Machine Learning, 45, 171–186, 2001

A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems

DAVID J. HAND d.j.hand@ic.ac.uk
ROBERT J. TILL r.till@ic.ac.uk
Department of Mathematics, Imperial College, Huxley Building, 180 Queen's Gate, London SW7 2BZ, UK

Editor: David W. Aha

Abstract. The area under the ROC curve, or the equivalent Gini index, is a widely used measure of performance of supervised classification rules. It has the attractive property that it side-steps the need to specify the costs of the different kinds of misclassification. However, the simple form is only applicable to the case of two classes. We extend the definition to the case of more than two classes by averaging pairwise comparisons. This measure reduces to the standard form in the two class case. We compare its properties with the standard measure of proportion correct and an alternative definition of proportion correct based on pairwise comparison of classes for a simple artificial case and illustrate its application on eight data sets. On the data sets we examined, the measures produced similar, but not identical results, reflecting the different aspects of performance that they were measuring. Like the area under the ROC curve, the measure we propose is useful in those many situations where it is impossible to give costs for the different kinds of misclassification.

Keywords: receiver operating characteristic, ROC curve, AUC, Gini index, error rate

1. Introduction

This paper is concerned with supervised classification problems, in which the aim is to devise a method or construct a rule for assigning objects to one of a finite set of classes on the basis of a vector of variables measured on the objects. The information on which the rule is to be based is a *design* or *training* set of objects with known vectors of measurements and known classifications. In particular, this paper discusses problems involving more than two classes, and is concerned with measures of performance of such rules. We restrict ourselves to situations in which the rules yield estimates of the probability that a point belongs to each class, or yield scores indicating strength of class membership.

For the special case in which there are only two classes, there are many distinct criteria for comparing the performance of classification rules (e.g., see the detailed discussions in Hand, 1997, 2000). Amongst the most popular are misclassification (or error) rate, and the criterion with which this paper is concerned, the area under the Receiver Operating Characteristic (ROC) curve (e.g., see Hanley & McNeil, 1982; Zweig & Campbell, 1993; Bradley, 1997).

Misclassification rate is simply the expected proportion of future cases which the rule will classify incorrectly. More generally, if the cost of misclassifying a class i point

is c_i, $i = 0, 1$, then the overall expected loss is

$$L = \pi_0 p_0 c_0 + \pi_1 p_1 c_1 \tag{1}$$

where p_i is the probability of misclassifying a class i object, and π_i is the probability that an object comes from class i (we are assuming no costs associated with correct classifications). Misclassification or error rate is the special case in which $c_0 = c_1$ (and both are then taken to be equal to 1, without loss of generality). If $\hat{p}(x)$ is the estimated probability that an object with measurement vector x belongs to class 0, then a standard result shows that minimum loss is achieved by choosing the classification threshold such that points are classified into class 0 if $\hat{p} > t = c_1/(c_0 + c_1)$.

This is all very well if one knows the values of the costs c_0 and c_1. Typically, however, these costs are difficult to determine (e.g., see Bradley, 1997; Provost, Fawcett, & Kohavi, 1998; Adams & Hand, 1999, 2000) and the references in Turney (1996).

If the costs cannot be determined, an alternative strategy is simply to compare the overall distributions of $\hat{p}(x)$ for class 0 points and class 1 points. The classification rule will generally be better the more these two distributions differ. As we show in detail below, the area under the ROC curve (AUC) is such a measure of the difference between these two distributions. By focusing on a comparison of the distributions of the $\hat{p}(x)$, the AUC ignores the costs and also (a consequence of the way costs and priors appear together in Eq. (1) the class priors: it concentrates attention on how well the rule differentiates between the distributions of the two classes, and is not influenced by external factors which depend on the use to which the classification is to be put.

All of the above refers to the two class case. Often, however, we are faced with problems in which there are more than two classes. In principle, such problems may be tackled readily enough: one simply chooses a set of costs and defines a classification rule which minimises a multiple class extension of Eq. (1). (For convenience, we shall use the expression 'multiple class' to signify problems with more than two classes.) In practice, however, this is almost never feasible. The major problems of choosing realistic costs in the two class problem, described in Adams and Hand (1999), are compounded in the multiple class problem. The default (and, indeed, popular) choice of equal costs for the various different kinds of misclassification, leading to overall misclassification rate, is in fact very rarely really suitable. What is needed is a multiple class extension of the AUC approach, which side-steps the problem of choosing costs altogether. This paper describes such a method.

In Section 2, we formally define the AUC measure and show how to estimate it from a sample of data. Common methods are based on explicit integration of areas under the ROC curve, and these risk introducing unnecessary error into the estimate. An alternative, based on the relationship to the Mann-Whitney-Wilcoxon test statistic is described. This relationship is well-known (e.g., see Hanley & McNeil, 1982), but both it and the implications for estimation seem not always to be appreciated, so we describe it here for completeness.

In Section 3, we describe a straightforward multiple class extensions of the AUC measure. In Section 4 we illustrate some of the properties of the measure, and in Section 5 we present some examples.

2. Estimating the AUC coefficient

The AUC is defined in terms of the Receiver Operating Characteristic curve. Let $\hat{p}(x)$ be the estimate of the probability that an object with measurement vector x belongs to class 0. Let $f(\hat{p}) = f(\hat{p}(x) \mid 0)$ be the probability function of the estimated probability of belonging to class 0 for class 0 points, and let $g(\hat{p}) = g(\hat{p}(x) \mid 1)$ be the probability function of the estimated probability of belonging to class 0 for class 1 points. Let $F(\hat{p}) = F(\hat{p}(x) \mid 0)$ and $G(\hat{p}) = G(\hat{p}(x) \mid 1)$ be the corresponding cumulative distribution functions. Then the ROC curve is defined as a plot of $G(\hat{p})$, on the vertical axis, against $F(\hat{p})$, on the horizontal axis. Clearly this plot lies in a unit square. A good classification rule is reflected by an ROC curve which lies in the upper left triangle of the square. This follows since any point above the diagonal corresponds to a situation in which $G(\hat{p}) > F(\hat{p})$, so that the class 1 points have *lower* estimated probability of belonging to class 0 than do the class 0 points. This is equivalent to moving a threshold, t, from 0 to 1, and plotting $G(t)$ against $F(t)$ for each value of t. The better such a curve is, the closer it gets to the top left corner: perfect separation is indicated by a curve which follows the left hand and top edges of the square. A classification rule no better than chance is reflected by an ROC curve which follows the diagonal of the square, from the lower left corner to the top right corner. A classification rule which is worse than chance would produce an ROC curve which lay below the diagonal—but in this case performance superior to chance could be obtained by inverting the labels of the class predictions. The AUC is then simply the area under the ROC curve.

Not all representations of ROC curves use the same axes, though the principles are the same. In particular, some authors (e.g., Provost & Fawcett, 1997) plot $F(\hat{p})$ (estimated by the 'true positive rate') on the vertical axis and $G(\hat{p})$ (estimated by the 'false positive rate') on the horizontal axis—that is, our axes are interchanged in their plots. A simple interchange of axes would result in a good ROC curve (one which does better than chance) lying in the lower half of the ROC square. This is avoided by letting the threshold t referred to in the previous paragraph move from 1 to 0. This seems artificial to us, and to go against convention, which is why we choose the former strategy. However, both are merely conventions.

The AUC measure of performance is closely related to the *Gini coefficient*, which is sometimes used as an alternative measure. This is most commonly defined as twice the area between the ROC curve and the diagonal (with this area being taken as negative in the rare event that the curve lies below the diagonal). Elementary geometry shows that $\text{Gini} + 1 = 2 \times \text{AUC}$. In this paper we work in terms of AUC, but the results apply equally to the Gini coefficient.

ROC curves are typically estimated either by applying the classification rule to a test set of points with known classes, or by using a design sample reuse method, such as cross-validation or bootstrap methods. In the examples in Section 4, we use a test set of points, independent of the design set.

The most common way of estimating the AUC is to produce an explicit plot of the ROC curve from the test set and estimate the area using quadrature. Let n_0 be the number of points in the test set which belong to class 0, and n_1 be the number which belong to class 1. Assuming that there are no ties in the estimated probabilities, then the ROC curve is a step

function, moving $1/n_0$ units to the right whenever the threshold t above becomes equal to $\hat{p}(x)$ for a member of class 0 and $1/n_1$ units up when t becomes equal to $\hat{p}(x)$ for a member of class 1. Sometimes a smoother curve is plotted, either by taking several points at a time (so that the horizontal and vertical steps of the curve are replaced by diagonal steps) or by explicitly smoothing the curve. If the AUC is estimated from such smoothed curves, then there is a risk of bias being introduced into the estimate of the area under the underlying true ROC curve. For example, the strategy of taking several points at a time will typically lead to underestimating the true AUC. Sometimes, however, the AUC is estimated directly from the basic step function. We return to this below, but first consider a more fundamental perspective.

For an arbitrary point $\hat{p}(x) = t$, the probability that a randomly chosen class 1 point will have a $\hat{p}(x)$ smaller than t is $G(t)$. Suppose that t is chosen randomly according to the distribution F. That is, t is the value of $\hat{p}(x)$ for points randomly chosen from the distribution of class 0 points. Then the probability that the randomly chosen class 1 point will have a *smaller* value of $\hat{p}(x)$ than the randomly chosen class 0 point is $\int G(u) f(u)\, du$.

However, from the definition of the ROC curve, we see that the area under this curve, the AUC, is $\int G(u)\, dF(u) = \int G(u) f(u)\, du$. Thus the AUC is equivalent to the probability that a randomly chosen member of class 1 will have a smaller estimated probability of belonging to class 0 than a randomly chosen member of class 0.

To obtain an estimate of AUC we replace the theoretical functions G and f in this integral by the observed distributions of sample values. Thus, let $f_i = \hat{p}(x_i)$ be the estimated probability of belonging to class 0 for the ith class 0 point from the test set, for $i = 1, \ldots, n_0$. Define $g_i = \hat{p}(x_i)$ similarly for the n_1 test set points which belong to class 1. Then $\{g_1, \ldots, g_{n_1}\}$ and $\{f_1, \ldots, f_{n_0}\}$ are samples from the g and f distributions, respectively.

Rank the combined set of values $\{g_1, \ldots, g_{n_1}, f_1, \ldots, f_{n_0}\}$ in increasing order. Let r_i be the rank of the ith class 0 test set point. There are $(r_i - i)$ class 1 test points with estimated probabilities of belonging to class 0 which are smaller than that of the ith class 0 test point. Summing over the class 0 test points, we see that the total number of pairs of points, one from class 0 and one from class 1, in which the class 1 point has smaller estimated probability of belonging to class 0 than does the class 1 point, is

$$\sum_{i=1}^{n_0} (r_i - i) = \sum r_i - \sum i = S_0 - n_0(n_0 + 1)/2 \tag{2}$$

where S_0 is the sum of the ranks of the class 0 test points. Since there are $n_0 n_1$ such pairs of points altogether, our estimate of the probability that a randomly chosen class 1 point has a lower estimated probability of belonging to class 0 than a randomly chosen class 0 point is

$$\hat{A} = \frac{S_0 - n_0(n_0 + 1)/2}{n_0 n_1} \tag{3}$$

It is not difficult to see that the area under the raw step function ROC described above is composed of elements of area $(r_i - i)/n_0 n_1$, with one element for each point from class 0. Adding these over the class 0 points leads to $\hat{A}$ as above.

This provides a very straightforward way of estimating the AUC, and one which is immune to errors introduced by smoothing procedures. $\hat{A}$ is equivalent to the test statistic used in the Mann-Whitney-Wilcoxon two sample test, thus demonstrating the equivalence of the AUC and Gini coefficient to this test statistic.

The important point to note here is that no mention of a threshold has been made. The measure $\hat{A}$ is an overall measure of how well separated are the estimated distributions of $\hat{p}(x)$ for class 0 and class 1.

Further insight into this is obtained from the following. Suppose we put a threshold at a position $\hat{p}(x) = t$, classifying all points with estimated probability (of belonging to class 0) below t as class 1 and all those with estimated probability above t as class 0. Let class 0 have prior π_0 and class 1 have prior π_1. Then the Bayes correct rate (=1 − Bayes error rate) is

$$r_B = \pi_1 \int_0^t g(x)\,dx + \pi_0 \int_t^1 f(x)\,dx. \tag{4}$$

If we now integrate this over the entire range of t, weighting the integral according to the mixture distribution of the two classes, $\pi_1 g(t) + \pi_0 f(t)$, we obtain

$$\begin{aligned} r &= \int_0^1 \left\{ \pi_1 \int_0^t g(x)\,dx + \pi_0 \int_t^1 f(x)\,dx \right\} \{\pi_1 g(t) + \pi_0 f(t)\}\,dt \\ &= \frac{1}{2}\left(\pi_1^2 + \pi_0^2\right) + 2\pi_1\pi_0 \text{ AUC} \end{aligned} \tag{5}$$

Thus the measure r, which is an average of the Bayes correct rates over all possible values of the classification threshold, is linearly related to the AUC measure. Put another way, if we randomly chose the threshold from the overall mixture distribution $\pi_1 g(t) + \pi_0 f(t)$ then the expected proportion correctly classified by the rule is given by r, a linear function of the AUC.

The very advantage of the AUC as a measure of the performance of a classification rule, that it is independent of choice of classification threshold, can sometimes be a disadvantage when comparing rules. In particular, if two ROC curves cross each other, then (in general—there are pathological exceptions) one will be superior for some values of the classification threshold and the other will be superior for other values of the classification threshold (see Hand, 1997 for details). The AUC is a global measure and it fails to take account of this. Based on this observation, Provost and Fawcett (1997, 1998) and Scott (1999) have independently suggested using the convex hull of a set of ROC curves for comparing the performance of a set of rules and defined *randomised rules*, which involve choosing randomly between rules which are both optimal at certain threshold values. Despite this, the AUC can be useful and is very widely used: one should bear in mind that, as discussed in detail in Hand (1997), there is no single perfect numerical performance measure for rule comparison. Adams and Hand (1999, 2000) describe another approach to tackling the problem of crossing ROC curves, based on taking a weighted sum of classification rules for a range of classification thresholds.

Referring to Hanley and McNeil (1982) we obtain the standard error of $\hat{A}$ to be

$$se(\hat{A}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta}) + (n_0 - 1)(Q_0 - \hat{\theta}^2) + (n_1 - 1)(Q_1 - \hat{\theta}^2)}{n_0 n_1}} \tag{6}$$

with

$$\hat{\theta} = \frac{S_0}{n_0 n_1} \quad \text{and} \quad Q_0 = \frac{1}{6}(2n_0 + 2n_1 + 1)(n_0 + n_1 + 1)(n_0 + n_1) - Q_1$$

where

$$Q_1 = \sum_{j=1}^{n_0} (r_j - 1)^2.$$

Alternatively, and we recommend this for the extension to more than two classes below, one might use a bootstrap method to estimate the standard error.

3. Multiple class extensions

Broadly speaking, there are two kinds of approaches to choosing multiple class performance measures. They differ according to whether or not the other classes are taken account of when computing the difference between classes in each pair. Take overall error rate as an example. The standard measure of this is obtained by dividing the sum of the off-diagonal elements of the confusion matrix by the total number of test points. In this calculation, each point is classified into the class which has the largest estimated probability of class membership *over all classes*. Thus, even though each cell of the confusion matrix makes reference to only two classes (it only includes class i points which are misclassified as class j), this is in the context of estimated probabilities for all classes. An alternative kind of confusion matrix could be defined in which each cell of the matrix was based solely on whether the estimated probability of belonging to class i was larger than the estimated probability of belonging to class j. This would lead to a measure which averaged the pair-wise comparisons of the classes. For error rate, the first strategy seems more natural—it yields a measure of the overall misclassification rate of the multiclass rule. However, for measures such as AUC, which avoid choosing the largest estimated probability at each x (since this requires specification of the priors and implicitly the costs), the second strategy seems more natural. It yields an overall measure of how well each class is separated from all of the others. The first approach is only feasible if priors can be specified for the class sizes. Without this, different choices of prior could lead to different estimated distributions dominating at any given $\hat{p}(x)$. The second approach sidesteps this necessity. For this reason, this is the method we have adopted. This second kind of confusion matrix also allows one to identify the fact that certain pairs of classes or certain pairs of groups of classes can be well separated, even if, overall, the classes cannot be well separated. To illustrate, the overall proportion correct of a diagnostic rule may be small, but if the rule can accurately assign

patients to types of disease (these being *groups* of similar disease classes) then the rule may be very valuable. Techniques for identifying rules which do well in separating groups of classes are described in Taylor and Hand (1999).

Suppose that the multiple classes have been labelled $0, 1, 2, \ldots, c-1$ ($c > 2$), with the order of the labels not reflecting any intrinsic order to the classes. Our classification rule will give us estimates of the probability that each test point belongs to each class $\hat{p}(i \mid x)$ for $i = 0, \ldots, c-1$. For any pair of classes i and j, we can compute the measure $\hat{A}$ using either $\hat{p}(i \mid x)$ or $\hat{p}(j \mid x)$. We therefore define $\hat{A}(i \mid j)$ as the probability that a randomly drawn member of class j will have a lower estimated probability of belonging to class i than a randomly drawn member of class i. Note that the first index serves as the base of the estimated probability, so that $\hat{A}(j \mid i)$ is the probability that a randomly drawn member of class i will have a lower estimated probability of belonging to class j than a randomly drawn member of class j. This means that, in the two class case, $\hat{A}(0 \mid 1) = \hat{A}(1 \mid 0)$. In general, however, for more than two classes $\hat{A}(i \mid j) \neq \hat{A}(j \mid i)$. Moreover, since any one-to-one transformation of the class labels preserves their uniqueness and yields an equally legitimate set of labels, we are unable to distinguish between $\hat{A}(i \mid j)$ and $\hat{A}(j \mid i)$. We tackle both problems by adopting $\hat{A}(i, j) = [\hat{A}(i \mid j) + \hat{A}(j \mid i)]/2$ as the measure of separability between classes i and j.

The overall performance of the classification rule in separating the c classes is then the average of this over all pairs of classes:

$$M = \frac{2}{c(c-1)} \sum_{i<j} \hat{A}(i, j) \tag{7}$$

Deriving an expression for the standard deviation of this multiclass measure is difficult because of the relationships between the $\hat{A}(i \mid j)$ terms. We therefore recommend the use of bootstrap resampling methods (Efron & Tibshirani, 1993).

4. Properties of M

Classification rules often yield estimated probabilities, which are, of course, bounded by 0 and 1. However, one of the attractive properties of M (and, of course, AUC) is that it is invariant to monotonic transformations of the estimated probabilities (which is why the methods we are describing are not restricted to methods which yield probability estimates, but also work for simple class membership 'scores'). We take advantage of this fact, in order to gain a feel for the properties of the M measure, by transforming the interval [0, 1] to the real line, and assuming that the transformed estimated probabilities of belonging to class 0 are normally distributed for each of the classes. This may seem artificial, but no alternative is obviously more realistic—the untransformed distributions on the interval [0, 1] will depend on the particular classification rule employed and on the properties of the process producing the data, so that no generally valid distributions can be given. In particular, to gain a feel for the behaviour of M, we will take the special case in which the c classes are distributed as $N(i\mu, 1)$, for $i = 0, \ldots, c-1$, that is, univariate normal distributions with unit variance and means which are integer multiples of μ. Note that the

assumption of unit variance implies no loss of generality for the case in which the classes have equal variance, since it is the ratio of mean to standard deviation which is relevant to determining separability.

For the situation in which $c = 2$, so that the means are separated by μ, we have

$$M = \hat{A}(0, 1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2} \int_{-\infty}^{x} e^{-y^2/2}\, dy\, dx = \Phi\left(\mu/\sqrt{2}\right) \tag{8}$$

where $\Phi(z)$ is the cumulative standard normal distribution.

For c classes, applying this result to pairs of classes with means arranged as indicated above yields

$$M = \frac{2}{c(c-1)}\left[(c-1)\Phi\left(\frac{\mu}{\sqrt{2}}\right) + (c-2)\Phi\left(\frac{2\mu}{\sqrt{2}}\right) + (c-3)\Phi\left(\frac{3\mu}{\sqrt{2}}\right) + \cdots + \Phi\left(\frac{(c-1)\mu}{\sqrt{2}}\right)\right] \tag{9}$$

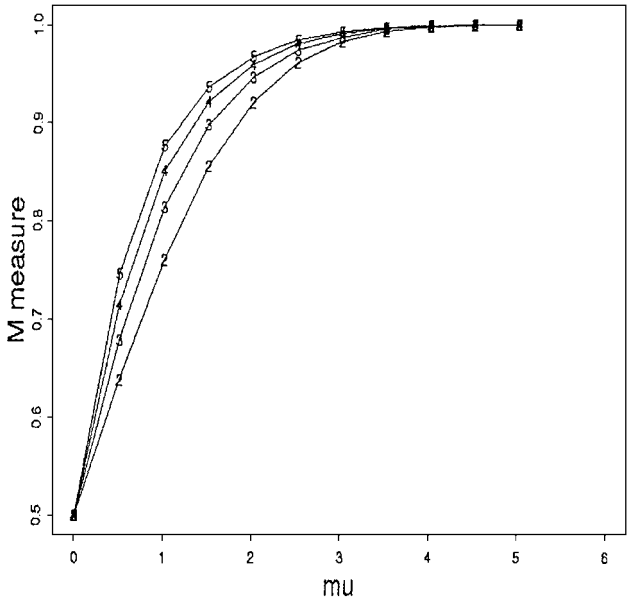
Figure 1(a) shows a plot of M (vertical axis) against μ for $c = 2, \ldots, 5$ (from bottom curve to top curve). As one would wish in a measure of separability, M increases with increasing μ. When all classes coincide, for any pair of classes, i and j, the probability that a randomly chosen point from class i will have a lower value than a randomly chosen point from class j is $1/2$. That is, $\hat{A}(i, j)$ takes the value $1/2$ for all pairs. Thus $M = 1/2$, as we observe in figure 1(a). For fixed μ, as the number of classes increases so M increases. One can also see from the figure that the rate of increase of M with μ is greater for larger c than for smaller c when μ is small. One striking thing about this figure is the steep slope of the curves for small μ. This means that a slight increase in μ will yield a quite substantial improvement in separability. In terms of choosing between rules, this means that a rule which has slightly better pairwise separability between the classes has substantially greater M score.

As we pointed out in Section 1, the AUC is an alternative measure to error rate. It has distinct properties (in particular, not requiring the specification of costs and a classification threshold). This means that the AUC and our M measure should not be regarded as competitors to error rate, but as qualitatively distinct measures, tapping into different aspects of performance. Nevertheless, it will help provide a feel for the properties of the generalised AUC if we relate its behaviour to error rate. Despite the fact that they measure different aspects of performance, it will be reassuring if they are not completely different. In fact, it is more convenient to relate its behaviour to the proportion correctly classified, rather than the proportion incorrectly classified, simply because both M and proportion correct increase with increasing separation between the classes (whereas error rate decreases).

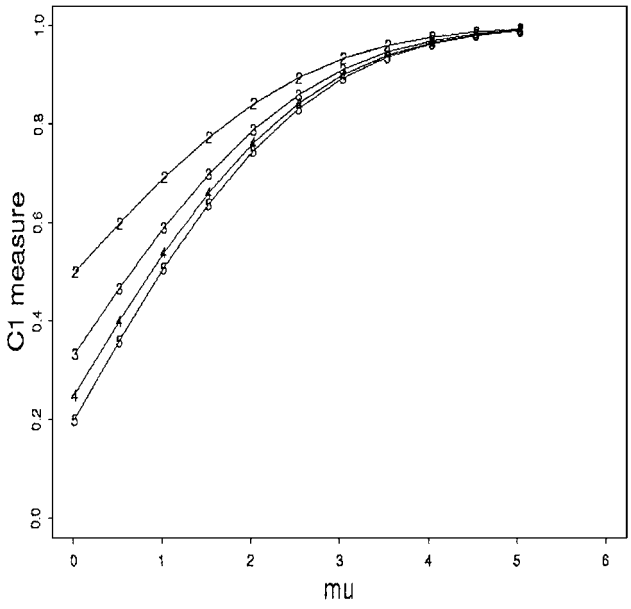
The Bayes proportion correct for the distributions described above is

$$C_1 = \frac{2(c-1)}{c}\Phi\left(\frac{\mu}{2}\right) - 1 + \frac{2}{c} \tag{10}$$

Figure 1(b) shows the plots of these curves corresponding to the curves of M in figure 1(a). Here the case of 2 classes corresponds to the top curve, and 5 classes to the bottom curve, the



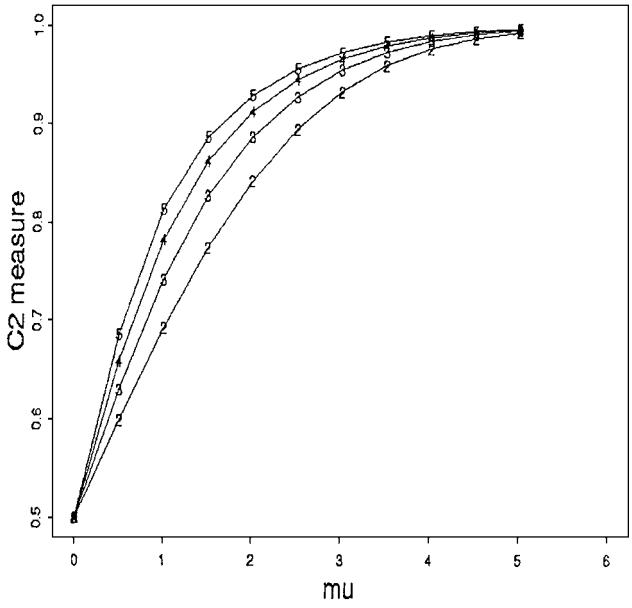
(a)



(b)

Figure 1. The *M* performance measure defined in this paper, and the two proportion correct measures, plotted against μ for 2 to 5 classes.

(*Continued on next page.*)



(c)

Figure 1. (*Continued*).

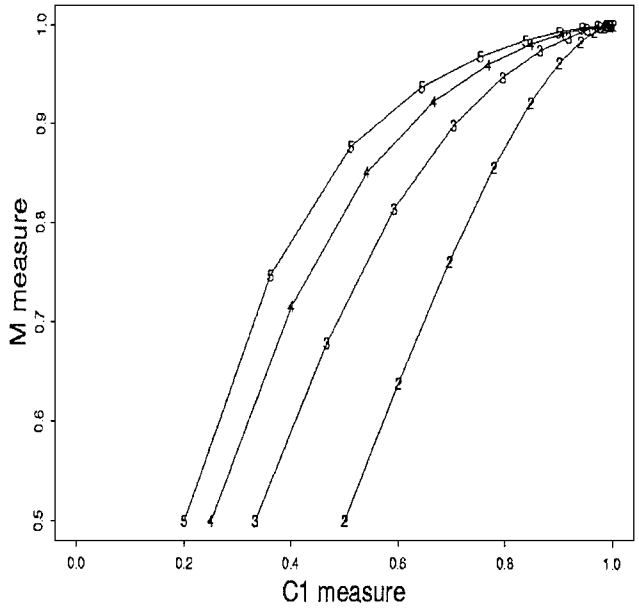
opposite of the M case. Note also that, when there are c classes with equal priors, random classification of points to classes leads to a Bayes proportion correct of $1/c$, as observed in the figure—not to the same value for all c, as is the case for the M measure.

In Section 3, we pointed out that there were two broad ways of defining multiclass extensions for two-class separability measures. The standard measure of proportion correct, as used in C_1, assigns classes according to the dominant estimated probability. M, however, combines the pairwise comparison of classes into a global measure. It is possible to define a measure of proportion correct in a way analogous to that of the M measure. To do this one simply averages over the proportions correct obtained by comparing classes a pair at a time—that is, obtained from the 'pairwise confusion matrix' described in Section 3. Adopting this approach leads to the measure

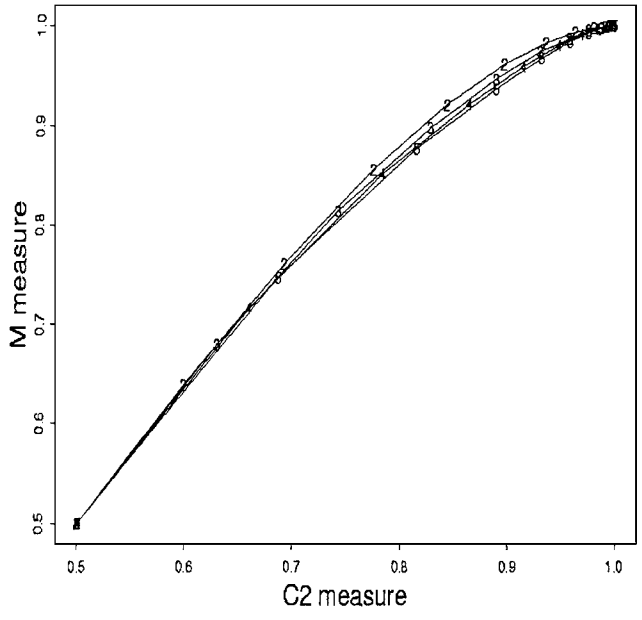
$$C_2 = \frac{2}{c}\left[(c-1)\Phi\left(\frac{\mu}{2}\right) + (c-2)\Phi\left(\frac{2\mu}{2}\right) + \cdots (c-k)\Phi\left(\frac{k\mu}{2}\right) + \cdots \Phi\left(\frac{(c-1)\mu}{2}\right)\right] \tag{11}$$

for the distributions above.

Figure 1(c) shows plots of curves of C_2 against μ corresponding to those of figure 1(a) and (b). The similarity to those of figure 1(a) is striking, including the order of the curves, although the curves for M are steeper. The close relationship between the M and C_2 measures is revealed in figure 2(b). They differ most at large values. The similarity between these



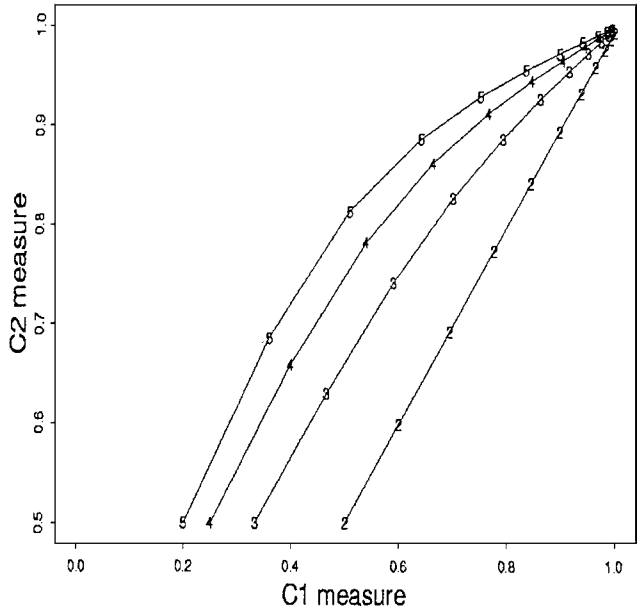
(a)



(b)

Figure 2. Pairwise comparisons of the M measure and the two proportion correct measures for 2 to 5 classes.

(*Continued on next page.*)



(c)

Figure 2. (*Continued*).

measures explains why figure 2(a) and (c) are similar. These show M and C_2, respectively, plotted against C_1 . From figure 2(a), (b), and (c), we see that the main difference (for the special case we have looked at here) between the measures arises from the pairwise comparisons used to define the confusion matrix, rather than the distinction between error rate and AUC. A key reason for this is that we have taken equal priors—equal sizes—for the classes. The M measure, being based on the AUC, is independent of class priors, so that the curves in figure 1(a) would remain the same if the priors were changed. Indeed, this is one of the important properties of the AUC, a property which is attractive in many situations. This is not the case for the other two measures.

The differential impact on C_2 and M of changing the priors is most simply illustrated by considering the special case of two classes, labelled 0 and 1. Suppose that the distributions of the classes are identical, apart from their locations, with the mean of class 0 being at 0 and the mean of class 1 being at μ. Suppose also that the priors, π_0 and π_1 are such that $\pi_0 > \pi_1$. Then, when μ is large, $C_2 \approx 1$ and $M = 1$, as in figure 2(b), but when $\mu = 0$, $C_2 = \pi_0$ and $M = 0.5$. This means that, when $\pi_0 \gg \pi_1$, the curve corresponding to that in figure 2(b) begins near the lower right hand corner of the square (and ends at the top right hand corner). That is, the two measures give quite different indicators of performance: when the prior for one class is very small, C_2 indicates that a high proportion correct is obtained while M indicates that (when the locations are similar) a randomly chosen member of class 1 has a probability of only about 1/2 of being greater than a randomly chosen member of class 0.

5. Examples

The merit of the M measure lies in the fact that it can be applied to situations involving more than two classes, and that it is a cost-independent measure, in the same way that the AUC is cost-independent. That is, its merits are intrinsic, and are not based on its relative performance compared to any other measure (such as, for example, error rate). Nonetheless, as remarked above, it is helpful to gain a feel for the behaviour of the measure by comparing it to other measures which might be used. To achieve this, we applied M, C_1, and C_2 to eight data sets, with numbers of classes ranging from two to ten, using three qualitatively very different kinds of classification rule, each off-the-shelf functions in Splus (Venables & Ripley, 1994): multiple logistic regression, a k-nearest neighbour rule, and the Splus implementation of the CART recursive partitioning rule (labelled 'tree' in the table)

Table 1. Rank orders of M, C_1, and C_2 on eight data sets using three classification rules.

No. of classes	Data set	Measure	Multiple logistic regression	K-nearest neighbour ($K = 9$)	Tree
2	WDBC	M	1	2	3
		C_1	1	3	2
		C_2	1	3	2
3	TAE	M	1	2	3
		C_1	1	2	3
		C_2	2	1	3
3	IRIS	M	1	2	3
		C_1	1	2.5	2.5
		C_2	1	2	3
4	CAR	M	3	1	2
		C_1	1	2	3
		C_2	3	1.5	1.5
5	PEN	M	2	1	3
		C_1	2	1	3
		C_2	2	1	3
5	GLASS	M	2	3	1
		C_1	2	3	1
		C_2	2	3	1
7	OPTICAL	M	1	2	3
		C_1	2	1	3
		C_2	1	3	2
10	CHRO'SOME	M	1	2	3
		C_1	2	1	3
		C_2	2	1	3

(Breiman et al., 1984). All except the CHROMOSOME data set were obtained from the UCI Repository of Machine Learning Databases (Blake & Merz, 1998). The former is discussed in Piper and Granum (1989). The abbreviations in Table 1 refer to the following data sets: WDBC: Wisconsin diagnostic breast cancer; TAE: Teaching assistant evaluation; IRIS: Iris plants; CAR: Car evaluation; PEN: Pen digits; GLASS: Glass identification; OPTICAL: Optical recognition of handwritten digits; CHROMOSOME: Chromosome data. In these comparisons, we used a simple split into roughly equal sized training and test sets, with, of course, the same split being used for each classification rule and each measure. As we noted above, in practical application of any of the measures, one would normally use a more sophisticated estimation method, typically based on sample reuse (such as the 632+ bootstrap in error rate estimation (Efron & Tibshirani, 1995)). We chose not to do this here for practical simplicity—our aim here is not to investigate how to estimate the different measures.

The eight data sets were chosen to span a range of situations. The minimum and maximum values for M over the eight data sets were 0.524 and 0.991, for C_1 they were 0.237 and 0.971, and for C_2 they were 0.491 and 0.970.

The rows of Table 1 show the ranks each measure gives to each of the three classification rules. What is interesting is that, in four out of the eight data sets, measure M gives a different rank ordering to the classifiers than does C_1, and in four of the eight data sets, measure M also gives a different rank ordering than does C_2. That is, although the M and C measures are both measuring performance, the different aspects are such that they will sometimes favour different rules.

6. Conclusion

Misclassification rate is often a poor criterion by which to assess the performance of classification rules. It makes the implicit assumption that the costs of the different kinds of misclassification are equal, an assumption which is rarely justified in practice. For the two class case it is sometimes possible to provide more appropriate cost values, though this is typically difficult. When there are more than two classes, it is extremely difficult, and usually impossible, to provide realistic assessments of the relative severity of the different kinds of misclassification which can arise.

Because of these problems, one frequently adopted measure of performance for the two class case is the area under the ROC curve. This is equivalent to the probability that a randomly chosen member of one class has a smaller estimated probability of belonging to the other class than has a randomly chosen member of the other class, and so is a natural measure of separability between the two estimated probability distributions. This measure has the singular merit that it is independent of costs, priors, or (consequently) any classification threshold. There is a huge literature associated with ROC curves, spread across medicine, statistics, chemistry, psychology, and, most recently, machine learning and data mining.

When expressed in the form of the area under the ROC curve, the measure has no obvious generalisation to multiple classes. However, when expressed in its equivalent probabilistic form, it has a straightforward generalisation obtained by aggregation over all pairs of

classes. This reduces to the standard form in the case of two classes, and so is a natural generalisation.

We showed, in an artificial situation, that the generalised form can behave in a way similar to standard measures such as error rate, but the fact that it is independent of costs and priors means that it is tapping different aspects of performance, and so will not always behave in the same way. To illustrate this, we evaluated the performance of three classification methods on eight real data sets, using the generalised measure developed above, and also two versions of error rate. In half of the cases that we investigated (which were chosen purely on the basis of the number of classes), the generalised measure produced a different rank order for the three classification methods.

Acknowledgments

We would like to express our appreciation to three anonymous referees, who provided detailed and extensive reports on earlier versions of this paper. One of the authors (DJH), as a former Editor of the *Journal of the Royal Statistical Society, Series C* and as Editor-in-Chief of *Statistics & Computing*, would like to add that these reports were exemplary (even if he did not necessarily agree with everything in them!). RJT was supported during this work by research grant GR/N08704 from the UK Engineering and Physical Sciences Research Council.

References

Adams, N. M. & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition, 32*, 1139–1147.

Adams, N. M. & Hand, D. J. (2000). Improving the practice of classifier performance assessment. *Neural Computation, 12*, 305–311.

Blake, C. & Merz, C. J. (1998). *UCI Repository of Machine Learning Databases*. Irvine, CA: University of California, Department of Information and Computer Science. [www.ics.uci.edu/∼mlearn/MLRepository.html].

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*, 1145–1159.

Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to The bootstrap*. London: Chapman & Hall.

Efron, B. & Tibshirani, R. J. (1995). Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical Report 176. Stanford, CA: Stanford University, Department of Statistics.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Chichester: Wiley.

Hand, D. J. (2000). Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica, 53*, 1–14.

Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Piper, J. & Granum, E. (1989). On fully automatic feature measurement for banded chromosome classification. *Cytometry, 10,* 1–14.

Provost, F. & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 43–48). Menlo Park, CA: AAAI Press.

Provost, F. J. & Fawcett, T. (1998). Robust classification systems for imprecise environments. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 706–713). Madison, WI: AAAI Press.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing classifiers. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453). Madison, WI: Morgan Kaufmann.

Scott, M. (1999). Parcel: Feature selection in variable cost domains. Doctoral Dissertation, Engineering Department, Cambridge University, UK.

Taylor, P. C. & Hand, D. J. (1999). Finding superclassifications with acceptable misclassification rates. *Journal of Applied Statistics, 26*, 579–590.

Turney, P. (1996). Cost sensitive learning bibliography. [www.iit.nrc.ca/bibliographies/cost-sensitive.html]

Venables, W. N. & Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

Zweig, M. H. & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots. *Clinical Chemistry, 29*, 561–577.

Received August 24, 1999
Revised July 7, 2000
Accepted February 23, 2001
Final manuscript June 18, 2001