Multiclass ROC Analysis

MATTHEW S. WANDISHIN AND STEVEN J. MULLEN

Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona

(Manuscript received 23 January 2008, in final form 11 September 2008)

ABSTRACT

Receiver operating characteristic (ROC) curves have become a common analysis tool for evaluating forecast discrimination: the ability of a forecast system to distinguish between events and nonevents. As is implicit in that statement, application of the ROC curve is limited to forecasts involving only two possible outcomes, such as rain and no rain. However, many forecast scenarios exist for which there are multiple possible outcomes, such as rain, snow, and freezing rain. An extension of the ROC curve to multiclass forecast problems is explored. The full extension involves high-dimensional hypersurfaces that cannot be visualized and that present other problems. Therefore, several different approximations to the full extension are introduced using both artificial and actual forecast datasets. These approximations range from sets of simple two-class ROC curves to sets of three-dimensional ROC surfaces. No single approximation is superior for all forecast problems; thus, the specific aims in evaluating the forecast must be considered.

1. Introduction

The receiver operating characteristic (ROC) curve has become increasingly popular as a measurement of forecast discrimination, that is, the ability of a forecast system to distinguish between an event and nonevent. The ROC is based on signal detection theory in which the event is the signal and the nonevent is the noise. The upper-left panel in Fig. 1 shows a hypothetical signal (p2) and noise (p1) distributions. It is typically assumed that p1 and p2 are Gaussian distributions, as pictured in Fig. 1, but oftentimes this is not the case. For example, frequently the x axis in Fig. 1 (sometimes referred to as the evidence or decision variable) is a probability, such as the output from an ensemble forecast, and thus bounded in the region [0,1]. Fortunately, the subsequent ROC analysis is not sensitive to deviations from normality (Mason 1982), in large part because the ROC curve is unchanged for any monotonic transformation of the evidence variable.1

A transformation is monotonic if it preserves the ranking of the event and nonevent samples; that is, f is monotonic if for each $x_2 > x_1$, $f(x_2) > f(x_1)$. Adjusting the tuner (or decision threshold; the vertical line in the top panels in Fig. 1) to the left will increase the signal, but also increase the noise. In forecasting terminology, an increase in the probability of detection (POD; the area beneath p2 to the right of the threshold) comes at the expense of an increase in the probability of false detection (POFD, i.e., false alarms; the area beneath p1 to the right of the threshold). The ROC curve is drawn by moving the threshold from left to right, creating a series of (POFD, POD) pairs that run from the upper-right corner (all cases are classified as events) to the lower-left corner (all cases are classified as nonevents) of the ROC graph (Fig. 1, bottom-left panel). The ROC curve thus provides a quantitative estimate of the probabilities of possible forecast outcomes (namely, the POD and POFD) for different decision thresholds and of the trade-offs between these outcomes as the decision threshold varies (i.e., how many more false alarms must one endure in order to increase the number of events captured?) (Mason 2003). The area under this curve (AUC) is then a scalar measure of forecast discrimination.

There is, however, another way of conceiving of the AUC implicit in the discussion above about the necessity of transformations to preserve the ranking of the samples. Bamber (1975) noted that the AUC is equivalent

1 The insensitivity of the ROC curve to monotonic transformations of the evidence variable is implicit in statements in Harvey et al. (1992, p. 874) and Mason and Graham (2002, p. 2160) and stated explicitly in Kupinski et al. (2001, p. 887) and Nakas and Yiannoutsos (2004, p. 3440).

Corresponding author address: Matthew Wandishin, NSSL, National Weather Center, 120 David L. Boren Blvd., Norman, OK 73072. E-mail: matt.wandishin@noaa.gov

DOI: 10.1175/2008WAF2222119.1

© 2009 American Meteorological Society
to the Mann–Whitney U statistic (see also Mason and Graham 2002; Nakas and Yiannoutsos 2004). That is, within the context of Fig. 1, the AUC is the probability that a randomly chosen member of the signal distribution (p2) will be to the right of a randomly chosen member of the noise distribution (p1). [This is also known as a two-alternative forced-choice (2AFC) test; the AUC is equivalent to the probability of correctly distinguishing an event from a nonevent in a 2AFC test (Scurfield 1996; Mason and Graham 2002; Mason 2003).] So, the AUC can be calculated by repeated sampling of the event and nonevent cases, tallying the frequency with which the event received a higher score (e.g., forecast probability) than the nonevent. Note that while this method will produce the same AUC as the previous method, it cannot be used to generate a ROC curve and so does not provide information for an individual user to optimize his or her decision making based on the forecast.

Thus, it is seen that ROC analysis is designed for binary situations, event and nonevent, and this is suitable for many meteorological forecasts. Nevertheless, there are forecast problems that move beyond this two-class framework, such as National Weather Service forecasts of precipitation type or Climate Prediction Center (CPC) seasonal forecasts of temperature and precipitation. An extension of the ROC approach to multiple classes is needed.

Of course, ROC analysis by itself is an incomplete measure of forecast performance. As many authors have noted (e.g., Harvey et al. 1992; Mason and Graham 2002), ROC curves based on probability forecasts are independent of the reliability of those probabilities. Indeed, this is to be expected given that the evidence variable need not be a probability even, as noted above. When probabilities are used, ROC analysis treats them only as ordinal measures of forecaster confidence (Harvey et al. 1992). While some authors cite this property as a shortcoming of ROC analysis (e.g., Glahn 2004), it can instead be viewed as the desirable trait of an evaluation tool designed to complement a measure of reliability (e.g., Mason 2003). Murphy and Winkler (1987) propose a verification framework within which the joint probabilities of the forecasts and observations are factored into separate measures that are concerned with different attributes of forecasts and observations: the calibration–refinement factorization that is conditioned on the forecasts and the likelihood–base rate factorization that is conditioned on the observations. Taken together, ROC curves and reliability diagrams (Wilks 2000) constitute such a complementary pair of measures.
Section 2 introduces an extension of ROC analysis for multiclass forecast problems that can be modeled as one-dimensional probability distributions, for example, when the forecast decision is based upon a single forecast score. Section 3 generalizes this extension to multiclass forecast problems that are modeled by multidimensional probability distributions, such as multiclass probability forecasts. Extensions that maintain the form of two-class ROC analysis by casting the multidimensional decision process as a set of two-class decisions are presented in section 4, followed by a brief summary.

2. Scalar decision variable

As indicated above, each point of the two-class ROC curve fully determines the $2 \times 2$ contingency table, also known as the confusion matrix (Fawcett 2003), associated with each threshold. If we normalize each column of the contingency table, then the POD is simply the diagonal element of the first column and the POFD is the off-diagonal element of the second column (Table 1). However, just as validly, the diagonal element of the second column could be used instead; the result is simply a reversal of the ordinate of the ROC diagram. That is, the ROC curve becomes a plot of POD against the probability of a null event (see Doswell et al. 1990), or more generally, a plot of the POD of p2 (POD2) as a function of the POD of p1 (POD1). Similarly, the ROC curve could become a plot (not shown) of the two off-diagonal elements (the POFDs, or error rates) in which case the curve is inverted and a perfect set of forecasts would have an $AUC = 0$.

Now consider the addition of a third distribution, p3 (Fig. 2). Using two thresholds to define three classes yields a $3 \times 3$ contingency table (Table 2). For the two-class problem, knowing one element from each column was sufficient, but for the three-class problem, two elements are needed to fully determine the column. Thus, whereas the two-class problem spans two dimensions, the three-class spans, not three, but six dimensions. More generally, $n$ classes result in an $(n^2 - n)$-dimensional problem. Some important concepts tied to the ROC curve extend to the $n$-class problem (Srinivasan 1999), including that optimal classifiers (i.e., those classifiers that maximize value for the decision task) lie on the convex hull of the points representing the error rates (i.e., the off-diagonal elements, the POFDs) of the classifiers being compared (see also Provost and Fawcett 2001). The two-class equivalent to this statement, for the traditional orientation of the ROC curve, is that for the collection of points in ROC space that compose a curve with monotonically decreasing slope (from nearly vertical at the lower-left corner to nearly flat at the upper-right corner), no other possible decision point exists that could provide more value to a forecast user. Unfortunately, Edwards et al. (2005) demonstrate that a perfect classifier and a no-skill classifier will yield ROC surfaces having identical (zero) volumes. In parallel with the two-dimensional problem, the volume under the surface defined exclusively by the error rates is zero for a perfect classifier. The no-skill classifier (e.g., random guessing) yields a degenerate curve that fails to span the hyperspace, giving, again, a volume of zero, in the same way that the area under a point is zero. Thus, the volume beneath the $(n^2 - n)$-dimensional ROC surface is all but useless as a performance measure for $n > 2$ classes. Thus, a straightforward extension of the ROC curve to a ROC surface, in which the surface represents the set of optimal decision points, is possible but not practical because of the difficulty in trying to visualize high-dimensional surfaces. Meanwhile, a straightforward use of the extension of the area under the ROC curve to a volume under the ROC surface appears not to be possible. As a result, one must resort to approximations and partial measures, as described below.

One solution to the difficulty presented by high dimensionality is to extend the approach described at the beginning of this section, in which only the diagonal elements of the contingency table are used (e.g., Nakas
and Yiannoutsos 2004). For the three-class problem, this involves the following decision criteria:

$$\text{VUS} = P(X_1 < X_2 < X_3).$$

(1)

Also as before, the unbiased nonparametric estimates of the volume can be calculated using an extension of the Mann–Whitney U statistic (Dreiseitl et al. 2000) or bootstrapping (Nakas and Yiannoutsos 2004). A perfect ordering of the three classes results in a VUS = 1, while random forecasts (i.e., the three distributions p1, p2, and p3 are identical) give a VUS = 1/6. This is the area beneath a plane intersecting the unit cube at the vertices (1, 0, 0), (0, 1, 0), and (0, 0, 1). More intuitively, assuming again that the three distributions are identical, there is a 1/3 chance that the class 3 sample will have a higher score than both the class 1 and class 2 samples; and there is a 1/2 chance that the class 2 sample will have a higher score than the class 1 sample; in other words, there are six possible orderings of the three samples, giving an overall probability of $1/3 \times 1/2 = 1/6$.

Varying the two decision thresholds, c1 and c2, then results in a three-dimensional ROC surface (Fig. 3). Analogous to the two-class problem, the volume under this surface (VUS) equals the probability that classes 1, 2, and 3 are correctly ordered (Scurfield 1996; Nakas and Yiannoutsos 2004)—as in a three-alternative forced-choice test (3AFC); that is,

$$\text{POFD}_{ij} = \text{P}(X_i \text{ classified as } j \text{ but } j \neq i).$$

For the three-class problem, this involves the following decision criteria:

- if $X < c_1$, then decide class 1, else
- if $c_1 < X < c_2$, then decide class 2, else
- if $X > c_2$, then decide class 3.

In this approach, this simplification comes at the cost of lost information; specifically, all information about misclassification errors—the off-diagonal elements of the $3 \times 3$ contingency table—is disregarded. A well-informed forecast user needs to know more than just how often the forecasts are correct. The loss incurred from a class 3 event incorrectly predicted to be a class 1 event might be very different from a class 2 event being incorrectly forecast as a class 1 event. For exemple, imagine three classes: no convection, nonsevere convection, and severe convection. Forecasting no convection on a day that nonsevere thunderstorms occur could have adverse economic impacts on some users, but the same forecast on a severe weather day includes the potential impact of physical harm. Related to the problem of neglecting the misclassification errors, the two-class decision task involves only one degree of freedom—the ROC curve is one dimensional—and so POD2 is completely determined by a choice of POD1, but the three-class problem has $(n^2 - n - 1) = 5$ degrees of freedom and so fixing two of the PODs does not, in general, determine the third (Edwards and Metz 2006).

A broader approach was introduced by Scurfield (1996). Given the task of ordering three samples, one from each class, there are six possibilities—123, 132, 213, 231, 312, and 321; and a ROC surface can be plotted for each. The volume under each surface is then the probability of occurrence for each respective ordering, with the five latter orderings each including misclassification errors. Since the six possibilities listed above are exhaustive and mutually exclusive, it follows that

$$\sum_{\alpha} V_{\alpha(123)} = 1,$$

(2)

where, following Scurfield (1996), $\alpha$ denotes a permutation mapping the set $\{1, 2, 3\}$ onto $\{1, 2, 3\}$ such that if $\alpha(1) = 2, \alpha(2) = 3,$ and $\alpha(3) = 1$, then $\alpha(123) = 231$; and $V_{\alpha(123)}$ is the volume beneath the $\alpha(123)$-ROC surface. For a perfect classifier, then, $V_{123} = 1$, and $V_{\alpha(123)} = 0$, for all other $\alpha$.

Conveniently, the areas beneath the ROC curves obtained by treating classes 1 and 2, classes 1 and 3, and classes 2 and 3 separately can be calculated directly from the ROC volumes (Scurfield 1996). Recall that the AUC is equal to the probability that the “yes” and “no” events are properly ordered; that is, $\text{AUC}_{12} = P(X_1 < X_2)$. Now, there are three possibilities for samples from

![FIG. 3. The three-dimensional ROC surface resulting from varying the decision thresholds c1 and c2 in Fig. 2.](image-url)
classes 1, 2, and 3 to be arranged such that \( X_1 < X_2 \), namely, either \( X_1 < X_2 < X_3 \), or \( X_1 < X_3 < X_2 \), or \( X_3 < X_1 < X_2 \). Therefore,

\[
P(X_1 < X_2) = P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2)
\]

which is to say,

\[
AUC_{12} = V_{123} + V_{132} + V_{312},
\]

and similarly for \( AUC_{13} \) and \( AUC_{23} \).

a. Example with artificial data

Now suppose the three ordered distributions in Fig. 2 are Gaussians such that \( p_1 = N(-1, 1) \), \( p_2 = N(0, 1) \), and \( p_3 = N(1, 1) \). Sliding the two thresholds as described above, and plotting the relevant elements of the \( 3 \times 3 \) contingency table produces the six ROC surfaces shown in Fig. 4 [e.g., the 132-ROC surface, upper right, is obtained by plotting \( P_{11} \) versus \( P_{23} \) versus \( P_{32} \), where \( P_{ij} \) is the probability that an object belonging to class \( j \) is classified as belonging to class \( i \); that is, the \( P_{ij} \)'s are the elements of the \( 3 \times 3 \) contingency table (Table 2), where \( P_{ij} \) is the POD when \( i = j \), and \( P_{ij} \) is the POFD when \( i \neq j \)]. Note that the intersection of the ROC surface with the sides of the unit cube is the two-class ROC curve for the classes composing the axes of the face. For example, for the 123-ROC surface (Fig. 4, top-left panel), the intersections show the \( A_{12} \) (right face), \( A_{23} \) (left face), and \( A_{13} \) (bottom face) ROC curves. Note also that the area beneath the \( A_{13} \) curve is bigger than those for the other two. Thus, because the means of \( p_1 \) and \( p_2 \) distributions are separated by a single standard deviation, as are the means of \( p_2 \) and \( p_3 \), while the means of \( p_1 \) and \( p_3 \) are separated by two standard deviations. Similarly, observe the ordering of volumes: \( V_{123} \), where each class is correctly identified, has the largest volume, as expected; followed by \( V_{132} \) and \( V_{213} \), each of which have two correctly ordered pairs and one misordered pair (i.e., for \( V_{132} \), \( 1 < 3 \) and \( 1 < 2 \), but \( 2 > 3 \)); followed by \( V_{231} \) and \( V_{312} \), each of which has only one correctly ordered pair; and finally \( V_{321} \), for which the ordering is completely reversed. (The correct or incorrect ordering of pairs of classes is manifested in the nature of the ROC curve on the faces of the unit cube, such that, for example, the 132-ROC surface has two convex curves and one concave curve, while all three curves are concave for the 321-ROC surface.) Of course, the fact that \( V_{132} = V_{213} \) and \( V_{231} = V_{312} \) is an artifact of the symmetry of these data. In general, these equalities will not hold.

Applying Eq. (3) to the volumes shown in Fig. 4 gives

\[
AUC_{12} = V_{123} + V_{132} + V_{312}
\]

\[
AUC_{12} = 0.53 + 0.20 + 0.03 = 0.76,
\]

\[
AUC_{23} = V_{123} + V_{213} + V_{231}
\]

\[
AUC_{23} = 0.53 + 0.19 + 0.03 = 0.75,
\]

\[
AUC_{13} = V_{123} + V_{312} + V_{213}
\]

\[
AUC_{13} = 0.53 + 0.20 + 0.19 = 0.92,
\]

which are exactly the areas expected by Gaussians separated by one and two standard deviations (see Marzban 2004).

b. Example with CPC seasonal temperature forecasts

Each month the National Weather Service’s Climate Prediction Center issues seasonal (i.e., 3 month) temperature and precipitation forecasts with lead times from 0.5 to 12.5 months (available online at http://www.cpc.noaa.gov/products/predictions/90day/). The half-month lead time is because the forecasts are issued on the third Thursday of each month. The forecasts consist of probability triplets representing the probability that the mean temperature (or total precipitation) will be below, near, or above normal, where these classes are defined by splitting the 1961–90 climatological distribution into thirds. The forecasts focus on the most likely class, with the difference between the probability for that class and the climatological probability (33%) subtracted from the opposite class. For example, a forecast for a 20% increase in above normal precipitation implies the probability triplet (53%, 33%, 13%) for the above-, near-, and below-normal classes, respectively. As a result of the focus on the most likely class, though the forecasts are presented as three separate probabilities, in essence the evidence is unidimensional. Therefore, the forecasts are transformed from probability triplets to a forecast score according to \( f = \frac{100[p(above) - p(below)]}{100} \); the forecast scores all fall between \(-700\) and \(700\).

Seasonal forecasts are evaluated for forecasts issued from December 1994 through July 2004 for all of the 102 climate divisions of the contiguous United States. A large percentage of the forecasts, 45% for the 0.5-month lead time, are simply the unaltered climatological probabilities with a forecast score \( f = 0 \). These forecasts have been eliminated from the subsequent analysis, so that the results should be considered conditioned on a nonzero forecast. As expected, the effect is to improve the performance of the forecasts since these are essentially nonforecasts; a forecast user has gained no additional information by consulting the forecast. This is not...
to say that the forecasts are completely useless; they may accurately represent that the current forecast is unpredictable. However, the presence of these non-forecasts clouds the measurement of the performance of the rest of the forecast set. It should be noted, as well, that the CPC does allow for an increased probability of near-normal conditions, for which the probabilities are reduced equally for the below- and above-normal classes. The forecast score defined above does not distinguish these forecasts from the unaltered climatological probabilities (33%, 33%, and 33%). These forecasts are retained but all carry the same zero forecast score regardless of the actual near normal class probability. This step is necessary to retain the one-dimensional character of the forecast score, but the effect is to underestimate the forecasts’ ability to distinguish the near normal class from the others. However, the maximum increased probability for the near normal class is just 10% (i.e., 28%, 43%, and 28%) compared to ~50% for the other classes; that is, little confidence is ever expressed in a forecast of the near normal class, and these forecasts compose a small fraction (~2% for the 0.5-month lead time) of the total number of forecasts. Also, it has previously been noted that the forecasts with an increased probability for near normal temperatures possess little skill (Wilks 2000). Thus, ignoring this small subset of the forecasts is not expected to negatively impact the assessment of the CPC forecasts, and the forecast score can be used as a reasonable substitute for the issued probability triplets.

Previous evaluations of the CPC seasonal outlooks have either examined skill scores, whether by treating them as simply categorical forecasts (Livezey and Timofeyeva 2008) or using the ranked probability score (Wilks 2000), or focused on the calibration–refinement factorization (see Murphy and Winkler 1987) using

**Fig. 4.** The six α(123)-ROC surfaces where the three classes are represented by the three normal distributions: \(N(-1, 1)\), \(N(0, 1)\), and \(N(1, 1)\). The numbers in the upper-left corner of each panel denote the ordering considered in each plot. Here, \(P_{ii}\) is the probability of detection for class \(i\) (POD\(_i\)), and \(P_{ij}\) is the probability of incorrectly classifying an observation from class \(j\) as belonging to class \(i\) (POFD\(_{ij}\)). The numbers in the upper-right corner of the plots is the volume under the ROC surface.
reliability diagrams (Wilks 2000). The likelihood–base rate factorization (Murphy and Winkler 1987), in the form of ROC curves, has been used to evaluate three-category seasonal probability forecasts similar to those examined here (Mason and Graham 1999; Kharin and Zwiers 2003), but only by essentially considering the outlooks as a set of two-class forecasts.

An examination of the distribution of forecast scores given the subsequent observed class for the 0.5-month lead-time temperature forecasts (Fig. 5) reveals little separation of the means, particularly between the below- and near-normal classes. This is revealed by the ROC surfaces, as well (Fig. 6). Note the similar appearance of each of the six surfaces and the small spread in the corresponding volumes beneath the surfaces. The volume for the correctly classified 123-ROC surface is only marginally better than the no-skill volume (0.25 versus 0.167). Also, the fact that the below- and near-normal class distributions are nearly identical is evident in the nearly diagonal curve on the right faces of the 123- and 213-ROC surfaces, or alternatively, in the nearly equal volumes beneath those two surfaces. Recall that, for this set of forecast data, the 213-ROC surface represents those forecasts that correctly classify the above-normal events but swap the below- and near-normal events. Hence, the CPC forecasts have some marginal ability to distinguish the above-normal events from the others, but cannot discriminate between near- and below-normal events.

It has been demonstrated that the skill in the seasonal temperature forecasts is limited to (imperfectly) capturing long-term warming trends and, for the winter months, strong El Niño events (Livezey and Timofeyeva 2008). In other words, the skill in these forecasts is confined to forecasts of above-normal temperatures, just as the forecasts’ ability to discriminate among the three events is confined to the above-normal category. The latter finding can be further highlighted by computing the areas under the two-class ROC curves from the volumes, as described at the end of section 2a, yielding $A_{AB} = 0.65$, $A_{AN} = 0.61$, and $A_{BN} = 0.53$. Incidentally, the volume for the correctly classified surface drops steadily with increasing lead time, reaching a low of 0.18 for the 6.5-month lead-time forecasts. This result is in contrast to Livezey and Timofeyeva (2008), who find the skill to be nearly independent of lead time, with the exception of forecasts of El Niño winters. At longer lead times, seasonal temperature forecasts retain marginal skill, but lose the ability to distinguish above-normal events from the other categories. Seasonal precipitation forecasts show little dependence on lead time; volumes for the 123-ROC surfaces are between 0.20 and 0.23 for all lead times.

3. Multidimensional decision variables

In the preceding section, the two-class ROC was extended to three classes simply by adding an additional distribution along the single decision axis (e.g., Fig. 2). However, many forecasts do not lend themselves to the use of a single decision axis. For example, the unidimensional character of the CPC seasonal forecasts was a result of the relatively infrequent adjustment of the near-normal forecast probability; that is, practically speaking, the CPC seasonal forecasts had only a single degree of freedom [i.e., $p(\text{above})$] with the other two probabilities tied to the first. More generally, a three-class forecast problem will have two degrees of freedom, with $p(3) = 1 - p(1) - p(2)$. As a result, the unidimensional decision axis is replaced by a two-dimensional decision space (Scurfield, 1998).

One form of the two-dimensional decision space, introduced by Murphy (1972) and proposed independently by Mossman (1999), plots each forecast probability triplet as a point within the probability triangle. More broadly, the forecasts lie within the probability cube, but since the three probabilities sum to unity, the points are restricted to a triangle connecting the vertices (1,0,0), (0,1,0), and (0,0,1) (Fig. 7), where the vertices represent classes 1, 2, and 3, respectively. Thus, each vertex represents a 100% forecast for that class while each side represents a 0% forecast for the class represented by the opposing vertex. More generally, the probability for each class is equal to the distance from the point to the side opposite that class vertex. A decision rule based on this form of the decision space, but different than that presented in Mossman (1999), is proposed in Lachiche and Flach (2003).
For the more commonly used form of the two-dimensional decision space, consider the goal of maximizing the expected utility. Utility is a measure of the benefit obtained or cost incurred by a forecast user. More specifically, event misclassifications (in 2D, false alarms and missed events) typically cause a user to incur losses: a vendor makes a large quantity of lemonade after a forecast of warm temperatures only to have the product wasted when temperatures remain low. Correct classifications (in 2D, hits and correct null events) typically yield a benefit to a user: a vendor maximizes profit by making extra quantities of lemonade after a correct forecast of record high temperatures. Very simply, one should choose class \( p_i \) when the expected utility of choosing class \( p_i \) is greater than the expected utility of choosing class \( p_j \) for all \( j \neq i \). That is,

\[
E[U(d = p_i | x)] = \max_j \{E[U(d = p_j | x)]\},
\]

where \( x \) represents the multidimensional data used to classify the observations (e.g., radar and satellite data, or numerical model output), \( d \) represents the decision or classification, \( U(d = p_i | x) \) is the utility obtained by choosing class \( \pi_i \) given the evidence \( x \), and \( E[\cdot] \) is the expectation operator. If \( U_{ij} \) is the utility of deciding an object belongs to class \( \pi_i \) when it is actually from class \( \pi_j \) (for \( i \neq j \), \( U_{ij} \) is a misclassification cost), then the expected utility of deciding an object belongs to class \( \pi_i \) when evidence \( x \) is observed is

\[
E[U(d = \pi_i | x)] = \sum_{j=1}^{N} U_{ij} p(t = \pi_j | x),
\]

where \( t \) represents the “truth” and \( p(t = \pi_j | x) \) is the conditional probability that an object belongs to class \( \pi_j \) given the observations \( x \). The expected utility of choosing class \( \pi_i \) is simply the sum of the utilities for each of the \( N \) possible outcomes multiplied by the probability that each outcome occurs. Equation (5) is a straightforward extension to multiple classes of the estimate of forecast value [Thompson and Brier (1955); see also Richardson (2000), which draws the connection between ROC curves and forecast value estimates].
Dividing both sides by \( p(x|t = \pi_i) \) gives

\[
\sum_{j=1}^{N} U_{kjl}p(t = \pi_j)p(x|t = \pi_j) = \max_{k} \left[ \sum_{j=1}^{N} U_{kjl}p(t = \pi_j)p(x|t = \pi_j) \right].
\]

(7)

where

\[
\text{ LR}_{jN} = \frac{p(x|t = \pi_j)}{p(x|t = \pi_N)}
\]

is the likelihood ratio of \( \pi_j \) to \( \pi_N \). Assuming \( N = 3 \) for simplicity, expanding (8), and rearranging the terms gives the decision rule: choose class \( \pi_i \) if and only if

\[
(U_{i1} - U_{i1})p(t = \pi_1)LR_{13} + (U_{i2} - U_{i2})p(t = \pi_2)LR_{23} > (U_{i3} - U_{i3})p(t = \pi_3),
\]

(10)

or

\[
\gamma_{i1}LR_{13} + \gamma_{i2}LR_{23} > \gamma_{i3},
\]

(11)

where \( \gamma_{ijk} = (U_{ik} - U_{jk})p(t = \pi_k) \). Using the identities \( \gamma_{ij} = \gamma_{kij} - \gamma_{ikj} \) and \( \gamma_{ijk} = -\gamma_{jik} \), we then get that the classification regions are defined by the following three decision lines:

\[
\gamma_{121LR_{13} + \gamma_{212LR_{23} = \gamma_{313} - \gamma_{323}},
\]

(12)

\[
\gamma_{131LR_{13} + \gamma_{232LR_{23} = \gamma_{313} + \gamma_{323}},
\]

(13)

\[
(\gamma_{131} - \gamma_{121})LR_{13} + \gamma_{232LR_{23} = \gamma_{323}},
\]

(14)

which are referred to as the “1 versus 2” line, the “1 versus 3” line, and the “2 versus 3” line, respectively (Edwards and Metz 2005a). The \( \gamma_{ij} \) terms represent the difference in utility between a correctly classified \( \pi_i \) event and a \( \pi_i \) event misclassified as \( \pi_j \). The \( \gamma_{ij} - \gamma_{jki} \) terms represent the difference in utilities between two misclassifications. In both cases the difference in utilities is scaled by the prevalence of the true class. The decision space is defined by the orthogonal likelihood ratio axes; with class \( \pi_1 \) more likely as \( LR_{13} \) increases, class \( \pi_2 \) more likely as \( LR_{23} \) increases, and class \( \pi_3 \) more likely as \( LR_{13} \) and \( LR_{23} \) decrease.

Rearranging (12)–(14) to solve for \( LR_{23} \) reveals that the slope and intercept of these decision lines are determined completely by the six \( \gamma_{ij} \) terms, that is, by the misclassification costs and the correct classification utilities, along with the class prevalence. Here is seen a major difference between the two- and \( n \)-class problems. For the standard two-class ROC, the decision threshold is independent of the misclassification costs or utilities, but for the multiclass scenario, the two are inseparable. Since the three lines intersect at a single point, there are in fact only five degrees of freedom. For example, if the slope and intercept of the 1-versus-2 and 1-versus-3 lines are fixed, then the slope of the 2-versus-3 line determines the location of that line’s intercept. Thus, we have the expected \( n^2 - n - 1 = 5 \) degrees of freedom within a six-dimensional ROC space.

An observer or classification system that partitions the decision space according to (12)–(14) is called the
ideal observer—the observer that maximizes the expected utility, or equivalently minimizes Bayesian risk (Edwards et al. 2004; He et al. 2006). In ROC terms, no observer, or decision strategy, can achieve a superior ROC curve to that achieved by the ideal observer; that is, for each POFD, the ideal observer will achieve a POD greater than or equal to any other observer. (Note that this is a stronger claim than simply that the ideal observer will achieve a higher AUC.) This property of the ideal observer has been extended to the multiclass problem (Edwards et al. 2004). Therefore, any classification system that does not partition the decision space according to (12)–(14) will have an inferior ROC surface. For example, it can be shown that the decision template used by Mossman (1999) and Dreiseitl et al. (2000) does not correspond to (12)–(14) and so cannot represent ideal observer performance (Edwards and Metz 2005b). This shortcoming is a direct result of the fact that the Mossman template is a sequence of binary decisions; that is, first decide whether an observation belongs to class $\pi_3$ or not; if not, then decide whether it belongs to class $\pi_2$; if not, then it belongs to class $\pi_1$ (Edwards and Metz 2005b). Each step involves a decision using only partial information; for example, the first decision is based solely on the probability that a given event belongs to class $\pi_3$, neglecting all information about classes $\pi_1$ and $\pi_2$. The ideal observer, on the other hand, makes optimal use of all available information. Incidentally, Mason (1982) also suggests treating the multiclass problem as a series of binary decisions.

The preceding discussion has established a framework for a suitable decision template and provides an example from the literature of a decision rule that does not fit this template. Still, given that the decision lines are a function of a user’s utilities, what shape should the decision template take when these utilities are unknown or when the classification system that is being evaluated will impact users with a variety of utilities? A full evaluation through the six-dimensional ROC space would require the slopes and intercepts to explore the full range of possible users. As established in section 2, this leads to a ROC surface that cannot be visualized and the volume under this surface is not a reliable performance measure unlike the area under the ROC curve. As a result, some constraints on the utilities are necessary.

If the misclassification errors are taken to have equal utilities (costs), then $\gamma_{ij} = \gamma_{ki}$ and so Eqs. (12)–(14) reduce to

$$\gamma_{121}LR_{13} + \gamma_{212}LR_{23} = 0,$$  \hspace{1cm} (15)

$$\gamma_{131}LR_{13} = \gamma_{313}, \text{ and}$$  \hspace{1cm} (16)

$$\gamma_{232}LR_{23} = \gamma_{323}.$$  \hspace{1cm} (17)

That is, the 1-versus-3 line is horizontal, the 2-versus-3 line is vertical, and the 1-versus-2 line passes through both the origin and the intersection of the other two lines (Scurfield 1998; He et al. 2006). [Both Scurfield (1998) and He et al. (2006) use the log of the likelihood ratio and not the form in (15)–(17). However, the ROC curve is invariant to monotonic transformations of the decision variable, so a monotonic transformation of an ideal observer decision variable is itself an ideal observer decision variable.] Under this constraint of equal utilities, the number of degrees of freedom is reduced from five to two, because now $\gamma_{ij} = \gamma_{ki}$, and the two-dimensional decision space problem becomes a straightforward extension of the one-dimensional decision space problem. Specifically, instead of moving the location of the pair of decision lines, as in the one-dimensional decision space, one varies the intercepts of the horizontal and vertical decision lines to produce a three-dimensional ROC surface, or the set of six three-dimensional ROC surfaces examined in section 2. While restricting the misclassification errors is unfortunate, it is necessary in order to visualize the problem with the three-dimensional surfaces. Furthermore, the suitability of this approach is supported by the fact that the observer that maximizes the ROC-123 surface is a special case of the general three-class ideal observer (Edwards and Metz 2006).

\subsection*{a. Example with artificial data}

Consider three bivariate normal distributions with means $(0, 0)$, $(1, 0)$, and $(0.5, \sqrt{3}/2)$ and unit variance and zero covariance; that is, they are each separated from the other two by one standard deviation. As indicated above, and paralleling the 1D-evidence case, sliding the two decision thresholds and plotting the relevant elements of the subsequent $3 \times 3$ contingency tables yields six ROC surfaces (Fig. 8), representing the six possibilities for classifying a set of observations, one from each class. The same permutation notation from section 2 will be used in this section.

Since each pair of density functions is separated by the same distance, the three ROC curves on the faces of the ROC-123 surface (Fig. 8) are identical and the overall volume is somewhat lower compared to the 1D Gaussians (Fig. 4). Notice as well that without the extra separation between two of the classes, the ROC surfaces containing two-category errors (ROC-231, ROC-312, and ROC-321) have larger volumes; that is, they are more likely to occur. Indeed, the notion of a two-category system does not even apply. Instead, the volumes break down into categories of all classes correctly classified (ROC-123), one class correctly classified (ROC-132, ROC-213, and ROC-321), and no classes correctly
classified (ROC-231 and ROC-312). It should also be noted that the connection between the volumes and areas for the 1D-evidence case expressed in (3) does not apply to the 2D-evidence case.

b. Example with National Weather Service forecasts of precipitation type

The National Weather Service issues winter weather forecasts of precipitation type: rain, freezing rain, ice pellets, or snow. Ice pellets compose a minute fraction of the total number of precipitation-type forecasts or observations and so are included with the snow forecasts herein, leaving three categories. Easy forecasts of rainfall, that is, events for which the surface temperature is greater than $5^\circ C$, are not considered; as a result, the dataset is dominated by snow events, $p(t = \pi_3) = 0.72$, with a substantial number of rain events, $p(t = \pi_R) = 0.25$, and relatively few freezing rain events, $p(t = \pi_Z) = 0.03$. Wandishin et al. (2005) evaluate probabilistic forecasts of precipitation type from a multiforecast model–multiprecipitation-type algorithm ensemble, given that some type of precipitation is both forecast and observed (see, especially, their Fig. 6). The ensemble produces very skillful forecasts, as measured by the Brier skill score (BSS; Wilks 1995), of rain (BSS = 0.64) and snow (BSS = 0.62) but the freezing rain forecasts (BSS = 0.20) possess only marginal skill. Despite possessing much less skill, the maximum relative expected value (Richardson 2000) of the freezing rain forecasts is quite high (0.66 compared to 0.77 for rain, and 0.78 for snow), highlighting the fact that skill measures only capture one aspect of forecast performance. [The ROC curves examined in Wandishin et al. (2005) are two-class approximations to the full multiclass problem, as described in section 4.] Furthermore, even the separate reliability diagrams do not capture completely the forecast performance from just the calibration–refinement perspective. For example, a reliability diagram gives no information about what does occur when the forecast event itself does not occur. A 20% forecast of rain might precede a rain event 30% of the time (an underforecast) but what event should a user expect after such a forecast? Obviously, this should...
depend on what forecast probabilities the other events received along with that 20% chance of rain, but the reliability diagram is not equipped to provide such information.

The ensemble is underdispersive, resulting in a large number of 100% and 0% forecasts. The transformation from a posteriori forecast probabilities to likelihood ratios is simple: one takes the ratio of the forecast probabilities and applies Bayes’s theorem. After cancellations, one is left with the likelihood ratio times the inverse of the class prevalence. However, ratios obviously are problematic with 0% forecast probabilities. Furthermore, the precipitation-type forecasts produce a significant number of ties. To handle these problems, a small random number $\mathcal{O}(10^{-5})$ is added to each forecast probability.

The ROC surfaces reflect the limited range of the forecast probabilities (Fig. 9). The ensemble possesses substantial ability to discriminate the different precipitation types. Recall that the no-skill volume is 0.167, and that a 1σ separation of the means leads to a volume of 0.46—a volume of 0.79 is substantially more impressive than an AUC of 0.79. However, while the volume is large, and thus the overall ability to correctly distinguish among the three events is high, there exists a large set of possible users for whom the precipitation-type forecasts are not nearly so useful. For example, consider a user whose cost sensitivity to freezing rain forecasts is such that he or she should take action at the threshold for which the POD for freezing rain ($P_{Z}$) exceeds 0.3. Because of the limited range of the forecasts, this user must instead act when the threshold is exceeded for which the $P_{Z}$ is 0.5 (the lowest nonzero POD achieved by the correct forecasts), thereby enduring many more false alarms. (The alternative is to never act, yielding a POD = 0, in which case the forecasts are of no value at all.) For snow events, the situation is even more extreme, as $P_{S}$ does not drop below 0.7.

Note that for the ROC surface representing cases in which just snow observations are correctly forecast (RSZ; middle left), the ROC curves on the left and right faces show considerable ability, but the low probability of mixing the rain and freezing rain forecasts keeps the total volume low (0.11), though it is still the second-most likely classification. Alternatively, the ROC surfaces show that

![Fig. 9. As in Fig. 4 but for NWS precipitation-type forecasts of freezing rain (Z), snow (S), and rain (R).](http://journals.ametsoc.org/waf/article-pdf/24/2/530/4645034/2008waf2222119_1.pdf)
the RSZ surface is the second-most common classification because of the skill of the forecasts in distinguishing snow events. Similarly, the strong ability to distinguish rain from other events makes the SZR classification (only the rain events correctly forecast; upper right) slightly more common (0.05) than the remaining possible classifications. Comparing the left and right faces of the RSZ ROC surface to those of the correctly classified ROC surface (ZSR; upper-left panel), it is apparent that misclassifying the rain and freezing rain events does not reduce greatly their separation from the snow events. In contrast, the ensemble is much better at distinguishing between rain and snow than between rain and freezing rain, and so the ROC curve on the bottom face of the top-right ROC surface (SZR; bottom face = rain versus snow events) is much better than the ROC curve on the left face (rain versus freezing rain events.)

**4. Two-class approximations to multiclass problems**

An alternative approach to high-dimensional ROC surfaces is to examine subsets of the total data in order to apply more directly the two-class methodology. There exist two forms of this alternative approach: “one versus all” and “one versus one”. In the one-versus-all approach (Provest and Domingos 2001; Fawcett 2003), each observation is classified as either belonging or not belonging to class $i$, $i = 1, 2, \ldots, n$ and a ROC curve is then produced for each of $n$ classes. This is also termed the class-reference formulation (Fawcett 2003). An AUC is calculated for each class and the final AUC is the weighted average of the three class-reference AUCs, where the weight is simply the prevalence of the reference class. Mathematically,

$$AUC_{CR} = \sum_{i=1}^{n} AUC(\pi_i)p(t = \pi_i),$$

where $AUC(\pi_i)$ is the area under the class-reference ROC curve for class $\pi$, and the subscript “CR” denotes “class reference.” Examples of the one-versus-all ROC curves are found in Mason and Graham (1999), Kharin and Zwiers (2003), and Wandishin et al. (2005); in none of these examples were the individual AUCs combined into a single forecast measure like $AUC_{CR}$.
The advantage of the class-reference approach is that the complexity of the multiclass problem increases linearly with $n$; that is, there is one ROC graph for each of the $n$ classes. Related to that, the results are easy to visualize. The disadvantages include the loss of information on misclassification errors. Although, the misclassification costs are not known precisely for many forecast problems, in many cases there may be some knowledge of the simple event–nonevent misclassification costs. For example, consider once again forecasts for severe thunderstorms, nonsevere thunderstorms, and no thunderstorms. Presumably, the public may be somewhat more alert to severe thunderstorms warnings, knowing that convective activity is expected as opposed to the no-thunder forecast scenario. This difference, however, may be insignificant compared to the public’s response to a severe weather forecast. Thus, the simpler severe–no-severe assessment may be suitable for forecast evaluation. One of the attractive features of the ROC curve and the AUC is that they are independent of the class prevalence and, thus, less sensitive to the properties of the particular data sample used in evaluating the classification system. Another disadvantage of the class-reference formulation is that AUC$_{CR}$ is sensitive to changes in class prevalence in two ways. First, from (18) it is evident that simply increasing the prevalence of a class that is well separated from the other classes will increase AUC$_{CR}$ without any real change in forecast ability. Less apparent is that since all classes $j$, $j \neq i$ are lumped together for the calculation of AUC($\pi_i$); this area, itself, is sensitive to changes in class prevalence. Consider the scenario in which class $k$ is easily distinguished from class $i$, $i \neq k$. Again, simply decreasing the prevalence of class $\pi_k$ will cause a corresponding decrease in AUC($\pi_i$).

For the one-versus-one approach (Hand and Till 2001), each class is compared pairwise with each other class. First, define $A(i|j)$ as the probability that a randomly drawn member of class $\pi_j$ will have a lower estimated probability of belonging to class $\pi_i$ than will a randomly drawn member of class $\pi_i$. Since there is no a priori reason to favor one class probability over the other, the separation between classes $\pi_i$ and $\pi_j$ is then measured by $A(i|j) = A(j|i) + A(i|i)/2$. For the two-class problem, $A(i|i) = A(j|i)$, but this typically does not hold for the multiclass problem. For example, consider two forecast triplets: $f_1 = (60\%, 30\%, 10\%)$ and $f_2 = (40\%, 10\%, 50\%)$. If $f_1$ and $f_2$ are classified based on the probability of belonging to class $\pi_1$, then $f_1$ will be assigned to class $\pi_1$ and $f_2$ assigned to class $\pi_2$. However, if they are classified based on the probability of belonging to $\pi_2$, then these classification will be reversed. The combined measure is a simple unweighted average over all pairs of classes:

FIG. 11. Pairwise ROC curves for bivariate normal distributions with means $\mu_1 = [0, 0]$, $\mu_2 = [1, 0]$, and $\mu_3 = [1.38, 1.45]$. See text for definitions of the indicated areas.
where the subscript \( P \) denotes “pairwise.” The pairwise approach maintains the two-class ROC property of being insensitive to class prevalence and misclassification costs and allows some examination of the misclassification errors. A disadvantage is the complexity of the pairwise comparisons: as with the Scurfield approach, \( n^2 - n \) graphs are required, although in this case they are the more familiar two-class ROC curves and not three-dimensional ROC surfaces.

### a. Example with artificial data

Consider isotropic bivariate normal distributions as in section 3a, except that the means of the class \( \pi_1 \) and class \( \pi_2 \) distributions are separated by 1 standard deviation, the means of the class \( \pi_1 \) and class \( \pi_3 \) distributions are separated by 2 standard deviations, and the means of the class \( \pi_2 \) and class \( \pi_3 \) distributions are separated by 1.5 standard deviations; that is, \( \mu_1^T = [0, 0] \), \( \mu_2^T = [1, 0] \), and \( \mu_3^T = [1.38, 1.45] \). For the ROC curves shown in Figs. 10 and 11, 2000 samples are taken from each distribution; the classes are equally prevalent. As expected, class \( \pi_2 \) observations are the most likely to be misclassified and class \( \pi_3 \) is the most easily distinguished (Fig. 10), since class \( \pi_3 \) has the greatest separation from the other two class (1.5 and 2 standard deviations) and class \( \pi_2 \) has the least separation (1 and 1.5 standard deviations). That class \( \pi_2 \) is the least well separated is also seen by the fact that using the probability that an observation belongs to class \( \pi_2 \) as the decision variable yields a smaller AUC than the probability of belonging to the other paired class; for example, compare \( A(2|1) \) with \( A(1|2) \) and \( A(2|3) \) with \( A(3|2) \) (Fig. 11). The arrangement of the class distributions is more easily determined from the pairwise-ROC curves: just as \( d(\mu_1, \mu_2) < d(\mu_2, \mu_3) < d(\mu_1, \mu_3) \)—where \( d(a, b) \) denotes the distance between \( a \) and \( b \)—so too \( A(1, 2) < A(2, 3) < A(1, 3) \), with those areas nearly equal to the standard two-class AUCs for means separated by 1, 1.5, and 2 standard deviations, respectively. Note also that the \( \text{AUC}_{CR} = \text{AUC}_P \).

Now consider the same class distributions but with 6000 samples taken from class \( \pi_3 \); that is, \( p(\pi_1) = p(\pi_2) = 0.2 \) and \( p(\pi_3) = 0.6 \). As described above, the pairwise ROC curves are unaffected by the change in class prevalence; however, the class-reference areas have increased.
from 0.82 to 0.86 for $\text{AUC}(\pi_1)$ and from 0.74 to 0.77 for $\text{AUC}(\pi_2)$, while $\text{AUC}(\pi_3)$ remains unchanged. Only the class-reference areas that contain the altered prevalence for a class grouped in the nonevent class will be affected. The combined area has increased from 0.81 to 0.85—without any fundamental change in discriminatory ability—and so is no longer equal $AUC_p$.

b. Example with precipitation-type forecast data

Recall that the ROC surfaces for the precipitation-type forecasts were somewhat difficult to interpret because the data points occupied only a small portion of the $P_i$ axes. The same is true for the ROC curves for the two two-class approximations (Figs. 12 and 13), but the curves are easier to visualize than the surfaces. Just as $V_{123}$ is quite large for these data, so, too, are the areas, $\text{AUC}_{CR} = 0.90$ and $\text{AUC}_p = 0.95$, areas that are roughly equivalent to two-class AUCs for distributions separated by 2–2.5 standard deviations. The higher class reference AUC suggests the value is inflated due to the very low prevalence of the more difficult to distinguish freezing rain observations. Observe that both $\text{AUC}(R)$ and $\text{AUC}(S)$ are nearly equal to $A(R,S)$. The infrequent freezing rain events have little influence on the class reference areas. Since freezing rain events often have a substantial societal impact (e.g., hazardous travel, power outages), the insensitivity of the class reference ROC curves to these events may be considered problematic.

5. Summary

We have presented extensions of the popular ROC analysis to the case of multiple event classes, for both a scalar decision variable and for multidimensional decision variables. A full extension of the two-class ROC curve is theoretically possible, but visualization is impossible as the ROC curve becomes an $(n^2 - n - 1)$-dimensional hypersurface within an $(n^2 - n)$-dimensional hypercube. Furthermore, the volume under this hypersurface, the multiclass equivalent to the AUC, appears to be unable to differentiate between nearly perfect and nearly guessing classification systems and, thus, is an unsuitable performance measure.

Scurfield (1996, 1998) tackled this problem by breaking the single high-dimensional surface into $(n^2 - n)$ $n$-dimensional surfaces by restricting the form of the decision rule. The surfaces represent the correct classification and all of the possible misclassifications, in a forced-choice scenario; the AUC is equivalent to the probability of success in a two-alternative forced-choice problem (2AFC). For example, for three classes (3AFC) it is possible to arrange a triplet of observations as 123, 132, 213, 231, 312, and 321. The volume under each of these
six surfaces is equal to the probability that the classification system will produce that arrangement. For example, $V_{123}$ equals the probability of a correct classification, while $V_{312}$ and $V_{231}$ equal probabilities of incorrectly classifying each observation in the triplet. Furthermore, it has been shown that a classification system that optimizes the ROC-123 surface in an ideal observer sense, that is, maximizes the expected utility, is also an ideal observer for the five-dimensional hypersurface, under certain constraints. These properties make the Scurfield approach very appealing. However, there are some shortcomings. The slope of a given segment of a two-class ROC curve is equal to the ratio of the misclassification costs association with that decision threshold, for a fixed event prevalence (Fawcett 2003). Thus, it is a simple task for a forecast user to determine which decision threshold provides maximum value. This relationship is lost when using the three-dimensional ROC surfaces. Also, the simplification required to reduce the degrees of freedom from five to two is that all misclassification costs are assumed to be equal. Scenarios for which this is not the case are easy to imagine: the precipitation-type forecasts examined herein, for example.

Since a full multiclass extension of ROC analysis is not possible given our current level of knowledge, it may be desirable to maintain the framework of the two-class problem. Two forms of this approach have been explored: one versus all and one versus one. In the one-versus-all, or class reference, approach, when evaluating the classification performance for class $a$, all of the remaining classes are lumped together into a single “not $a$” category. Therefore, details about specific misclassification errors are not available. Also, the class reference ROC curves, as they are called, are sensitive to changes in class prevalence. The insensitivity to changes in class prevalence is considered an attractive trait of two-class ROC analysis. The results are less dependent on the particular data sample from which they are attained. For the one-versus-one, or pairwise, approach, each class is compared with each of the remaining classes, in turn. Thus, estimates for all of the misclassification errors are obtained. In addition, the pairwise approach retains the insensitivity to changes in class prevalence. The $n^2 - n$ AUCs should be good approximations to the area under the curves formed by the intersection of the full-dimensional ROC hypersurface with the walls of the hypercube.

A final question might be: Which approach is preferable? Unfortunately, there does not appear to be a straightforward answer to that question. For example, if the full set of misclassification errors is not necessary, but only a “yes–no” assessment for each class, then the class reference approach may be suitable. If a greater knowledge of misclassification errors is required and the assumption of equal misclassification costs is thought to be appropriate, then the Scurfield approach would be desired. The pairwise approach offers the greatest freedom concerning misclassification errors, but, by definition, ignores how the trade-off between the detection rates of classes $a$ and $b$ changes as the detection rate of class $c$ varies. Thus, in evaluating a set of forecasts it is necessary to consider what is known about misclassification costs and what is needed by forecast users.

**Acknowledgments.** We thank Vit Drga for pointing us to several additional references and Harold Brooks for many helpful discussions. This research was supported in part by the National Science Foundation under Grant ATM-0432232.

**REFERENCES**


