

A utility-based performance metric for ROC analysis of N-class classification tasks

Darrin C. Edwards* and Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

We have shown previously that an obvious generalization of the area under an ROC curve (AUC) cannot serve as a useful performance metric in classification tasks with more than two classes. We define a new performance metric, grounded in the concept of expected utility familiar from ideal observer decision theory, but which should not suffer from the issues of dimensionality and degeneracy inherent in the hypervolume under the ROC hypersurface in tasks with more than two classes. In the present work, we compare this performance metric with the traditional AUC metric in a variety of two-class tasks. Our numerical studies suggest that the behavior of the proposed performance metric is consistent with that of the AUC performance metric in a wide range of two-class classification tasks, while analytical investigation of three-class “near-guessing” observers supports our claim that the proposed performance metric is well-defined and positive in the limit as the observer’s performance approaches that of the guessing observer.

Keywords: ROC methodology, expected utility, three-class classification

1. INTRODUCTION

We are attempting to extend the well-known observer performance evaluation methodology of receiver operating characteristic (ROC) analysis^{1,2} to classification tasks with three or more classes. This could conceivably be of benefit, for example, in a medical decision-making task in which a region of a patient image must be characterized as containing a malignant lesion, a benign lesion, or only normal tissue.³

Unfortunately, a fully general but tractable extension of ROC analysis to tasks with more than two classes has yet to be developed. It is known that the performance of an observer in a classification task with N classes ($N \geq 2$) can be completely described by a set of $N^2 - N$ conditional error probabilities,^{4,5} and that the performance of the ideal observer (that which minimizes Bayes risk⁴) is completely characterized by an ROC hypersurface in which these conditional error probabilities depend on a set of $N^2 - N - 1$ decision criteria.⁵ Although analytic expressions for the ideal observer’s conditional error probabilities given reasonable models for the underlying observational data have been worked out in the two-class case,⁶ this has not yet been accomplished in a fully general manner for tasks with three or more classes.

Furthermore, we have shown that an obvious generalization of the area under the ROC curve (AUC) does not in fact yield a useful performance metric in tasks with three or more classes.⁷ In the formulation we advocate, the set of $N^2 - N$ conditional error probabilities serve as the axes of the observer’s ROC space. This is equivalent to plotting a two-class observer’s false-negative fraction (FNF), rather than the more conventional true-positive fraction (TPF), as a function of false-positive fraction (FPF) to construct the observer’s ROC curve. Since $\text{FNF} = 1 - \text{TPF}$, this yields an ROC curve which is simply an “upside-down” version of the conventional curve, and the area under this ROC curve (which we will denote \tilde{A}) is just one minus the conventionally defined AUC. Clearly this area will vary from 0.5, for a “guessing” observer, to 0, for a “perfect” observer. In a task with more than two classes, however, we showed that although the “hypervolume under the ROC hypersurface” (HUH) is again 0 for a perfect observer, the HUH of a guessing observer is, counterintuitively, also 0.⁷ (Briefly, the number of degrees of freedom of the guessing observer’s ROC hypersurface is $N - 1$ rather than $N^2 - N - 1$, yielding a “degenerate” hypersurface with no hypervolume, much as in three dimensions the integral under a “surface” which is actually a curve — *e.g.*, $z = f(x, y)$ where $y = g(x)$ — will be zero.)

*Correspondence: E-mail: d-edwards@uchicago.edu; Telephone: 773 834 5094; Fax: 773 702 0371

What is needed is a performance metric that shares the useful properties of AUC, namely its intuitive direct relationship to the “difficulty” of the observer’s task (“near-guessing” observers have an \tilde{A} near 0.5, “near-perfect” observers have an \tilde{A} near 0), without suffering from this drawback of degeneracy. We have begun to investigate a performance metric that has its origins in the “expected utility” concept fundamental to ideal observer decision theory,⁴ and which we have reason to believe is both related to HUH and yet not plagued by the degeneracy issues of the HUH. In the next section, we attempt to motivate this performance metric, the “surface-averaged expected cost” (SAEC), and derive theoretical properties of this quantity. In Sec. 3, we outline the simulation studies we implemented in a number of simple two-class classification tasks; the results of those studies are presented in Sec. 4. The implications and limitations of the proposed metric are discussed in Sec. 5, and we summarize our conclusions in Sec. 6.

2. THEORY

In a two-class classification task, with the classes labeled “ π_+ ” (“positive”) and “ π_- ” (“negative”), the expected utility of an observer can be written as⁴

$$E\{\mathbf{U}\} \equiv (U_{TP}TPF + U_{FN}FNF)P(\pi_+) + (U_{FP}FPF + U_{TN}TNF)P(\pi_-), \quad (1)$$

where TPF is the probability of deciding an observation is positive, conditional on it actually being drawn from class π_+ , more explicitly denoted as $P(\mathbf{d} = \pi_+ | \mathbf{t} = \pi_+)$; FNF is $P(\mathbf{d} = \pi_- | \mathbf{t} = \pi_+)$; FPF is $P(\mathbf{d} = \pi_+ | \mathbf{t} = \pi_-)$; and TNF is the true-negative fraction, or $P(\mathbf{d} = \pi_- | \mathbf{t} = \pi_-)$. Each U represents the utility of a particular decision under a particular truth condition. (We use a bold typeface to denote statistically variable quantities, and here \mathbf{t} denotes the true class to which a randomly sampled observation belongs, while \mathbf{d} denotes the decision made for that observation.)

In a classification task with an arbitrary number of classes N , with labels running from π_1 to π_N , the above expression is readily generalized to obtain

$$E\{\mathbf{U}\} \equiv \sum_{j=1}^N \sum_{i=1}^N (U_{i|j}P_{ij})P(\pi_j), \quad (2)$$

where we have written the observer’s conditional classification rates $P(\mathbf{d} = \pi_i | \mathbf{t} = \pi_j)$ simply as P_{ij} . From the rules for conditional probability,⁸ $\sum_i P_{ij} = 1$, and so we can rewrite this expression to obtain

$$\begin{aligned} E\{\mathbf{U}\} &= \sum_{i=1}^N U_{i|i}P(\pi_i) \\ &\quad - \sum_{j=1}^N \sum_{\substack{i=1 \\ i \neq j}}^N (U_{j|j} - U_{i|j})P(\pi_j)P_{ij} \\ &= U_0 - \sum_{j=1}^N \sum_{\substack{i=1 \\ i \neq j}}^N \gamma_{jij}P_{ij}, \end{aligned} \quad (3)$$

where U_0 is just the expression $\sum_i U_{i|i}P(\pi_i)$ (independent of the conditional error rates P_{ij} which describe the observer’s performance), and $\gamma_{jij} \equiv (U_{j|j} - U_{i|j})P(\pi_j)$ gives, to within an arbitrary scale factor, the set of $N^2 - N - 1$ decision criteria used by the ideal observer to make decisions.^{5,9–11} Note that the γ_{jij} are strictly positive if we impose the reasonable assumption that an incorrect utility will always have a smaller utility than the corresponding correct decision. If we now define the “normalized” utility (more precisely, if we choose particular units in which to “measure” utility) as

$$\mathbf{u} \equiv \frac{\mathbf{U}}{(\sum_{j=1}^N \sum_{\substack{i=1 \\ i \neq j}}^N \gamma_{jij}^2)^{1/2}}, \quad (4)$$

and similarly define $\gamma_0 \equiv U_0/(\sum \gamma_{jij}^2)^{1/2}$, we can simplify the expression for expected utility further to obtain

$$E\{\mathbf{u}\} = \gamma_0 - \hat{\gamma} \cdot \vec{P}. \quad (5)$$

Here \vec{P} is an $(N^2 - N)$ -dimensional vector whose components are the conditional error rates P_{ij} (with a specified ordering, *e.g.*, $(P_{12}, P_{13}, \dots, P_{1N}, P_{21}, \dots, P_{N(N-2)}, P_{N(N-1)})$) — *i.e.*, the coordinates of ROC space; and $\hat{\gamma}$ is a unit vector of the same dimensionality as \vec{P} , whose components are the corresponding values of γ_{jij} after normalization.

It is important to keep in mind that although this normalized expected utility is optimized only by the ideal observer, it is well-defined for any observer at a particular operating point \vec{P} and choice of (normalized) utilities *via* $\hat{\gamma}$. Furthermore, assuming the values of the observational priors $P(\pi_i)$ to be fixed and the values of the utilities to be determined externally to the observer (*i.e.*, not modifiable by the observer within a given experiment or set of experiments), maximizing the normalized expected utility is clearly equivalent to minimizing $\hat{\gamma} \cdot \vec{P}$. We will refer to this latter quantity as the expected cost; note that although “cost” has a far more general definition in the literature (as do “utility,” “risk,” etc.), we will attempt to avoid confusion here by using the term only in this restricted sense.

Suppose we have measured the set of all possible values of $P_{N(N-1)}$ for a given observer as a function of the other $N^2 - N - 1$ conditional error probabilities. (For the ideal observer, this can be conceived of as measuring \vec{P} for every possible value of $\hat{\gamma}$; for a non-ideal observer, we assume that we can modify whatever set of $N^2 - N - 1$ decision criteria it is actually using, even if these are not usefully related to the utilities.) We write this as

$$\begin{aligned} P_{N(N-1)} &= R(P_{12}, P_{13}, \dots, P_{1N}, P_{21}, \dots, P_{N(N-3)}, P_{N(N-2)}) \\ &= R(\vec{P}^*), \end{aligned} \quad (6)$$

where \vec{P}^* denotes the “reduced” vector, of dimensionality $N^2 - N - 1$, obtained by deleting the $(N^2 - N)$ th component of \vec{P} . The HUH can be defined⁷ as

$$\text{HUH} \equiv \int_{\Omega_R} R(\vec{P}^*) d^{N^2 - N - 1} \vec{P}^*, \quad (7)$$

or equivalently,

$$\text{HUH} \equiv \int_{V_R} d^{N^2 - N} \vec{P}, \quad (8)$$

where Ω_R denotes the set of \vec{P}^* for which $R(\vec{P}^*)$ is defined (the domain of the function defining the ROC hypersurface), and V_R denotes the set of all \vec{P} enclosed by that hypersurface and by the boundaries of the ROC space (given that $0 \leq P_{ij} \leq 1$). Note that in a two-class task, with the ROC curve given by $\text{FNF} = R(\text{FPF})$, this reduces to

$$\begin{aligned} \text{HUH} &= \int_{V_R} d^{N^2 - N} \vec{P} \\ &= \int_0^1 \int_0^{R(\text{FPF})} d\text{FNF} d\text{FPF} \\ &= \int_0^1 R(\text{FPF}) d\text{FPF} \\ &= \tilde{A}, \end{aligned} \quad (9)$$

as expected. (Note that, as stated in Sec. 1, this is one minus the conventional AUC that would be obtained by integrating TPF as a function of FPF.)

Despite the long-standing success of AUC as a summary performance metric for ROC analysis, we have shown the HUH not to be useful for this purpose in a classification task with three or more classes.⁷ Briefly, a “perfect” observer can achieve values of, say, $P_{N(N-1)} = 0$ for any achievable set of \vec{P}^* ; by Eq. 7, the HUH for such an observer will thus be zero (and will approach zero for a “near-perfect” observer). A “guessing” observer will assign observations to the N classes randomly, independent of the actual truth states of those observations; since the total probability of making a decision will be one, this leaves a set of only $N - 1$ degrees of freedom (each of the probabilities of assigning an observation to a given class). But it can be shown that in such a situation, the resulting domain of integration Ω_R is “degenerate,” and the integral in Eq. 7 is zero regardless of the value of the integrand (and will approach zero for a “near-guessing” observer). Thus, opposite extremes of performance result in similar or identical values of HUH, making this quantity useless even as a summary performance metric in classification tasks with more than two classes. In the two-class case, $N^2 - N - 1 = N - 1 = 1$, of course, and (amusingly or providentially, depending perhaps on one’s worldview) no such degeneracy is encountered.

Discouraging though this result may be, it immediately brings to the forefront the question of what motivated the choice of AUC as a summary performance metric to begin with. In the present context, it can be said that AUC averages directly over “performance description variables” (such as FNF) without regard to utility (or, equivalently, cost). For an experiment involving a human observer (the internals of whose decision-making process may be unavailable to experimenter control) or an algorithmic observer (trained on a finite sample of observational data), the actual “costs” may be unknown to the experimenter, or may not be available for modification in any practical sense. On the other hand, ideal observer decision theory demonstrates the tremendous theoretical and practical importance of Eq. 5, and it is natural to ask whether consideration of the expected cost, $\hat{\gamma} \cdot \vec{P}$, might not be worthwhile, given the difficulty in generalizing AUC just described.

For the ideal observer itself, this line of inquiry seems quite promising indeed. For each possible value of $\hat{\gamma}$, the ideal observer will choose an operating point \vec{P} that minimizes the expected cost. (It is possible, given particular forms of the observational data probability density functions (PDFs), that multiple operating points \vec{P} will be associated with a given $\hat{\gamma}$; it can be shown, however, that such points will always lie in a simply connected region, analogous to a straight line along a two-class ROC surface. We will not consider such special cases here.) By taking the ideal observer’s ROC hypersurface as given, one can proceed in the opposite direction: at any given point on the ideal observer’s ROC surface, the appropriate $\hat{\gamma}$ for that point is that which minimizes the expected cost. This, in turn, can be shown to imply that the appropriate $\hat{\gamma}$ is normal to the ideal observer’s ROC hypersurface at each point \vec{P} .

For non-ideal observers, the situation is much more confusing. Given that such an observer might not be basing its decisions on the utilities (available to the ideal observer) at all, it is unclear what value of $\hat{\gamma}$ to assign to a given \vec{P} on such an observer’s ROC surface. Arbitrarily, we choose to make the same assignment made by the ideal observer: at each point on the observer’s ROC hypersurface, we choose that value of $\hat{\gamma}$ that is normal to the ROC hypersurface at that point. Intuitively, this can be taken to be giving the non-ideal observer the “benefit of the doubt”: in determining a total expected cost for the observer, we will at each point take the contribution to that cost to be the “minimum” possible. Alternatively, we can say that the observer under this model is at least behaving “locally” optimally.

Thus, for the ROC hypersurface given in Eq. 6, we define the “local” utility vector to be

$$\hat{\gamma}_R \equiv \frac{(-\nabla R, 1)}{\sqrt{|\nabla R|^2 + 1}}, \quad (10)$$

where the expression in parentheses denotes a vector of dimension $N^2 - N$ whose first $N^2 - N - 1$ components are the negatives of the components of ∇R ; the sign is chosen because the components of $\hat{\gamma}_R$ must be positive, ruling out the possibility $(\nabla R, -1)$. We use this definition to construct the surface integral

$$\int_{\sigma_R} \hat{\gamma}_R \cdot \vec{P} \, d^{N^2-N-1} \sigma. \quad (11)$$

The integral is over the ROC hypersurface σ_R , that is, the set of points \vec{P} such that $P_{N(N-1)} = R(\vec{P}^*)$. The differential element on this hypersurface is denoted by $d^{N^2-N-1}\sigma$, where the superscript reminds us of the dimensionality “within” that surface.

In the two class case, the differential element reduces to the differential arc length, which we can define as

$$ds \equiv \sqrt{1 + \left(\frac{d\text{FNF}}{d\text{FPF}}\right)^2} d\text{FPF}. \quad (12)$$

The integral in Eq. 11 can then be written as

$$\begin{aligned} \int_0^1 \frac{\left(\frac{-d\text{FNF}}{d\text{FPF}}, 1\right)}{\sqrt{\left(\frac{d\text{FNF}}{d\text{FPF}}\right)^2 + 1}} \cdot (\text{FPF}, \text{FNF}) \sqrt{1 + \left(\frac{d\text{FNF}}{d\text{FPF}}\right)^2} d\text{FPF} &= \int_0^1 \left(-\text{FPF} \frac{d\text{FNF}}{d\text{FPF}} + \text{FNF}\right) d\text{FPF} \\ &= \int_0^1 -\text{FPF} \frac{d\text{FNF}}{d\text{FPF}} d\text{FPF} + \int_0^1 \text{FNF} d\text{FPF} \\ &= \int_0^1 \text{FPF} d\text{FNF} + \int_0^1 \text{FNF} d\text{FPF} \\ &= 2\tilde{A}. \end{aligned} \quad (13)$$

Note that in the next to last step, the negative sign has disappeared because $\text{FNF} = 0$ when $\text{FPF} = 1$ and vice versa, so that the order of the limits of integration will be reversed. It is also vital to remember that \tilde{A} here denotes the area under the “upside-down” ROC curve (FNF plotted against FPF), and is thus one minus the conventional AUC.

Clearly the quantity we have defined is directly related to performance — in fact, far more closely than we had reason to hope: despite our *ad hoc* choice of $\hat{\gamma}_R$, the relation in Eq. 13 holds for arbitrary observers, and not just ideal observers. Even more surprisingly, the generalization of this relationship can be shown to hold for observers in tasks with arbitrary numbers of classes. Returning to Eq. 11, we rearrange terms to obtain

$$\begin{aligned} \int_{\sigma_R} \hat{\gamma}_R \cdot \vec{P} d^{N^2-N-1}\sigma &= \int_{\partial V_R} \hat{\gamma}_R \cdot \vec{P} d^{N^2-N-1}\sigma \\ &= \int_{\partial V_R} \vec{P} \cdot \hat{\gamma}_R d^{N^2-N-1}\sigma \\ &= \int_{\partial V_R} \vec{P} \cdot \left[\frac{(-\nabla R, 1)}{\sqrt{|\nabla R|^2 + 1}} d^{N^2-N-1}\sigma \right] \\ &= \int_{V_R} \text{div} \vec{P} d^{N^2-N} \vec{P} \\ &= (N^2 - N)\text{HUH}. \end{aligned} \quad (14)$$

Here we have used the n -dimensional extension of the divergence theorem (known in three dimensions as Gauss’s theorem);¹² div is the operator $\sum_i (\partial/\partial P_i)$, which when applied to the vector \vec{P} will simply yield the dimensionality $N^2 - N$ of \vec{P} . Note also that in the first step, we have “closed” the ROC hypersurface with the boundary ∂V_R of the ROC hypervolume; this can be done for the given integrand, because the “bottom” surface $P_{N(N-1)} = 0$ will contribute nothing to the surface integral.

Unfortunately, we are now back where we started: since it is equal (to within a proportionality constant) to the HUH, the surface integral defined above will have exactly the same drawbacks as that quantity. However, writing

the performance metric in this form — as an integral of the scalar quantity $\hat{\gamma}_R \cdot \vec{P}$ over the ROC hypersurface — suggests a different approach, namely, considering an “average” of this quantity over the hypersurface:

$$\overline{C}_\sigma \equiv \frac{\int_{\sigma_R} \hat{\gamma}_R \cdot \vec{P} d^{N^2-N-1}\sigma}{\int_{\sigma_R} d^{N^2-N-1}\sigma} \quad (15)$$

where we have divided the previous quantity by the “surface area” of the ROC hypersurface. The quantity \overline{C}_σ is the SAEC referred to in Sec. 1; the overline reminds us that it is an expectation value, and the subscript σ reminds us that it is averaged over a surface (the ROC hypersurface). This can be considered analogous to the concept from univariable calculus of the “average” of a function over an interval:

$$f_{\text{avg}} \equiv \frac{1}{b-a} \int_a^b f(x) dx. \quad (16)$$

In particular, it should be immediately clear that \overline{C}_σ is bounded by the maximum and minimum values of $\hat{\gamma}_R \cdot \vec{P}$, and that if $\hat{\gamma}_R \cdot \vec{P}$ were constant over a given ROC hypersurface, then \overline{C}_σ would be equal to this constant value.

Further analysis will need to be performed to confirm that this quantity remains well-defined for guessing or even “near-guessing” observers. We have reason to believe that an extension of L’Hôpital’s rule should be applicable in this case; *i.e.*, although the numerator and denominator will both converge to zero in the limit of approach to a guessing observer, the limit of \overline{C}_σ itself should still be a non-zero quantity. Our results in this regard, however, are still very preliminary. For the present work, we will consider only properties of this quantity in the two-class case, where the degeneracy issues involving HUH do not arise. In the two-class case, of course, we can use Eq. 13 to write

$$\overline{C}_\sigma \equiv \frac{2\tilde{A}}{S} \quad (17)$$

where S is the arc-length along the ROC curve.

3. MATERIALS AND METHOD

We numerically investigated the behavior of \overline{C}_σ compared with the conventional AUC under two models for the distributions of the observer’s latent decision variable data: the “conventional” binormal model,¹³ and the ideal-observer-related “proper” binormal model.⁶ Under the conventional model, the observer’s decision variables are assumed to be drawn from a pair of distributions which are an (unspecified) monotonic transformation of two normal distributions:

$$\mathbf{x}_+ \sim N(x; \mu_+ = a/b, \sigma_+ = 1/b) \quad (18)$$

and

$$\mathbf{x}_- \sim N(x; \mu_- = 0, \sigma_- = 1), \quad (19)$$

where $N(x; \mu, \sigma)$ is a normal density function with mean μ and standard deviation σ . The observer makes decisions by comparing an observation of unknown class \mathbf{x} with a threshold x_0 ; varying this threshold from $-\infty$ to ∞ will sweep out the observer’s ROC curve. This curve is completely specified by the two parameters a and b , and analytic forms exist for both individual operating points (FPF, TPF) and the conventional AUC (denoted A_z under this model) as functions of a and b .¹³

Under the “proper” binormal model, the observer is again assumed to make decisions using underlying data monotonically related to the pair of distributions given in Eqs. 18 and 19. However, the actual decisions are made by comparing the likelihood ratio of \mathbf{x} , rather than \mathbf{x} itself, with a threshold. The likelihood ratio is given by

$$\mathbf{y} \equiv \frac{N(\mathbf{x}; a/b, 1/b)}{N(\mathbf{x}; 0, 1)}. \quad (20)$$

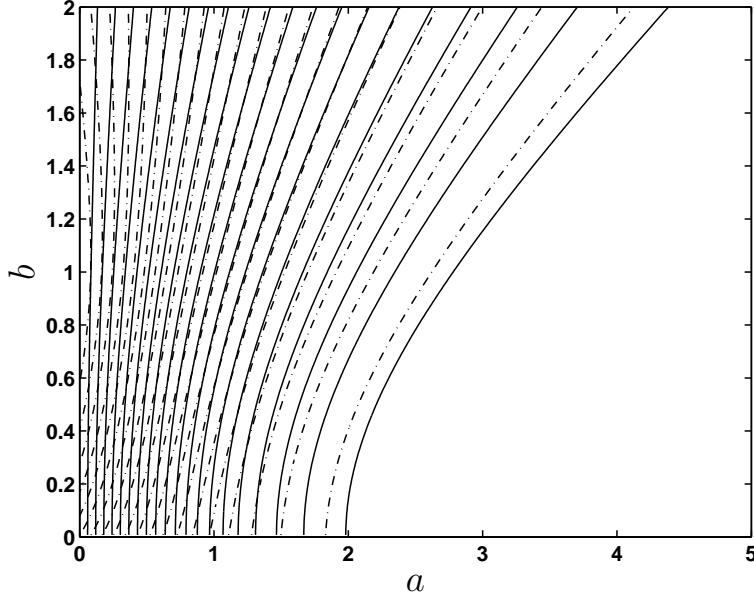


Figure 1. Isopleths of the A_z performance metric (solid lines) and of the proposed \overline{C}_σ metric (dash-dotted lines), for various values of the a and b parameters of the conventional binormal model.

Varying the threshold y_0 throughout its range will sweep out the observer’s ROC curve. For numerical purposes, it has been found convenient to parametrize this curve using the parameters

$$c \equiv \frac{b-1}{b+1} \tag{21}$$

and

$$d_a \equiv \frac{\sqrt{2}a}{\sqrt{1+b^2}} \tag{22}$$

rather than a and b directly. The observer’s ROC curve is completely specified by c and d_a , and analytic forms have been determined for both individual operating points (FPF, TPF) and the conventional AUC under this model as functions of those two parameters.⁶

We calculated the A_z of an observer assumed to operate under the conventional binormal model for 250 values of a distributed uniformly between 0 and 5, and (at each such value of a) for 250 values of b distributed uniformly between 0 and 2. For each of these 62,500 pairs of parameter values, we also calculated the corresponding value of \overline{C}_σ using the relation in Eq. 17. (The arc length S was calculated by generating a large number of operating points along the curve, and adding together the line segment lengths $\sqrt{(FPF_i - FPF_{i-1})^2 + (TPF_i - TPF_{i-1})^2}$.)

A similar procedure was performed for the proper binormal model. We calculated the conventional AUC for each of 250 values of c distributed uniformly between -1 and 1 , and (at each such value of c) for 250 values of d_a uniformly distributed between 0 and 4. For each of these 62,500 pairs of parameter values, we also calculated the corresponding value of \overline{C}_σ (again using the approximation for arc length described for the conventional model).

4. RESULTS

The calculated values of A_z and of \overline{C}_σ for the conventional binormal model are shown in isopleth (“contour”) plots in Fig. 1. Similarly, the calculated values of the conventional AUC and \overline{C}_σ for the proper binormal model are shown in isopleth plots in Fig. 2.

Although difficult to discern from the plot, the isopleths in Fig. 1 do in fact cross, particularly in the lower left region. For example, the parameter pair ($a = 0.4819, b = 0.5060$) corresponds to an A_z value of 0.6663 and

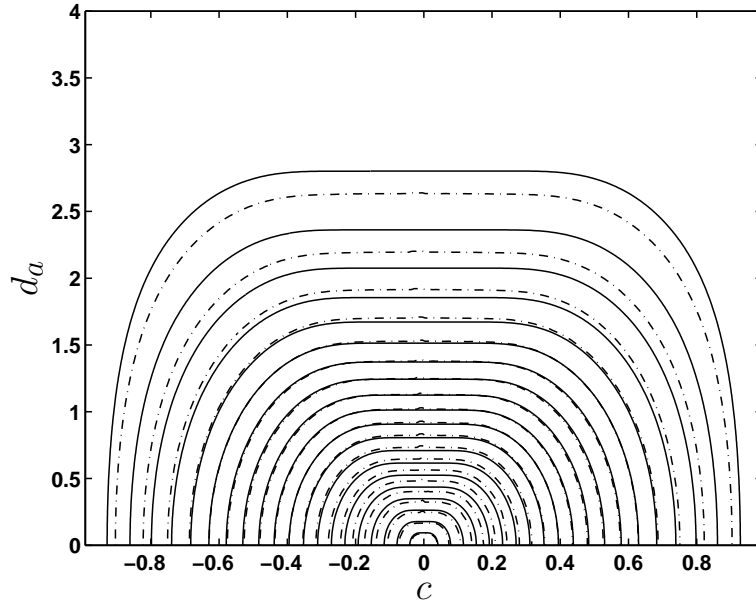


Figure 2. Isopleths of the conventional AUC performance metric (solid lines) and of the proposed \overline{C}_σ metric (dash-dotted lines), for various values of the c and d_a parameters of the proper binormal model.

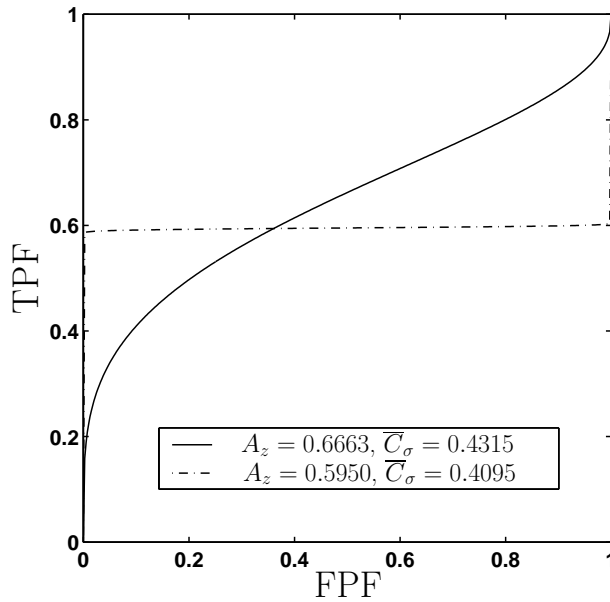


Figure 3. ROC curves generated under the conventional binormal model with parameter values of $(a = 0.4189, b = 0.5060)$ (solid curve), and $(a = 0.2410, b = 0.0080)$ (dash-dotted curve).

a \overline{C}_σ of 0.4315, while the parameter pair $(a = 0.2410, b = 0.0080)$ corresponds to an ROC curve which has both a lower A_z of 0.5950 and a lower \overline{C}_σ of 0.4095. (Recall that, as its name implies, the SAEC \overline{C}_σ is a “cost”, and thus lower values are intended to be “preferable,” in contrast to A_z and the conventional AUC.) These two ROC curves are plotted (conventionally, using TPF as the ordinate) in Fig. 3.

5. DISCUSSION

It is evident from Fig. 1 that the proposed performance metric \overline{C}_σ does not perform identically to the conventional AUC in all situations (*i.e.*, for arbitrary decision rules). This is illustrated in more detail in Fig. 3; if the two curves represented observers (radiologists or imaging systems, for example) which one wished to rank in order of performance, then the two performance metrics would disagree as to which were actually preferable. This is understandable given the shapes of the curves; the system with slightly lower A_z is so severely “hooked” that its arc length will be very close to two, driving down the “cost” \overline{C}_σ to a greater extent than the loss in conventional AUC.

It should be recalled, however, that in practical situations in which such a severe “hook” is seen in the ROC curve, the observational data themselves do not usually support such a fitting of the curve.⁶ Even aside from such data sampling and curve-fitting issues, comparing two systems when at least one of them has an ROC curve with such a large “hook” is often problematic (compare the well-known situation when two systems have very similar AUCs, but “cross,” making the decision of which system to prefer dependent on the region of ROC space in which one chooses to operate). In short, the fact that \overline{C}_σ does not agree exactly with a performance metric such as A_z , itself known to be imperfect, is not necessarily a fatal flaw.

The results presented in Fig. 2 are far more surprising. There appear to be no visible “crossings” of the isopleths for any choices of parameters c and d_a . Although this result still needs to be confirmed analytically, it would if found true imply that \overline{C}_σ and the conventional AUC under the proper binormal model are equivalent performance metrics. Whether this equivalence could be extended to arbitrary ideal observer models (*i.e.*, those for arbitrary PDFs rather than the binormal model) would also be an important area for further investigation.

The extensibility of the proposed performance metric to tasks with more than two classes is quite plausible, but much remains to be done here as well. Preliminary work in this direction suggests that it may be possible to apply an extension of L’Hôpital’s rule to the integrals in Eq. 15 in the situation where they approach 0 due to dimensionality considerations. However, the resulting limit appears to depend strongly on the underlying data PDFs (a counterintuitive result given the behavior of two-class near-guessing observers, whose ROC curves all approach the diagonal line regardless of the data PDFs). More careful work will be necessary to validate or refute these claims.

Related to the issue of dimensionality just mentioned is the situation of the “discrete” observer, *i.e.*, an observer which operates only at discrete operating points in ROC space (this applies to the two-class observer as well as those with more classes). We have so far been unable to usefully generalize the definition of $\hat{\gamma}_R$ and thus Eq. 15 to this situation, even in the two-class case. It remains to be seen whether this last issue is an important one or not.

6. CONCLUSIONS

We have proposed a novel ROC performance metric, the SAEC. Although grounded in the same theoretical framework as the expected utility of the ideal observer, its practical realization involves readily comprehensible quantities — the AUC and the arc length along the ROC curve in a two-class task, and the surface-averaged integral of a well-defined scalar in a task with more than two classes.

Although the properties of this performance metric have yet to be thoroughly investigated, preliminary results are quite encouraging. We have high hopes that this performance metric will allow comparison of observers in classification tasks of varying complexity, without suffering from the drawbacks that other performance metrics, such as the HUH, have been shown to possess.

ACKNOWLEDGMENTS

This work was supported by grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D. C. Edwards, principal investigator). Charles E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

REFERENCES

1. J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
2. C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine* **VIII**(4), pp. 283–298, 1978.
3. D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.* **31**, pp. 81–90, 2004.
4. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
5. D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N -class classification," *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.
6. C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, pp. 1–33, 1999.
7. D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N -class classification tasks," *IEEE Trans. Med. Imag.* **24**, pp. 293–299, 2005.
8. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., New York, 1991.
9. D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.* **50**, pp. 478–487, 2006.
10. D. C. Edwards and C. E. Metz, "Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities," in Proc. SPIE Vol. 6146 *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Miguel P. Eckstein, eds., pp. 61460A1–61460A7, (SPIE, Bellingham, WA), 2006.
11. D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.* , 2006. (submitted).
12. S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations: Second Edition*, Harcourt Brace Jovanovich, San Diego, CA, 1986.
13. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statist. Med.* **17**, pp. 1033–1053, 1998.