerits of error rate have stimulated the search for measures which look at other aspects of performance.

For simplicity suppose that we have just two classes (the result below can be generalised), with class membership shown by an indicator variable c taking values 0 or 1. Then many classification rules are based on comparing the estimated probability, $\hat{p}(x)$, that an object with measurement vector x will belong to class 1, with a threshold t. Of course, it is not necessary that probability estimates be involved in such a procedure - any monotonic transformation of the

probabilities would be equally effective. Indeed, any transformed version of the predicting function which was greater than the transformed threshold if and only if $\hat{p}(x)$ was above t would be equally effective. However, depending on the precise nature of the problem, there are often advantages in basing rules on explicit probability estimates (see Hand, 1997, for details).

Two performance measures which have been fairly widely adopted are the *Brier score* and *log score*. Brier score is defined as $\sum_i (\hat{p}(x_i) - c_i)^2$, where the subscript i signifies the ith object in a test set. That is, it is the sum of squared errors between the predicted probability and true class indicator. Log score is defined as $-\sum_i c_i \log \hat{p}(x_i) - \sum_i (1-c_i)\log(1-\hat{p}(x_i))$.

In population terms, Brier score is the mean squared error between the predictions and the true class indicators, $\underset{x}{E}\,\underset{c|x}{E}(\hat{p}-c)^2$, where $\underset{c|x}{E}$ signifies expectation over true class c for a given x. Focussing attention on a fixed x, standard decomposition of this gives

$$\underset{c|x}{E}(\hat{p}-c)^2 = \underset{c|x}{E}\left(\hat{p}-\underset{c|x}{E}\,c\right)^2 + \underset{c|x}{E}\left(\underset{c|x}{E}\,c-c\right)^2$$

where the first term is the squared bias of the estimator and the second is the variance of the true class indicator at x. Note that here the second term on the right is independent of the classification rule, so that it provides a lower bound on the accuracy which an estimate can obtain. It also follows from this that if the Brier scores for two classifiers applied to the same data sets are subtracted then the result is an estimate of the difference between their mean (over x) squared biases. That is, we can easily obtain estimates of their relative (though not absolute) squared bias.

There is an important point to be made about this decomposition. This is that the above expression, when the expectation over x is taken, measures accuracy of an *estimate*. It is *conditional* on the design set. For a given design set, it will tell us how well the chosen method performs and is certainly something we will be interested in so that, for example, we can compare competing classification rules. However, we also want to know about the accuracy of *estimators*. That is, we want to know whether an estimation *procedure* (not conditioning on a particular design set) can be relied on to give good estimates, regardless of whether a particular instance turns out to be good. To explore this, we need to evaluate the expectation of the above over different design sets. Since the second term on the right hand side of the above expression is invariant to design set changes, we will focus on the first term. We have, still fixing attention on a given x

$$\underset{D}{E}\,\underset{c|x}{E}\left(\hat{p}-\underset{c|x}{E}\,c\right)^2 = \underset{D}{E}\left(\hat{p}-\underset{D}{E}\,\hat{p}\right)^2 + \left(\underset{D}{E}\,\hat{p}-\underset{c|x}{E}\,c\right)^2,$$

that is, a decomposition into the variance and the (squared) bias of the *estimator*.

Decompositions such as these, some more naturally suited to classification problems, have also recently been developed from the perspective of improving classification rule performance (Breiman, 1996; Tibshirani, 1996). By partitioning simple error measures, it is possible to focus on and remedy particular weaknesses of the rules. These decompositions split the measures into

three components which (following the above illustration) we can term variance of the estimator (over design sets), (squared) bias of the estimator, and variance of the true class. Which of these are important will depend on the application, with different components being important for different kinds of classification problem (this is the driving force behind the work on assessment). For example, in some problems bias is irrelevant provided most classifications are correct, whereas in others confidence in the accuracy of the estimated probabilities is needed.

Although decompositions of assessment/performance criteria have been investigated only recently, their components have a long history of exploration (which has not always been acknowledged in the more recent literature). They also have a history of being explored in different disciplines - so that, as a consequence, different terms have been coined for the same measure. Hand (1997, chapter 6) collates these different terms. Some important early work, in the context of classification rules, is that by Habbema *et al* (1978) and Hilden *et al* (1978a,b). For the special case of the two class case, much of the formalism parallels that of work on Bayesian scoring, on which there is a considerable body of advanced material (see, for example, Murphy, 1972a,b, 1973; Yates, 1982; DeGroot and Fienberg, 1983).

We mentioned two criteria above, the Brier and logarithmic scores. When viewed from the perspective of being criteria to be optimised to yield a classification rule (rather than as performance measures) they are immediately seen to be log-likelihoods derived from normal and Bernoulli distributions, respectively. When looked at from this viewpoint the log score seems more natural: the indicator function variable c defined above is Bernoulli distributed with parameter $p(x)$. We might also look to see if there are other properties which will allow us to choose between them. One such property is called *properness*.

A *proper* criterion is one which is minimised by the true probabilities - any other values yield a larger (equals poorer) measure. Both Brier score and log score are proper. However, we have already noted that classification rules typically introduce bias deliberately so as to avoid the excessive variance implicit with using the raw design set. This means that they are unlikely to achieve the minimum possible score. This being the case one might wonder if 'properness' is a useful criterion in this regard. The properness criterion has recently been rediscovered by the neural networks community (and given different names).

The relationship between Brier score and logarithmic score can be seen as follows. The population version of the negative of log score is

$$\int p(x)\log\hat{p}(x)dx+\int\big(1-p(x)\big)\log\big(1-\hat{p}(x)\big)dx=\int p\log\hat{p}+\int(1-p)\log(1-\hat{p}).$$

Subtracting $\int p\log p+\int(1-p)\log(1-p)$, which is the same for all classifiers and so will not affect any comparisons between classifiers, leads to

$$\begin{aligned}&\int p\log\hat{p}+\int(1-p)\log(1-\hat{p})-\int p\log p-\int(1-p)\log(1-p)\\&=\int p\log(\hat{p}/p)+\int(1-p)\log\big((1-\hat{p})/(1-p)\big)\end{aligned}$$

$$= \int p \log\left(1+\frac{\hat{p}-p}{p}\right) + \int (1-p)\log\left(1+\frac{p-\hat{p}}{(1-p)}\right)$$

$$\approx \int p\left\{\frac{\hat{p}-p}{p}\right\} - \frac{1}{2}\int p\left\{\frac{\hat{p}-p}{p}\right\}^2 + \int (1-p)\left\{\frac{p-\hat{p}}{1-p}\right\} - \frac{1}{2}\int (1-p)\left\{\frac{p-\hat{p}}{1-p}\right\}^2$$

$$= -\frac{1}{2}\int \left\{\frac{1}{p}+\frac{1}{1-p}\right\}(\hat{p}-p)^2$$

The $\{1/p+1/(1-p)\}$ factor means that, in comparisons between rules, the log score puts greater emphasis on probability estimates where p is near 0 or 1 than does the Brier score. In many problems the threshold will be set at 1/2, so that it is in these regions where accuracy is of more significance, not in the extremes of p. This suggests that Brier score might be the more appropriate measure.

On the other hand, misclassifications in the vicinity of the true decision surface (with a threshold of 1/2) are less serious than elsewhere: in regions where p = 1/2 one will inevitably misclassify 50% of the points whichever classes one assigns them to. Effort in such regions is wasted. Clearly an ideal measure would place most weight at an intermediate position between the extremes of p = 1/2 and p = 0 or 1.

3. Some simulation comparisons on neural networks

In the past the limitations of the restricted shapes of the decision surfaces of the more classical methods of supervised classification have been overcome by an explicit and user-directed process of *feature extraction*, in which combinations and transformations of the raw measurement variables are created to define a feature space in which the restricted decision surface form is effective in separating the classes. The key here is that the effectiveness of this feature selection process depends on the expertise of the developer of the system - in particular, on their knowledge and understanding of the data. The developer needs to know which combinations and transformations are likely to contribute to linear (for example) separation between the classes. In contrast, methods such as neural networks automatically find the features. Combinations and transformations of the raw measurements are included as a basic part of the models. Of course, this means that the space of possible models which they span is much greater than that of the more restricted models. This is the main reason why they have only recently attracted attention, despite the basic forms having been around since the early 1960s (Palmieri and Sanna, 1960; Gamba *et al*, 1961): they needed the power of modern computers before practical implementations could be produced.

The fact that neural networks will automatically find good features, without requiring expert knowledge or good understanding of the data, is one reason why such methods have attracted so much attention. They are an ideal black box – throw the data at them and they will give an answer. Of course, it also follows that if one does have a good understanding of the data, then it is likely that neural networks, and similar flexible classifiers, will not add to the performance of methods which make use of that understanding. After all, neural networks have to search the space of possible derived features, and the immensity of this space gives them

plenty of opportunity to capitalise on chance - to model aspects of the data which are peculiar to the finite, randomly selected design set, rather than being aspects of the true underlying decision surface.

These complementary aspects of the power and the problems arising from the flexibility of neural networks bring home the difficulties of separating the systematic components of decision surface variation from the random components which are specific solely to the design set. We have described above how different criteria can be used to tap different aspects of this variation. To explore this further, we carried out some simulation studies to compare different criteria for estimation and assessment. The different criteria: Brier score, Log score, Bhattacharyya score $\sum_i 1-\left(c_i\hat{p}(x_i)+(1-c_i)(1-\hat{p}(x_i))\right)^{\frac{1}{2}}$, and L10 score $\sum_i \left(\hat{p}(x_i)-c_i\right)^{10}$, produced very similar results. Moreover, measuring performance using one criterion while estimating parameters using another (as is often the case in practice) did not lead to significant degradation of performance. An extract from our simulation is given in Table 1.

Table 1: Simulation results. Columns are the criteria used in estimation. Rows are criteria used in assessment - yielding the values in the body of the table. Standard errors are given in brackets.

	Brier	Log	Bhatt	L10
Brier	0.0839 (1.4E-3)	0.0827 (1.3E-3)	0.0859 (1.2E-3)	0.1825 (4.9E-3)
Log	0.2792 (3.9E-3)	0.2821 (4.0E-3)	0.2845 (4.1E-3)	0.5521 (1.1E-2)
Bhatt	0.0972 (2.1E-3)	0.0919 (1.6E-3)	0.0971 (9.4E-3)	0.2389 (4.3E-3)
L10	0.0130 (6.6E-4)	0.0069 (2.7E-4)	0.0018 (7.0E-5)	0.0005 (2.0E-5)
Err. Rate	11.21 (2.4E-1)	10.69 (2.1E-1)	10.94 (1.2E-1)	12.79 (3.8E-1)

The simulations summarised in table 1 are based on design sample sizes of 50, randomly sampled from the synthetic data described in Ripley (1996). The classifier used was a neural network with 6 hidden nodes, with no additional shrinking. For each sample size 50 design sets were generated and 20 network optimisation runs carried out, those that converged to a local minimum being retained. Performance is assessed on an independent test set of 1000 observations. The columns of the table show the criteria used for estimating the parameters, and the rows the value of the indicated criterion at the estimated parameter values. The values in the body of the table show averages, over the 50 design sets, of the minima of the 20 runs.

Smoothing the classification rule using weight decay produces quite striking effects. The horizontal axes of figures 1 to 4 show the values of the (log) criterion used in estimating the parameters and the vertical axes show the assessment criterion. The letters a, b, c, and d correspond to four different extents of weight decay (0, 0.001, 0.01, and 0.1, respectively). All of these figures demonstrate the effectiveness of shrinking using weight decay in terms of improved predictive performance on the test set (the reduction in vertical range as the weight decay increases), but the effect is particularly striking when the log score is also used as the assessment criterion (figure 3). The increased horizontal range as weight decay is increased shows the deteriorating fit to the design set.

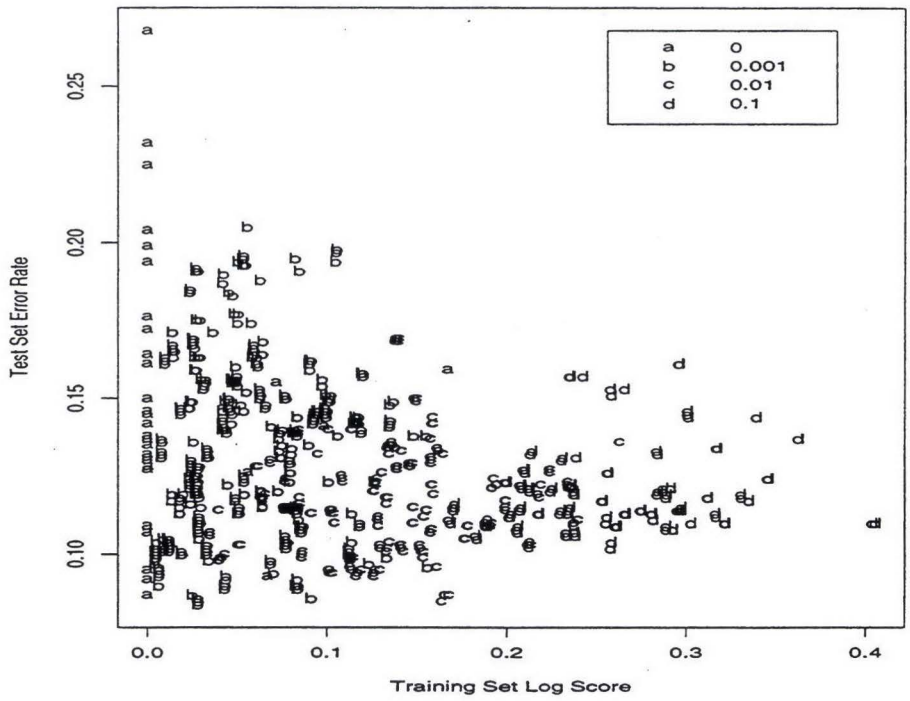

Figure 1: *Test set error rate against training set log score.*

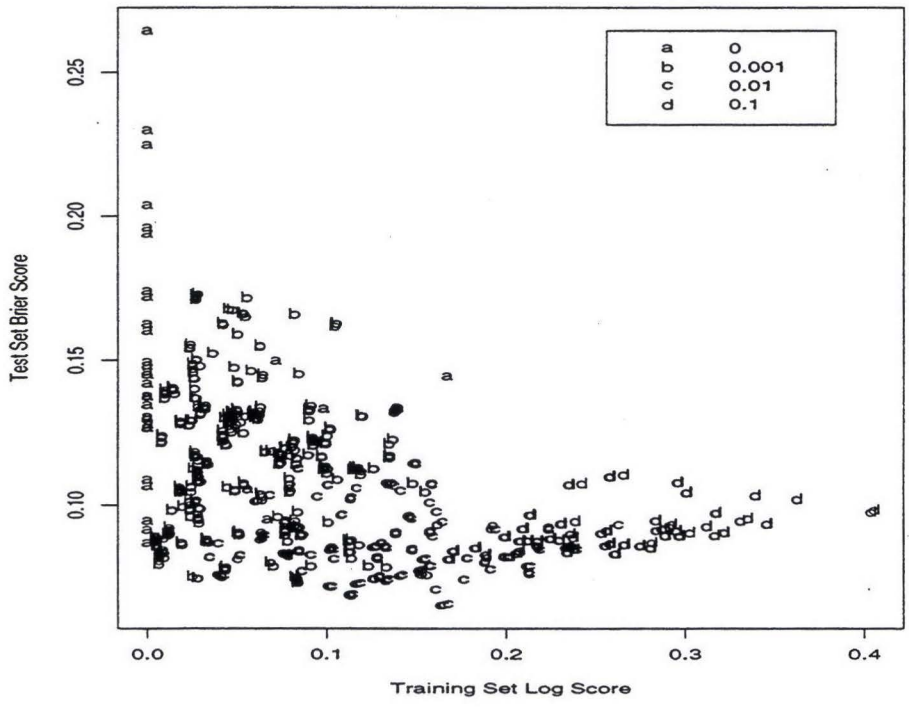

Figure 2: *Test set Brier score against training set log score.*

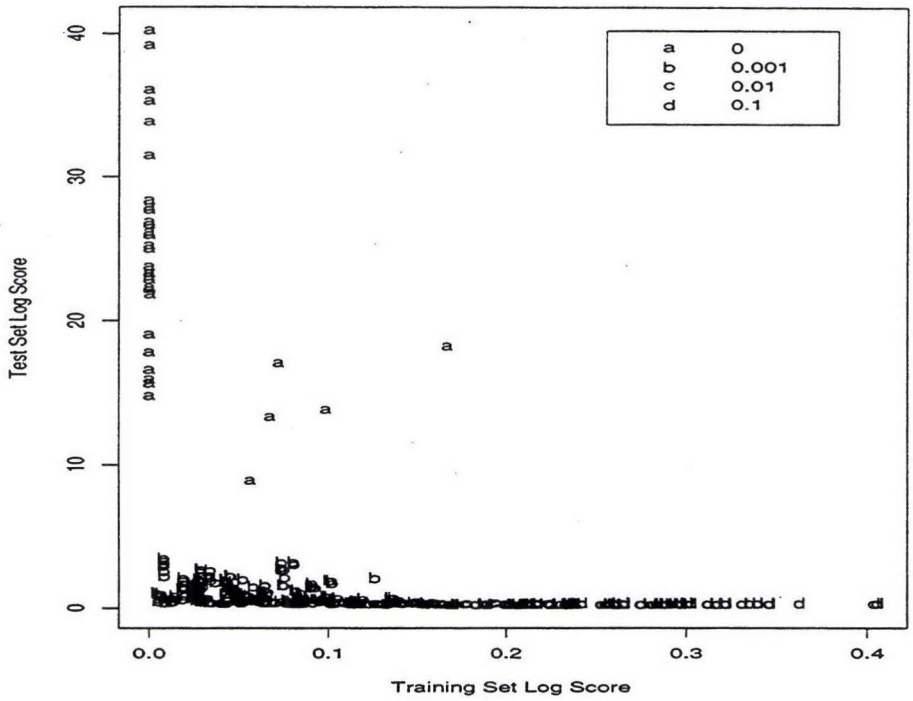

Figure 3: *Test set log score against training set log score.*

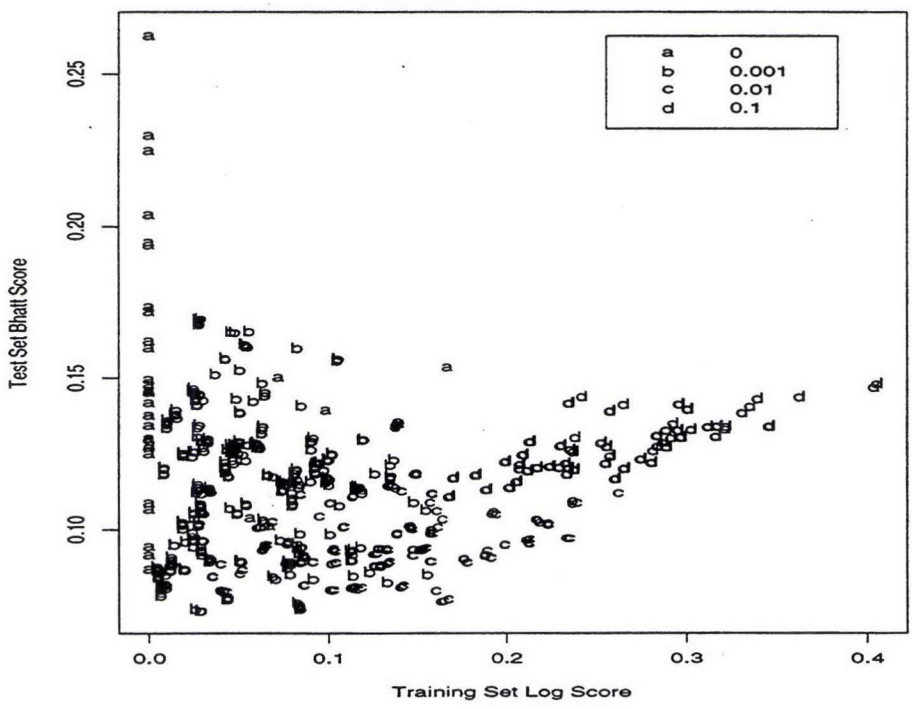

Figure 4. *Test set Bhattacharyya score against training set log score.*

4. Resemblance

The Bayes error is the smallest error rate which can be achieved by a classification rule for a given set of variables. It is thus a measure of *separability* (sometimes *inseparability* - see Hand, 1997), describing how different are the distributions of the classes. Separability measures are useful - in variable selection, for example. On the other hand, they are of limited value in assessing rules. There we want to know whether classification rules built using the *predicted* distributions really serve to separate the true distributions. That is, we want to know 'do the predicted probabilities induced by the rule *really* serve to separate the true classes well?' We call this property *resemblance*. Resemblance thus measures the variation between the true probabilities conditioned on the estimated ones, and ideally these probabilities will be very different. Letting $\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x}) = (\hat{p}(1|\mathbf{x}),\ldots,\hat{p}(C|\mathbf{x}))$, we want to know how different are the elements of the vector $\mathbf{p}(\mathbf{j}|\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x})) = (p(1|\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x})),\ldots,p(C|\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x})))$. An example of a possible measure would be the variance of the $p(k|\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x}))$, telling us how different these are from each other. Note that this is not the same as the variance of the $\hat{p}(k|\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x}))$, which would tell us how well the well the classifier 'thinks' it separates the true classes, and not how well it actually does separate them. Such a classifier could be hopelessly optimistic. In order to estimate the variance of the $p(k|\hat{\mathbf{p}}(\mathbf{j}|\mathbf{x}))$ we need access to another data set (the test set) or to use sample reuse methods such as bootstrap. The key point here is that resemblance measures evaluate true probabilities conditioned on the estimated probabilities.

5. Estimating bias

We noted above that the relative bias of rules could be found by subtracting probability scores for the rules - a common component, solely due to the particular test set in question, is cancelled by the subtraction. However, this does not permit us to obtain absolute estimates of bias. Yu, Hand, and Webb (1996) describe approaches for finding absolute estimates. Essentially, the integration over the complete measurement space is split into two parts: an integration over this space conditional on fixed values of the estimated conditional probability, $\hat{p}$, of belonging to class 1, followed by an integration of this estimated probability over the interval [0,1]. This is based on the strategy adopted by researchers involved in assessing Bayesian probabilities. The difference is that they carry out the second integration by splitting $\hat{p}$ into (typically) ten intervals and summing the squared bias estimates in these. This may be a legitimate way to proceed in Bayesian probability scoring, but is less suitable for assessing classifier performance. Yu *et al* (1996) therefore apply a smoothed estimator over the range of the $\hat{p}$.

6. Conclusion

Many comparisons of classification methods have now been carried out. Some use real data, but perhaps the majority use simulated data. Most use error rate, though a few use more sophisticated measures such as Brier score. However, different measures of performance tap different aspects of performance. The measure should be matched to the objectives. This applies also at high levels, not merely the low levels we have discussed in this paper. For example, in comparing methods to choose which to use in a commercial application, it is wise that the likely

users, rather than specialist experts, make the comparisons. Experts, almost by definition, can obtain better performance from the methods in which their expertise lies than can non-experts.

References

Breiman L. (1995) Bagging predictors. Department of Statistics, University of California, Berkeley, Technical Report.

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

DeGroot M. and Fienberg S.E. (1983) The comparison and evaluation of forecasters. *The Statistician*, **32**, 12-22.

Friedman J.H. (1991) Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, 1-141.

Friedman J.H. and Stuetzle W. (1981) Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817-823.

Gamba A.L., Gamberini G., Palmieri G. and Sanna R. (1961) Further experiments with PAPA. *Nuovo Cimento Suppl.*, **20**, 221-231.

Habbema J.D.F., Hilden J. and Bjerregaard B. (1978) The measurement of performance in probabilistic diagnosis I. The problem, descriptive tools, and measures based on classification matrices. *Methods of Information in Medicine*, **17**, 217-226.

Hand D.J. (1994) Assessing classification rules. *The Journal of Applied Statistics*, **21**, 3–16.

Hand D.J. (1995) Comparing allocation rules. *Statistics in transition*, **2**, 137–150.

Hand D.J. (1997) *Construction and Assessment of Classification Rules*. In press - expected January. Chichester: John Wiley and Sons.

Hastie T.J. and Tibshirani R.J. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Hertz J., Krogh A., and Palmer R.G. (1991) *Introduction to the Theory of Neural Computation.* Redwood City, California: Addison-Wesley.

Hilden J., Habbema J.D.F., and Bjerregaard B. (1978a) The measurement of performance in probabilistic diagnosis II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of Information in Medicine*, **17**, 227-237.

Hilden J., Habbema J.D.F., and Bjerregaard B. (1978b) The measurement of performance in probabilistic diagnosis III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine*, **17**, 238-246.

Kohavi R. and Wolpert D.H. (1996) Bias plus variance decomposition for zero-one loss functions. Department of Computer Science, Stabnford University Technical Report.

Murphy A.H. (1972a) Scalar and vector partitions of the probability score: Part I, Two state situation. *Journal of Applied Meteorology*, **11**, 273-282.

Murphy A.H. (1972b) Scalar and vector partitions of the probability score: Part II, N-state situation. *Journal of Applied Meteorology*, **11**, 1183-1192.

Murphy A.H. (1973) A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595-600.

Palmieri G. and Sanna R. (1960) *Methodos*. **12**, No.48.

Ripley B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

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

Yates F. (1982) External correspondence: decompositions of the mean probability score. *Organizational Behaviour and Human Performance*, **30**, 132-156.

Yu K., Hand D.J., and Webb A. (1996) Estimating the imprecision of classification rules. In preparation.