

Optimality of a utility-based performance metric for ROC analysis

Darrin C. Edwards* and Charles E. Metz

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

ABSTRACT

We previously introduced a utility-based ROC performance metric, the “surface-averaged expected cost” (SAEC), to address difficulties which arise in generalizing the well-known area under the ROC curve (AUC) to classification tasks with more than two classes. In a two-class classification task, the SAEC can be shown explicitly to be twice the area above the conventional ROC curve ($1 - \text{AUC}$) divided by the arclength along the ROC curve. In the present work, we show that for a variety of two-class tasks under the binormal model, the SAEC obtained for the proper decision variable (the likelihood ratio of the latent decision variable) is less than that obtained for the conventional decision variable (*i.e.*, using the latent decision variable directly). We also justify this result using a readily derived property of the arclength along the ROC curve under a given data model. Numerical studies as well as theoretical analysis suggest that the behavior of the SAEC is consistent with that of the AUC performance metric, in the sense that the optimal value of this quantity is achieved by the ideal observer.

Keywords: ROC methodology, expected utility, ideal observer estimation

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

More recently, we proposed⁸ a novel performance metric called the “surface-averaged expected cost” (SAEC), defined as

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

where \vec{P} is a vector of the observer’s conditional error probabilities, any one of which, considered as a function of the others, can be used to define the ROC hypersurface; $\hat{\gamma}_R$ is a unit vector normal to the ROC hypersurface; σ_R denotes the boundary of the region enclosed by the ROC hypersurface and the boundaries of the ROC space; and $d^{N^2-N-1}\sigma$ is a differential element of surface area on σ_R . This expression, although complicated in appearance, is motivated by straightforward considerations of the expected utility (or, equivalently, the “costs”) of decisions made by an observer; furthermore, because its form is that of an average of a quantity over a region, it is hoped the SAEC will prove more useful in classification tasks with more than two classes than the HUH itself. (We note that although the definition of $\hat{\gamma}_R$ is, for the ideal observer, equivalent to a normalized version of the decision utilities and class priors used by the ideal observer at that operating point,⁸ the definition in terms of the ROC surface itself is well-defined for observers which do not make use of an ideal, or even cost-based, decision strategy. To put it another way, although the form of the SAEC is, as just stated, *motivated* by considerations of observer expected utility, its calculation is not dependent on a particular observer’s choice of utility or cost structure, even assuming the observer makes decisions based on such a structure; nor is it dependent on the class priors, *i.e.*, the prevalences.)

We also showed previously that, in a two-class classification task, the SAEC reduces to $2\tilde{A}/S$, where S is the arclength along the ROC curve.⁸ Empirically, this quantity was found to have behavior consistent with the AUC: under an ideal observer decision model, with observational data drawn from a binormal distribution, the AUC and SAEC were found to be monotonically related; while under a conventional binormal model, discrepancies between SAEC and A_z were found to arise in situations where A_z itself does not provide an unambiguous measure of observer performance (*e.g.*, where curves cross, or are strongly “hooked”⁶).

The previous work described above was limited in that, for each simulation study, a single decision strategy was considered (either the “proper” binormal model⁶ or the “conventional” binormal model⁹). In the work described here, we wish to compare the behavior of the SAEC across decision strategies. That is, for a given data model (a particular choice of a pair of normal distributions for the observational data), we compare the value of SAEC obtained under the “proper” binormal observer model (which estimates the behavior of the ideal observer) with the value of SAEC obtained under the “conventional” binormal model. It is well known that the ideal observer obtains a larger value for the AUC than any observer using a different decision strategy.⁴ By the same token, if it can be shown that the ideal observer obtains a lower value of SAEC than that obtained by any observer using a different decision strategy, this should serve to support the claim that SAEC could prove useful as a performance metric (*e.g.*, for ranking observers by performance).

2. MATERIALS AND METHOD

We numerically investigated the behavior of \overline{C}_σ under both the conventional and proper binormal models. 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:

$$\mathbf{x}_+ \sim N(x; \mu_+ = a/b, \sigma_+ = 1/b) \quad (2)$$

and

$$\mathbf{x}_- \sim N(x; \mu_- = 0, \sigma_- = 1), \quad (3)$$

where $N(x; \mu, \sigma)$ is a normal probability 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 .⁹

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. 2 and 3. 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

$$\mathbf{y} \equiv \frac{N(\mathbf{x}; a/b, 1/b)}{N(\mathbf{x}; 0, 1)}, \quad (4)$$

where $N(x; \mu, \sigma)$ is a normal probability density function with mean μ and standard deviation σ . 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

$$c \equiv \frac{b-1}{b+1} \quad (5)$$

and

$$d_a \equiv \frac{\sqrt{2}a}{\sqrt{1+b^2}} \quad (6)$$

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

Using the relation $\bar{C}_\sigma = 2\tilde{A}/S$ given in Sec. 1, we calculated $\bar{C}_\sigma^{\text{conv}}$ for an observer assumed to operate under the conventional binormal model for each of 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 0.008 and 2. (The arc length S was calculated by generating a large number of operating points along the curve, and adding together the line segment lengths $\sqrt{(FPF_i - FPF_{i-1})^2 + (TPF_i - TPF_{i-1})^2}$.)

For each of these 62,500 pairs of parameter values, we obtained the corresponding values of the proper binormal parameters c and d_a using the relations in Eqs. 5 and 6. For each such resulting parameter pair, we then calculated $\bar{C}_\sigma^{\text{prop}}$ for an observer assumed to operate under the proper binormal model, again using the approximation for arc length described for the conventional model.

Finally, for each pair of parameter values a and b , we looked for “perverse” situations in which $\bar{C}_\sigma^{\text{prop}}$ for the proper binormal model observer was larger than the value of $\bar{C}_\sigma^{\text{conv}}$ for the corresponding conventional binormal model observer. Since \bar{C}_σ is defined as a cost, smaller values are “preferable” to larger ones (as is the case for \tilde{A} , one minus the conventional AUC). That is, we looked for situations in which the SAEC index ranked the ideal observer as inferior to another decision strategy.

3. RESULTS

Figure 1 shows $\Delta_C \equiv \bar{C}_\sigma^{\text{prop}} - \bar{C}_\sigma^{\text{conv}}$ as a function of the two binormal model parameters a and b for those values of $\Delta_C > 0$. That is, negative values of Δ_C , corresponding to the expected situation in which $\bar{C}_\sigma^{\text{prop}}$ is less than $\bar{C}_\sigma^{\text{conv}}$, are set to zero in the plot so that the (relatively smaller) positive values are more readily discerned.

Because Fig. 1 is potentially misleading, in that a given difference may be of greater or lesser importance depending on the actual magnitude of the values of \bar{C}_σ in question, we plot in Fig. 2 the natural logarithm of the ratio $\bar{C}_\sigma^{\text{prop}}/\bar{C}_\sigma^{\text{conv}}$. This can equivalently be thought of as the difference $\ln \bar{C}_\sigma^{\text{prop}} - \ln \bar{C}_\sigma^{\text{conv}}$, and we therefore denote this quantity $\Delta_{\ln C}$. As in Fig. 1, negative values of $\Delta_{\ln C}$, corresponding to the expected situation in which $\bar{C}_\sigma^{\text{prop}}$ is less than $\bar{C}_\sigma^{\text{conv}}$, are set to zero in the plot so that the (relatively smaller) positive values are more readily discerned.

4. DISCUSSION

In the vast majority of situations for which Δ_C was found greater than zero, this difference was itself quite small (less than or equal to 0.001). As is evident in Fig. 1, these situations generally correspond to values of the a parameter greater than three. From Fig. 2, the majority of these situations are seen not to have a large value of the ratio $\bar{C}_\sigma^{\text{prop}}/\bar{C}_\sigma^{\text{conv}}$. That is, the difference $\bar{C}_\sigma^{\text{prop}} - \bar{C}_\sigma^{\text{conv}}$ was actually rather small compared to the

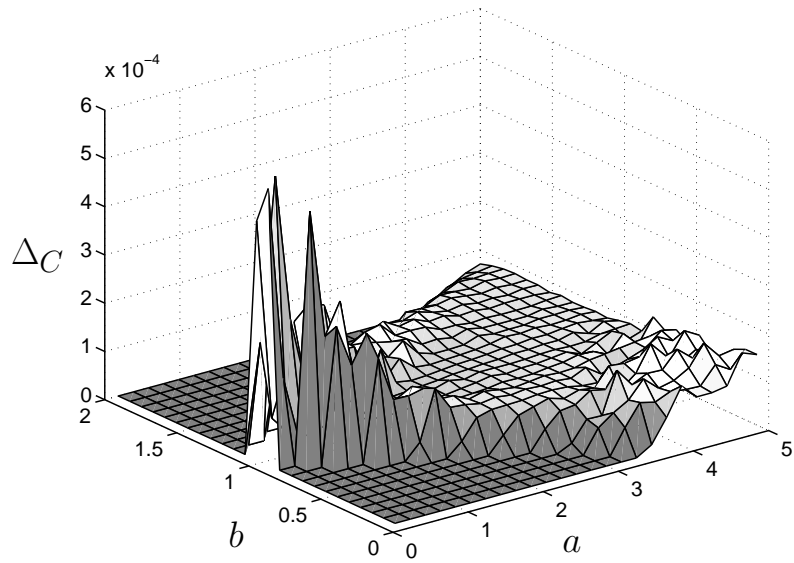


Figure 1. The difference Δ_C in SAEC values for the proper and conventional binormal models as a function of the binormal model parameters a and b , with negative values of Δ_C set to zero to allow for a reasonable dynamic range in the graph.

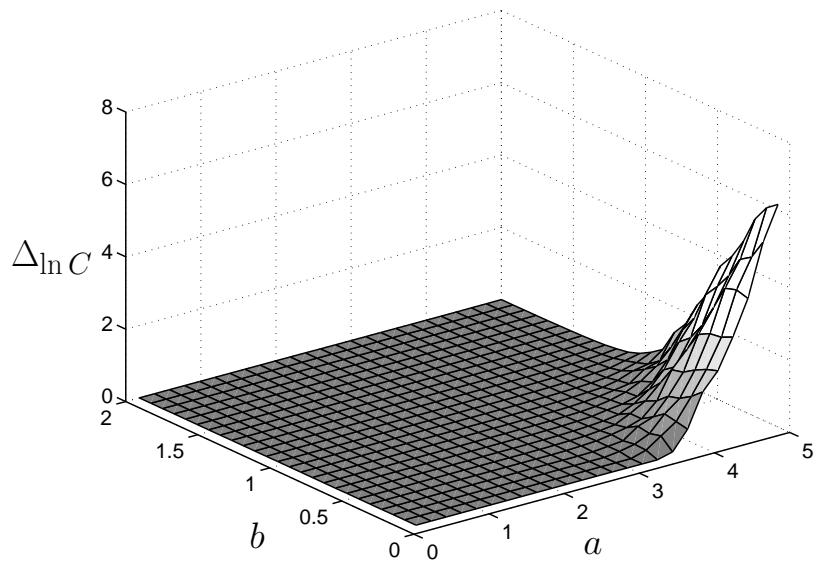


Figure 2. The difference $\Delta_{\ln C}$ in natural logarithms of SAEC values for the proper and conventional binormal models as a function of the binormal model parameters a and b , with negative values of $\Delta_{\ln C}$ set to zero to allow for a reasonable dynamic range in the graph.

magnitude of either $\overline{C}_\sigma^{\text{prop}}$ or $\overline{C}_\sigma^{\text{conv}}$. We claim that this measured discrepancy in $\overline{C}_\sigma^{\text{prop}}$ and $\overline{C}_\sigma^{\text{conv}}$ (in particular, $\overline{C}_\sigma^{\text{prop}}$ being greater than $\overline{C}_\sigma^{\text{conv}}$, contrary to theoretical expectations) is attributable, in one form or other, to rounding error.

The exception to this trend, *i.e.*, large values of $\Delta_{\ln C}$ visible in Fig. 2, are clearly grouped in the range of the graph where $b \leq 1$ and a is particularly large. These are situations in which the AUC is generally close to 1 for both the conventional and proper binormal models, and thus $\overline{C}_\sigma^{\text{prop}}$ and $\overline{C}_\sigma^{\text{conv}}$ will both be close to zero since $\overline{C}_\sigma \equiv 2\tilde{A}/S$. This discrepancy is also most likely attributable to rounding error.

There were eight pairs of values of the a and b parameters for which Δ_C was greater than 0.001. In each of these situations, the value of the b parameter was close to one. This indicates a situation in which the proper and conventional models yield nearly identical results, and the discrepancy can again be attributed to rounding error.

These results are readily justified from a theoretical standpoint given the definition of S and hence \overline{C}_σ . The arclength along any ROC curve can be expressed as

$$S \equiv \int_0^1 \sqrt{1 + \left(\frac{d\text{TPF}}{d\text{FPF}}\right)^2} d\text{FPF}. \quad (7)$$

But

$$\text{TPF} = \int_{x_0}^{\infty} p(x|\pi_1) dx \quad (8)$$

$$\text{FPF} = \int_{x_0}^{\infty} p(x|\pi_2) dx \quad (9)$$

where $p(x|\pi_1)$ is the probability density function for the decision variable values under the actually positive hypothesis (π_1), and $p(x|\pi_2)$ is the probability density function for the decision variable values under the actually negative hypothesis (π_2). The slope of the ROC curve is thus

$$\frac{d\text{TPF}}{d\text{FPF}} = \frac{\frac{d\text{TPF}}{dx_0}}{\frac{d\text{FPF}}{dx_0}} \quad (10)$$

$$= \frac{-p(x_0|\pi_1)}{-p(x_0|\pi_2)} \quad (11)$$

$$= \text{LR}(x_0), \quad (12)$$

where $\text{LR}(x_0)$ is the likelihood ratio of x evaluated at the particular value x_0 of the decision threshold corresponding to the point at which the slope of the ROC curve is being evaluated. This allows us to write Eq. 7 as

$$S = \int_{-\infty}^{\infty} \sqrt{1 + \text{LR}^2(x)} p(x|\pi_2) dx \quad (13)$$

$$= \left\langle \sqrt{1 + \text{LR}^2(\mathbf{x})} \right\rangle_{\pi_2}, \quad (14)$$

that is, the expected value of the square root of one plus the square of the likelihood ratio decision variable under the actually negative hypothesis. This implies that a given decision variable and its likelihood ratio will yield ROC curves of equal arclength, and thus differences in \overline{C}_σ can arise only from differences in AUC (appearing in the numerator of the definition of \overline{C}_σ). We should thus expect \overline{C}_σ to be lower for an ideal observer decision strategy than for any other decision strategy (since the AUC will be higher).

5. CONCLUSIONS

We have previously proposed a novel performance metric for ROC analysis in an attempt to address known difficulties in generalizing AUC to classification tasks with more than two classes. Numerical studies as well as theoretical analysis suggest that its behavior is consistent with that of the AUC performance metric, in the sense that the optimal value of this quantity is achieved by the ideal observer.

Although much work remains to be done in investigating the properties of this performance metric, our preliminary results are so far quite encouraging. We have high hopes that the SAEC 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

Charles E. Metz receives royalties from 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. D. C. Edwards and C. E. Metz, "A utility-based performance metric for ROC analysis of N -class classification tasks," in Proc. SPIE Vol. 6515 *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Berkman Sahiner, eds., pp. 6515031–65150310, (SPIE, Bellingham, WA), 2007.
9. 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.