Published in final edited form as:

IEEE Trans Med Imaging. 2009 February ; 28(2): 185–193. doi:10.1109/TMI.2008.928919.

The Validity of Three-Class Hotelling Trace (3-HT) in Describing Three-Class Task Performance: Comparison of Three-Class Volume Under ROC Surface (VUS) and 3-HT

Xin He [Member, IEEE] and

Department of Radiology, Johns Hopkins School of Medicine, Baltimore, MD 21287 USA (e-mail: xinhe@jhmi.edu).

Eric C. Frey [Senior Member, IEEE]

Department of Radiology, Johns Hopkins School of Medicine, Baltimore, MD 21287 USA (e-mail: efrey@jhmi.edu).

Abstract

In order to describe multiclass classification performance, several figures of merit (FOM) have been proposed. Among the earliest and most widely known of these is the three-class Hotelling trace (3-HT). The goal of this paper is to present theoretical and empirical data demonstrating the failure of 3-HT as a measure of three-class task performance. To help do this, we contrast it to a newly proposed three-class FOM, the volume under the three-class receiver operating characteristic (ROC) surface (VUS). The VUS is obtained from a decision theory based three-class ROC analysis method which has been proved to extend the decision theoretic, linear discriminant analysis (LDA), and psychophysical foundations of binary ROC analysis to a three-class paradigm. We demonstrate empirically that the VUS and 3-HT do not have a monotonic relationship in general when describing three-class task performance. Numerical experiments demonstrated that the VUS provided reasonable results, while the 3-HT failed to distinguish between the case where all objects could be perfectly classified from the case where only one pair of the classes could be perfectly classified. We have provided theoretical explanations of this failure of 3-HT. The significance of this work goes beyond merely demonstrating the problems of the 3-HT, it demonstrates that a FOM that is mathematically correct and has a strong theoretical basis can provide results that violate a common sense understanding of three-class task performance. This fact raises the question of “how to evaluate a classification performance evaluation method?” We believe the answer to this question lies in the theoretical foundations of binary ROC analysis. We have thus contrasted the two FOMs in terms of three fundamental theories underlying binary ROC analysis: decision theory, binary linear discriminant analysis, and the equivalence of two psychophysical classification procedures. These theoretical investigations demonstrated the importance of extending and unifying all the fundamental theories of binary classification in the development of a three-class FOM; violating one of these fundamental binary classification theories may, as it did for the L-HT, provide predictions of three-class task performance that do not agree with a common sense understanding of three-class task performance.

Index Terms

L-class Hotelling trace; L-class linear discriminant analysis; receiver operating characteristic (ROC) analysis; three-class classification

I. Introduction

Binary receiver operating characteristic (ROC) analysis has been a standard for describing binary classification performance. In order to describe multiclass classification performance, several figures-of-merit (FOM) have been proposed [1]–[23]. Different FOMs have different theoretical foundations. Based on decision theory, Edwards and Metz derived the general ROC surface in a 6-D hyperspace [3]–[8], [14]. However, as yet there is no practical index that can be used to summarize task performance. In fact, the volume under the hypersurface, a natural extension of the binary FOM has been shown not to be useful [3]–[8], [14]. Sahiner reduced the 6-D space to a 3-D space by adding constraints on utility and introduced the normalized volume under surface (NVUS) as a FOM [18]. Everson introduced a multiclass ROC analysis method based on multiobjective optimization, and resulted in a generalization of the Gini coefficient as a FOM [9]. Fukunaga extended the binary linear discriminant analysis to multiclass, and proposed the Hotelling trace as a FOM [24]. Scurfield extended the psychophysical foundation of binary ROC analysis into a three-class paradigm, and proposed to use a set of six 3-D surfaces spanned by the classification rates (e.g., true class 1 fraction, false class 1 fraction when the truth is class 2, etc.) to describe task performance [21]. In addition to three-class methods that are based on the extension of existing binary classification theories, there are many *ad hoc* extensions of binary ROC analysis. Examples are Hand's FOM which is a average of pairwise AUC values [10], Mossman's method, which also results in a three-class ROC surface and volume under this surface was suggested FOM [15], and the ordinal three-class ROC analysis that are proposed by many authors [13], [16], [20], [22], etc. Subsequent theoretical and statistical studies have also demonstrated many of the desirable properties of these three-class methods.

The aforementioned three-class methods provide different FOMs. Different FOMs serve as metrics for three-class performance, and different FOMs may provide different results on task performance ranking. As a result, one question related to all these methods is “*how to evaluate a classification performance evaluation method?*” One might think that a FOM would be valid if it has a strong theoretical foundation. The work presented in this paper clearly shows that 3-HT, which has a solid foundation in linear discriminant analysis (LDA), fails to be a useful three-class FOM. On the other hand, some of the three-class methods that do not have a specific theoretic foundation seem to work very well, and are thus widely used, such as the ordinal three-class ROC.

We still believe the answer to the question about evaluation of evaluation methods lies in the theoretical foundations of binary ROC analysis—not only one of the foundations, but more than one of them. In this paper, we have contrasted 3-HT and VUS in terms of three fundamental theories underlying binary ROC analysis: decision theory, binary LDA, and the equivalence of two psychophysical classification procedures. These theoretical investigations demonstrated the importance of extending and unifying all the fundamental theories of binary classification in the development of a three-class FOM; violating one of these fundamental binary classification theories may, as it did for the L-HT, result in rankings of three-class task performance that do not agree with a common sense understanding and/or are not useful for practical applications.

In following, we first theoretically show the limitations of 3-HT as a FOM and then demonstrate the limitations experimentally by comparing it with the VUS in predicting three-class task performance. We show that the two FOMs do not have a monotonic relationship and that the predictions from the L-HT do not adequately explain the differences in class performance in a logical way. Finally, we contrast the two FOMs under comparison in terms of the theoretical foundations for binary ROC analysis. This comparison indicates the importance of extending and unifying all the fundamental theories of binary classification in the development of a three-class FOM since, as seen in the case of the L-HT, violating one of the theoretical foundations can lead to FOMs that provide predictions of three-class task performance that are not meaningful.

A. L-Class Hotelling Trace (L-HT)

LDA was pioneered by Fisher and was originally only defined for the binary problem [25], where high-dimensional data are projected onto a 1-D space and the classification is performed in this 1-D space. This projection operation maximizes the distance between the means of the two classes relative to the average variance within each class. The Fisher criterion maximizes over all linear projections, $\vec{\omega}$

$$\text{SNR}^2 = c \frac{\vec{\omega}^T S_1 \vec{\omega}}{\vec{\omega}^T S_2 \vec{\omega}} \quad (1)$$

where $c = 4$, and S_1 and S_2 , are the interclass and intraclass scatter matrixes, respectively, which are given by

$$S_1 = \frac{1}{L} \sum_{i=1}^L (\vec{g}_i - \vec{g})(\vec{g}_i - \vec{g})^T \quad (2)$$

and

$$S_2 = \frac{1}{L} \sum_{i=1}^L \langle (\vec{g} - \vec{g}_i)(\vec{g} - \vec{g}_i)^T \rangle_i = \frac{1}{L} \sum_{i=1}^L K_i. \quad (3)$$

In (2) and (3), in binary for LDA, the symbol $\vec{g}_i$ denotes the mean vector of class i , $\vec{g}$ is the ensemble mean of vectors from all the classes, and K_i is the covariance matrix of the i th class. The trace of matrix $S_2^{-1} S_1$ has been proved to be directly related to the criterion in (1), i.e.,

$$J_2 = \text{tr}[S_2^{-1} S_1] = \frac{1}{4} \text{SNR}^2. \quad (4)$$

Therefore, the maximization of the Fisher criterion reduces to the following eigenvalue decomposition problem [24]:

$$S_2^{-1} S_1 \vec{\omega}_0 = \mu \vec{\omega}_0. \quad (5)$$

As a result, the trace of the matrix $S_2^{-1} S_1$, which equals the eigenvalue in (5), is proposed as a FOM for describing binary task performance via its relationship to signal-to-noise ratio (SNR). This trace is often termed the Hotelling trace in medical imaging literature [26], [27].

Multiclass (more than two classes) LDA is a natural extension of the binary LDA [24], where the two scatter matrixes are defined exactly the same way as in (2) and (3). Solving the eigenvalue decomposition problem in (5), the L-class Hotelling trace (L-HT) is given by

$$J_L = \text{tr}[S_2^{-1} S_1] = \sum_{i=1}^{L-1} \mu_i \quad (6)$$

where μ_i is the i th eigenvalue of $S_2^{-1} S_1$, and the corresponding eigenvectors are considered the optimal linear observer (or classifier) according to multiclass LDA [24], [26]. However, there are no known closed form solutions for either the eigenvectors or the eigenvalues in multiclass LDA, where more than two classes are involved. In addition, unlike binary LDA, where the Hotelling trace has a natural and intuitive explanation as described by (4), a corresponding interpretation of the L-HT in multiclass LDA is not known.

B. Three-Class ROC Analysis

Three-class classification is considerably more complex than the binary classification [7], [8]. We have previously developed a three-class ROC analysis method, which has been proved to extend and unify the decision theoretic, LDA, and psychophysical foundations of binary ROC analysis in a three-class paradigm [28]–[31].

In the following, a decision model refers to a set of decision variables and the decision rules. In the proposed optimal three-class decision model, a pair of log likelihood ratios are used as the optimal decision variables that span the decision space, as shown in Fig. 1 (a). The decision rules are represented by the decision structure (constructed by the bold lines) that partitions the decision space into three areas, where the decisions are made. The triplet of $(T1F, T2F, T3F)$ computed at each decision structure location maps out a three-class ROC surface [Fig. 1(b)] (this process is parallel to the one in binary ROC analysis, where the TPF and FPF pairs computed at each threshold maps out a binary ROC curve). Here, TiF represents the probability of correct decisions on class i and is also termed the sensitivity of class i . Each operating point on the surface thus represents the sensitivity triplet corresponding to one particular set of decision rules (i.e., a decision structure at a particular position) [28].

We have proved the optimality of the decision model with respect to several decision criteria under decision theory. In particular, assuming incorrect decisions have equal utilities under the same hypothesis (the equal error utility assumption) and given a set of decision utilities and prior probabilities, the proposed decision model maximizes the expected utility (MEU); it maximizes the probability of making correct decisions (MC) given a set of prior probabilities; it satisfies the Neyman-Pearson (N-P) criterion in the sense that, given a pair of sensitivities, the third sensitivity is maximized; and the maximum likelihood criterion produces one operating point on the ROC surface [29]. The ROC surface is thus a collection of optimal operating points that satisfy MEU, MC, and N-P criteria. When comparing two three-class ROC surfaces obtained from two diagnostic tests, if they do not cross, the diagnostic test corresponding to the higher ROC surface or a larger volume under three-class ROC surface (VUS) provides better task performance according to all the aforementioned decision criteria. Thus, the volume under three-class ROC surface (VUS) was proposed as a FOM for three-class task performance [28].

We then investigated the LDA foundation of three-class ROC analysis and proposed an optimal three-class linear observer, three-class Hotelling observer (3-HO), that maximizes the SNR of the test statistics between each pair of the classes simultaneously when the data satisfy a certain linear relationship. Furthermore, if the data are multivariate Gaussian

distributed with equal covariance matrixes, the 3-HO estimates three-class ideal observer (3-IO) performance [30].

Finally, we have demonstrated the relationship between VUS, obtained from a rating procedure, and percent correct, obtained from a corresponding categorization procedure. In the latter procedure, three randomly sampled objects, one from each of three distinct classes, are presented to the observer simultaneously. The observer is instructed to categorize the three objects into each of the three hypotheses. The observer is said to make a correct decision when and only when all three objects are correctly categorized. We have proved that VUS equals percent correct for continuous data. The equivalence is not dependent on the decision variables used, but does depend on the shape of the decision structure [31].

II. Theory

In the following, we demonstrated the limitations of 3-HT theoretically. For a three-class classification task, we first rewrote the interclass scatter matrix as

$$S_1 = \frac{1}{9} [(\vec{g}_1 - \vec{g}_2)(\vec{g}_1 - \vec{g}_2)^t + (\vec{g}_3 - \vec{g}_2)(\vec{g}_3 - \vec{g}_2)^t + (\vec{g}_1 - \vec{g}_3)(\vec{g}_1 - \vec{g}_3)^t]. \quad (7)$$

We defined the differences between the class means, $\vec{\psi}_1$ and $\vec{\psi}_2$, as

$$\begin{cases} \vec{\psi}_1 = \vec{g}_1 - \vec{g}_2 \\ \vec{\psi}_2 = \vec{g}_3 - \vec{g}_2 \end{cases}. \quad (8)$$

In general, $\vec{\psi}_1 \neq \vec{\psi}_2$, otherwise the three-class classification problem reduces to a binary classification problem. The difference between $\vec{g}_1$ and $\vec{g}_3$ is a linear combination of $\vec{\psi}_1$ and $\vec{\psi}_2$, i.e.,

$$\vec{g}_1 - \vec{g}_3 = \vec{\psi}_1 - \vec{\psi}_2. \quad (9)$$

Then the interclass scatter matrix S_1 can be expressed as

$$S_1 = \frac{1}{9} [2\vec{\psi}_1\vec{\psi}_1^t + 2\vec{\psi}_2\vec{\psi}_2^t - \vec{\psi}_1\vec{\psi}_2^t - \vec{\psi}_2\vec{\psi}_1^t]. \quad (10)$$

Every vector, $\vec{\psi}$, in the range of S_1 can be expressed as a linear combination of $\vec{\psi}_1$ and $\vec{\psi}_2$, i.e.,

$$\begin{aligned} S_1 \vec{\psi} &= \frac{1}{9} [2\vec{\psi}_1\vec{\psi}_1^t + 2\vec{\psi}_2\vec{\psi}_2^t - \vec{\psi}_1\vec{\psi}_2^t - \vec{\psi}_2\vec{\psi}_1^t] \vec{\psi} \\ &= a\vec{\psi}_1 + b\vec{\psi}_2 \end{aligned} \quad (11)$$

where a and b are scalars given by

$$a = \frac{1}{9} [2\vec{\psi}_1^t - \vec{\psi}_2^t] \vec{\psi}$$

and

$$b = \frac{1}{9} [2\vec{\psi}_1 - \vec{\psi}_1'] \vec{\psi} \quad (12)$$

(Note that $\vec{\psi}' \vec{\psi}$ is a scalar). From this we see that the range of S_1 is spanned by $\vec{\psi}_1$ and $\vec{\psi}_2$ and, as long as the differences in class means in (8) are not linear combinations, the rank of S_1 is two [26]. As a result, the matrix $S_2^{-1} S_1$ has rank two, suggesting there are two features to be extracted from it. The two features are characterized by the two eigenvalues and the corresponding eigenvectors, respectively. That is to say, the 3-HT is expressed as

$$J_3 = \mu_1 + \mu_2 \quad (13)$$

where μ_i is the i th eigenvalue. It follows that as long as one eigenvalue is large, the 3-HT is large, suggesting near-perfect overall task performance, even if the other eigenvalue is sufficiently small, indicating a failure in distinguishing the other feature. Thus, the 3-HT cannot distinguish the cases where it can perfectly classify both features from the cases where it can perfectly distinguish only one feature. Consequently, the value of 3-HT as a FOM should be questioned. To provide better understanding of the limitations of 3-HT, we considered the following special case.

A Special Case: Data With Equal Covariance

Assume that all three classes have equal covariance, i.e.,

$$K_1 = K_2 = K_3 = K. \quad (14)$$

Substituting (7) in to (6), the 3-HT can be expressed as

$$\begin{aligned} J_3 &= \text{tr}[S_2^{-1} S_1] \\ &= \frac{1}{9} \text{tr}[K^{-1}(\vec{g}_1 - \vec{g}_2)(\vec{g}_1 - \vec{g}_2)'] \\ &\quad + \frac{1}{9} \text{tr}[K^{-1}(\vec{g}_3 - \vec{g}_2)(\vec{g}_3 - \vec{g}_2)'] \\ &\quad + \frac{1}{9} \text{tr}[K^{-1}(\vec{g}_1 - \vec{g}_3)(\vec{g}_1 - \vec{g}_3)']. \end{aligned} \quad (15)$$

Comparing (15) with (4), J_3 becomes

$$J_g = \frac{1}{9} (\text{SNR}_{12}^2 + \text{SNR}_{32}^2 + \text{SNR}_{13}^2) \quad (16)$$

where $\text{SNR}_{ij}^2 (i, j=1, 2, 3; i \neq j)$ is the square of the SNR of the two-class Hotelling test statistics between classes i and j . Equation (16) shows the 3-HT is proportional to quadrature sum of the SNR of each pair of the three classes. Each SNR ranges from 0 to infinity. Therefore, as long as one SNR goes to infinity, the 3-HT, J_3 , goes to infinity, suggesting perfect classification, independent of the value of the other SNR. Although in the meanwhile, another SNR might be zero, suggesting two of the three classes are completely inseparable. For example, when classes 1 versus 2 are perfectly classified, while classes 2 versus 3 are degenerate, J_3 is very large because the SNR between classes 1 and 2 is very large, although the SNR between classes 2 and 3 is zero. This clearly demonstrates a major limitation of 3-HT as a three-class FOM.

III. Methods

The purpose of the experiments is to compare the usefulness of 3-HT and VUS in describing three class task performances. The results of these experiments demonstrated the limitations

of 3-HT. We first designed a three-class task with easily manipulated task performance. We then performed a three-class channelized Hotelling observer (3-CHO) study to obtain three-class data and compute the 3-HT and VUS over a range of task difficulties.

A. Task Design

From decision theory, likelihood ratios are the optimal decision variables. Making a decision thus requires computing the likelihood ratios for a set of input data. In three-class LDA, the eigenvectors of the matrix $S_2^{-1}S_1$ are used to extract two features from the input data. The strength of each of these two features determines the 3-HT, and thus the task performance. Therefore, it is desirable to have a three-class task model that allows us to easily manipulate the distribution of the two likelihood ratios as well as the two features to control the task performance.

We propose the two-signal task model shown in Fig. 2(a). In this task model, each data element consists of two images: Image A and Image B. The three classes are defined by the combined presence or absence of signals in images A and B. Fig. 2 illustrates this task, where the gray dot denotes the presence of a circular mesa signal with magnitude S_A on the center of image A, and the black dot denotes the presence of the same shaped signal but with magnitude S_B on the center of image B. Note that classes 2 and 3 have the same signal on image A. In addition, we defined statistically identical backgrounds on image A for classes 1, 2, and 3, and we defined statistically identical backgrounds on image B for classes 1, 2, and 3. The statistical properties of the backgrounds of image A and B may be different. Note that the second column of images is defined as class 3 and the third column as class 2 for mathematical convenience.

We used Gaussian filtered white noise as the background. Random Gaussian noise with zero mean and variance σ_A and σ_B were then added to the backgrounds of image A, and B, respectively. All images were 32×32 pixels in size. To create an ensemble, 500 samples of objects were generated for each class for a particular set of signal/noise parameters. Fig. 2(b) shows sample images. It can be seen that the above design results in three classes of objects with same statistical background and noise properties. Since the signals are independent of the background or the noise, this task design results in three classes with equal covariance matrixes.

We now explain how this task design enabled us to manipulate the distributions of the two likelihood ratios and the two features of LDA easily to control the task performance. In particular, since image B of Classes 1 and 3 had the same statistical properties, the likelihood ratio of class 1 over class 3, LR_{13} , is mainly determined by the likelihood ratio for signal presence on image A. For the same reason, LR_{23} , is mainly determined by the likelihood ratio of signal presence on image B. Thus, manipulating the magnitudes of signals A and B in relative to the noise allowed us to manipulate the distributions of the likelihood ratios. In addition, we see that

$$\vec{\psi}_1 = \vec{g}_1 - \vec{g}_2 = \begin{bmatrix} -\vec{g}_A \\ 0 \end{bmatrix} \text{ and } \vec{\psi}_2 = \vec{g}_3 - \vec{g}_2 = \begin{bmatrix} 0 \\ \vec{g}_B \end{bmatrix} \quad (17)$$

where $\vec{g}_A$ and $\vec{g}_B$ are the noise-free images A and B with signals and S_A and S_B , respectively. Further, the additive noise is not dependent on the signals. The eigenvectors of S_1 can be treated as the linear combinations of $\vec{\psi}_1$ and $\vec{\psi}_2$. Manipulating the magnitudes of $\vec{\psi}_1$ and $\vec{\psi}_2$ would enable us to manipulate the two features extracted by LDA. Note that $\vec{\psi}_1$ and $\vec{\psi}_2$ are not the eigenvectors of $S_2^{-1}S_1$, and thus do not represent the two features directly.

Their magnitudes, however, determine the class separabilities. Manipulating them thus enabled us to manipulate the two features indirectly.

It can be seen that by increasing the signal magnitude on image A while decreasing the signal magnitude on image B, the separability of class 1 versus classes 2 and 3 will be improved, while the separability of classes 1 and 3 versus 2 will be degraded. The former will increase SNR_{12}^2 and SNR_{13}^2 , and the latter will decrease SNR_{12}^2 and SNR_{13}^2 . The overall task performance depends on the tradeoffs among the three SNRs of the three classes.

B. Obtaining the Three-Class Data

Although the images were small in size, the dimension of the image vector 2048×1 still made it very computationally difficult to precisely calculate the covariance matrixes for each class and to invert it to obtain the inverse of S_2 . To reduce dimensionality, we applied the three-class channelized Hotelling observer (3-CHO). In particular, a 3-CHO converts the data vector into low dimension feature vectors. For this reason, we conducted the following observer study using feature vectors rather than data vectors. To generate a feature vector, images A and B of each object were processed using an appropriate set of channels, $U = [\vec{u}_1 \ \vec{u}_2 \ \dots \ \vec{u}_N]$, where N is the number of channels. The feature vectors for both images A and B, $\vec{v}_A$ and $\vec{v}_B$, were then stacked together to form a feature vector $\vec{v}$ for the data vector $\vec{g}$. This process can be expressed as

$$\begin{aligned} \vec{v} &= \begin{bmatrix} \vec{v}_A \\ \vec{v}_B \end{bmatrix} \\ &= \begin{bmatrix} \vec{u}_1^\dagger \vec{g}_A & \vec{u}_2^\dagger \vec{g}_A & \dots & \vec{u}_N^\dagger \vec{g}_A & \vec{u}_1^\dagger \vec{g}_B & \vec{u}_2^\dagger \vec{g}_B & \dots & \vec{u}_N^\dagger \vec{g}_B \end{bmatrix}^t \end{aligned} \quad (18)$$

where $\vec{g}_A$ and $\vec{g}_B$ are the image vectors for images A and B, respectively. Equation (18) is equivalent to constructing a $2N$ channel set, Ψ

$$\Psi = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \dots & \vec{u}_N & \vec{0} & \vec{0} & \dots & \vec{0} \\ \vec{0} & \vec{0} & \dots & \vec{0} & \vec{u}_1 & \vec{u}_2 & \dots & \vec{u}_N \end{bmatrix} \quad (19)$$

where $\vec{0}$ is a zero column vector M of elements, and M is the number of pixels in image A or B. The feature vector $\vec{v}$ is then expressed as

$$\vec{v} = \Psi^\dagger \vec{g}. \quad (20)$$

The 3-HT and VUS values were calculated using the feature vectors instead of the original image vectors. The channels used in this were the four octave-wide rotationally symmetric frequency channels as shown in Fig. 3.

C. Computing the VUS and 3-HT

As described above, we first obtained the feature vectors of the image ensemble; the 3-HT and VUS values were estimated using the ensemble of resulting feature vectors instead of the original image vectors. To calculate VUS, we performed a 3-CHO study [30]. Using half of the feature vectors, we trained a 3-CHO to obtain two 3-CHO templates

$$\begin{aligned} \vec{\omega}_{13} &= (K_1 + K_3)^{-1} (\vec{v}_1 - \vec{v}_3) \\ \vec{\omega}_{23} &= (K_3 + K_2)^{-1} (\vec{v}_2 - \vec{v}_3) \end{aligned} \quad (21)$$

where $\vec{\bar{v}}_i$ is the mean feature vector of the i th class and $\vec{\omega}_{ij}$ is the Hotelling template between classes i and j . In the experiment, we used three classes of objects with equivalent covariance matrixes, i.e., $K_1 = K_2 = K_3$. Thus, the condition for the existence of a 3-HO that simultaneously maximizes the SNR between each pair of the classes [30], i.e.,

$$\vec{\omega}_{13} = \vec{\omega}_{12} - \vec{\omega}_{23} \quad (22)$$

could be easily satisfied using this task design. We then applied the 3-CHO templates to the remaining half of the feature vectors to obtain a pair of test statistics for each feature vector using and

$$\lambda_{13} = \vec{\omega}_{13}^t \vec{v}$$

and

$$\lambda_{23} = \vec{\omega}_{23}^t \vec{v}. \quad (23)$$

The test statistics of all feature vectors in the testing set were collected to obtain a three-class decision plane, and the statistical algorithm developed in [31] was used to compute the VUS.

To calculate the 3-HT, we first computed the mean and variance of the ensemble of feature vectors for each class, and then used (2), (3), and (6) to compute the 3-HT.

IV. Experiments and Results

In the following we estimated the VUS and 3-HT for ensembles of objects obtained with various combinations of parameters to investigate whether these two metrics provide information about the tradeoffs in image quality and provide reasonable predictions of task performance. We designed two experiments using the three classes illustrated in Fig. 2. We fixed the variance of the Gaussian noise on both images A and B, i.e.,

$$\sigma_A = 5 \text{ and } \sigma_B = 5 \quad (24)$$

while the magnitude of the signals on both images, S_A and S_B , were adjusted to vary the separabilities of the three classes. We refer to each combination of S_A and S_B as a data treatment. In the following experiments, we computed one VUS value and one 3-HT value for each data treatment.

A. Experiment 1

In this experiment, we fixed the magnitude of signal in image A, S_A , and varied the signal magnitude in image B, S_B . To be specific

$$S_A = 3$$

and

$$S_B = 0.001, 0.05, 0.1, 0.5, 0.8, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 50, 100, 200 \quad (25)$$

For each data treatment, we obtained one VUS value and the corresponding 3-HT. Fig. 4 and Fig. 5 demonstrate how the task performance as measured by 3-HT and VUS changes with the increase of S_B . Fig. 6 demonstrates the relationship between VUS and 3-HT as S_B increases.

Fig. 4 and Fig. 5 show that, as the magnitude of the signal on image B decreased, the VUS decreased and converged to 0.320, while the 3-HT decreased and converged to 0.234. Further, as the magnitude of the signal in image B increased, the VUS increased and converged to 0.765, while the 3-HT increased rapidly, seemingly without bound. Note that in Fig. 4, the two figures have different y -axis scales. Fig. 6 shows that the VUS and 3-HT are monotonically related, as is the case for two-class task, indicating that VUS and 3-HT predict the changes of task performance uniquely.

This result is not surprising, because when S_A is fixed and S_B decreases, the performance on the task should degrade as S_B decreases and converges when the classification into class 2 versus classes 1 and 3, which depends only on the ability to detect S_B , becomes a pure guess. To this end, both VUS and 3-HT behaved as predicted. On the other hand, when S_A is fixed and S_B increases, the performance on the task should improve as S_B increases and converges when the classification into class 2 versus classes 1 and 3 is perfect, while the classification into class 1 versus classes 2 and 3 is confined by the fixed S_A . As can be seen, the VUS successfully predicted this trend, but the 3-HT did not: as S_B increases, the 3-HT continues to increase rapidly.

B. Experiment 2

The parameters for Experiment 2 are shown in Table I. Compared with Experiment 1, an additional value of S_A was used. Therefore, we varied the signal-to-noise ratios in both Images A and B at the same time; the overall task performance was thus dependent on both images. The purpose of this experiment was to show how the task performance changes in terms of both 3-HT and VUS when S_A was decreased from 6.0 to 3.0 while S_B was increased from 0.001 to 200.

The VUS versus 3-HT plot is shown in Fig. 7. Note the two plots have different scales on x axis. It can be seen that, when signal magnitudes on both images change together, 3-HT and the VUS are not monotonically related, indicating they do not provide a consistent measure of task performance. More specifically, while 3-HT suggests that by changing the parameter, the task performance is getting better, the VUS might suggest otherwise.

To further demonstrate the limitations of 3-HT, we selected three data treatments from Experiment 2, denoted by i, ii, and iii; the values of the parameters for these sets are shown in Table II. Also shown in Table II are the 3-HT and VUS values for the corresponding dataset. It can be seen that VUS suggested that the task performance for set iii was almost the same as that for set ii and was significantly lower than that for set i, while 3-HT suggested differently.

For each set of the data treatment, we calculated the corresponding 2-HT and AUC for each pair of the classes, as shown in Table II. Before trying to explain the results, we first noted that the AUC value is equivalent to the percent correct in a 2-alternative forced choice (2AFC) task. By equivalent, we mean that, when an investigator calculates the AUC value from a rating procedure, “he is in fact, or at least in mathematical fact, reconstructing random pairs of images, one from a diseased subject and one from a normal subject, and using the reader’s separate ratings of these two images to simulate what the reader would have decided if these two images had in fact been presented together as a pair in a 2AFC experiment” [32]. When the AUC is close 1.0, task performance is close to perfect. Further

increasing the separability of the two classes results in the AUC asymptotically approaching 1, which indicates perfect classification.

As shown in Table III, classifications between Classes 1 versus 2 (second row) and between Classes 3 versus 2 (third row) are almost perfect for all data treatment i, ii, and iii (all have AUC values greater than 0.99, suggesting in all cases the observer makes over 99% correct decisions in the corresponding 2AFC task). Comparing data treatment sets ii and iii, the AUC values between each pair of the classes are very similar, suggesting similar task performance between sets ii and iii. Comparing parameter sets i and iii, the AUC value for classes 1 versus 3 is significantly higher in set i than in set iii. Intuitively, we would conclude that the data in set i should result in better performance than those in set iii, and set ii results in similar performance as compared to set iii. We see that the VUS successfully predicted these trends, while the 3-HT failed to do so.

V. Discussion

Although classic binary ROC analysis has been a standard method in medical diagnostic performance evaluations, it has many limitations for medical applications. Using medical imaging as an example, classic ROC analysis is limited in handling cases where more than two diagnostic alternatives are involved, etc. To address this limitation, many multiclass extensions of the classic ROC analysis have been proposed [1], [2], [4], [13], [18], [20], [33]. In addition to ROC-type approaches, many other approaches have been proposed to describe multiclass classification performance including the L-HT from multiclass LDA discussed in this work [24], [26]. Some of these methods have attempted to extend the relevant binary classification theories into multiclass, while others intuitively extended binary classification FOMs. While these approaches may well be mathematically and algorithmically correct, the results of this paper demonstrate for the L-HT, despite its mathematical correctness and strong theoretical foundation in LDA, provides misleading measures of three-class task performance.

The limitations of the L-HT as a measure of three-class task performance raise a general question about how to evaluate methods for evaluating task performance. We believe that the results of this paper suggest that one should look at the fundamental theoretical bases. In the following, we present a comparison of the theoretical bases of the L-HT and the VUS, which serves as the FOM for our proposed three-class ROC analysis method. It is our hope that this discussion will provide insight into the criteria that should be used in future evaluations of task performance metrics.

Binary ROC analysis has strong theoretical justifications that describe several fundamentally desirable properties of classification metrics. In particular, a doctor or decision maker always does his/her best to deliver a decision. The decision theoretic foundation of ROC analysis mathematically models the requirement that decision makers “do their best” by modeling the decision as a process of one of finding the maximum (or minimum) of a relevant scalar quantity. The scalar quantity could be the utility, the probability of making correct decisions, or the TPF for a given FPF, etc., for different decision criteria. The second foundational property is psychophysical. In binary ROC analysis, this means that the FOM, i.e., the AUC value, is equivalent to the percent correct in a corresponding 2AFC procedure. This equivalence provides a probabilistic meaning for the AUC which ranges from 0.5 (pure guessing) to 1.0 (perfect classification). The probabilistic meaning of AUC value can be described as the “probability of correctly ranking a (normal, abnormal)” pair [32], [34]. The third foundational theory is LDA. The LDA foundation of ROC analysis, in addition to its rigorous mathematical form, suggests another meaning of classification. Suppose the rating values of the two classes form two Gaussian distributions, the FOM of LDA, the LDA SNR,

is a ratio of the distance between the two Gaussian distributions and the sum of the variances. Intuitively, either increasing the distance or decreasing the variance improves the classification of the two classes. In other words, decreasing the overlap between the two Gaussian distributions improves the classification performance. In general, ROC analysis, despite its simple form, unifies not only the rigorous mathematical forms of the aforementioned theoretical foundations, but, most importantly, underlying meaning of classification implied by these theoretical foundations.

In previous publications, we have developed a three-class ROC analysis method that extends and unifies the decision theoretic [28], [29], LDA [30], and psychophysical foundations [31] of ROC analysis in a three-class paradigm. We believe the extension and unification of the mathematical forms of the theoretical foundations also extends and unifies the desirable properties of a classification metric in a three-class paradigm. Multiclass LDA, on the other hand, attempted to extend the mathematical form of the binary LDA. However, as demonstrated here, this fails to extend the psychophysical foundation. In other words, 3-HT did not distinguish between the case where images could be perfectly classified for all classes (percent correct = 1.0) from the case where only one of the classes could be perfectly classified (percent correct < 1.0). In addition, it fails to extend the underlying principle of binary LDA, i.e., that decreasing the overlap among the distributions improves the classification performance.

The comparison performed in this paper thus suggests the importance of extending and unifying the theoretical foundations of classic ROC analysis when developing classification performance measures: violating one of the theoretical foundations can lead to a failure to satisfy the desirable properties of a classification metric. More theoretical investigations should be devoted to understand the theoretical foundations of binary ROC analysis and the philosophical meaning of classification metrics that form its foundation.

VI. Conclusion

We have previously showed that the VUS can be used as a FOM for classification performance in a three-class task [28]–[31]. In this paper, we questioned the validity of 3-HT, which is a widely used three-class performance measure, as a FOM for task performance. In particular, we first showed theoretically that the 3-HT can be decomposed into the algebraic sum of two traces, both of which are unbounded. As a result, the 3-HT, which is the sum of these two unbounded traces, provides a FOM that, in some cases, does not rank performance in a meaningful way. We then demonstrated experimentally that while rankings of class performance provided by the VUS were reasonable, the 3-HT provided rankings that did not make intuitive sense and did not correlate with those provided by the VUS. This questions the validity of using 3-HT as a FOM in evaluating three-class classification tasks, and the VUS provided by the proposed three-class ROC analysis is a better FOM for describing three-class task performance.

The significance of this work goes beyond a mere comparison of two FOMs; it demonstrated a somewhat puzzling fact: a FOM that is mathematically correct with strong theoretical roots turns out to fail to satisfy the desirable properties of a classification metric. This fact further raised the question of “*how to evaluate a classification performance evaluation method?*” We suggested that the answer to this question lies in the theoretical and philosophical aspects of binary ROC analysis, and thus reviewed the two FOMs in terms of three fundamental theories of binary classification: decision theory, binary linear discriminant analysis, and the equivalence of two psychophysical classification procedures. These theoretical investigations demonstrated the importance of extending and unifying all the fundamental theories of binary classification in the development of a three-class FOM:

violating one of the fundamental binary classification theories might also lead to failure to satisfy the desirable properties of a classification metric.

Acknowledgments

This work was supported in part by the National Institutes of Health (NIH) under Grant K99 EB007620-01, Grant R01-EB000288, and Grant R01-HL068575. The content of this work is solely the responsibility of the authors and does not necessarily represent the official view of the NIH or its various institutes.

The authors would like to thank their colleagues, Dr. H. H. Barrett, Dr. K. J. Myers, Dr. J. M. Links, Dr. B. M. W. Tsui, and Dr. J. Xu for the helpful comments and thought-provoking discussions with regard to this work.

References

1. Chan, H-P.; Sahiner, B.; Hadjiiski, LM.; Petrick, N.; Zhou, C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study; Proc. SPIE; 2003. p. 567-578.
2. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med. Decision Making. 2000; vol. 20:323-331.
3. Edwards, DC. Ideal Observer Estimation and Generalized ROC Analysis for Computer-Aided Diagnosis. Chicago, IL: Univ. Chicago; 2003.
4. Edwards DC, Lan L, Metz CE, Giger ML, Nishikawa RM. Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions. Med. Phys. 2004; vol. 31:81-90. [PubMed: 14761024]
5. Edwards DC, Metz CE. Restrictions on the three-class ideal observer's decision boundary lines. IEEE Trans. Med. Imag. 2005 Dec; vol. 24(no. 12):1566-1573.
6. Edwards DC, Metz CE. Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule. J. Math. Psychol. 2006; vol. 50:478-487.
7. Edwards DC, Metz CE, Kupinski MA. Ideal observers and optimal ROC hypersurfaces in N-class classification. IEEE Trans. Med. Imag. 2004 Jul; vol. 23(no. 7):891-895.
8. Edwards DC, Metz CE, Nishikawa RM. The hypervolume under the ROC hypersurface of "near-guessing" and "near-perfect" observers in N-class classification tasks. IEEE Trans. Med. Imag. 2005 Mar; vol. 24(no. 3):293-299.
9. Everson RM, Fieldsend JE. Multi-class ROC analysis from a multi-objective optimisation perspective. Pattern Recognit. Lett. 2006; vol. 27:918-927.
10. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. 2001; vol. 45:171-186.
11. Heckerling PS. Parametric three-way receiver operating characteristic surface analysis using mathematica. Med. Decision Making. 2001; vol. 21:409-417.
12. Johnson AJ, Noga AJ, Kosoy O, Lanciotti RS, Johnson AA, Biggerstaff BJ. Duplex microsphere-based immunoassay for detection of anti-West Nile virus and anti-St. Louis encephalitis virus immunoglobulin M antibodies. Clinical and Diagnostic Lab. Immunol. 2005; vol. 12:566-574.
13. Kijewski MF, Swensson RG, Judy PF. Analysis of rating data from multiple-alternative tasks. J. Math. Psychol. 1989; vol. 33:428-451.
14. Metz, CE. Handout in the Course Mathematics for Medical Physicists. Chicago, IL: Univ. Chicago; 2000. The optimal decision variable: Likelihood ratio.
15. Mossman D. Three-way ROCs. Med. Decision Making. 1999; vol. 19:78-89.
16. Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Stat. Med. 2004; vol. 23:3437-3449. [PubMed: 15505886]
17. Patel, AC.; Markey, MK. Comparison of three-class classification performance metrics: A case study in breast cancer CAD; SPIE Med. Imag. Conf; San Diego, CA: 2005.
18. Sahiner B, Chan HP, Hadjiiski LM. Performance analysis of three-class classifiers: Properties of a 3-D ROC surface and the normalized volume under the surface for the ideal observer. IEEE Trans. Med. Imag. 2008 Feb; vol. 27(no. 2):215-227.

19. Sahiner, B.; Chan, H-P.; Hadjiiski, LM. Performance analysis of three-class classifiers: Properties of the 3-D ROC surface and the normalized volume under the surface; *Proc. SPIE*; 2006.
20. Scurfield BK. Multiple-event forced-choice tasks in the theory of signal detectability. *J. Math. Psychol.* 1996; vol. 40:253–269. [PubMed: 8979976]
21. Scurfield BK. Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks. *J. Math. Psychol.* 1998; vol. 42:5–31. [PubMed: 9606159]
22. Xiong CJ, Belle Gv, Miller JP, Morris JC. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Stat. Med.* 2006; vol. 25:1251–1273. [PubMed: 16345029]
23. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics.* 2005; vol. 21:2394–2402. [PubMed: 15713736]
24. Fukunaga, K. *Introduction to Statistical Pattern Recognition*. New York: Academic; 1990.
25. Fisher RA. The use of multiple measures in taxonomic problems. *Ann. Eugenics.* 1936; vol. 7:179–188.
26. Barrett, HH.; Myers, KJ. *Foundations of Image Science*. New York: Wiley; 2003.
27. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. *Proc. Nat. Acad. Sci. USA.* 1993; vol. 90:9758–9765. [PubMed: 8234311]
28. He X, Metz CE, Tsui BMW, Links JM, Frey EC. Three-class ROC analysis—A decision theoretic approach under the ideal observer framework. *IEEE Tran. Med. Imag.* 2006 May; vol. 25(no. 5): 571–581.
29. He X, Frey EC. Three-class ROC analysis—The equal error utility assumption and the optimality of three-class ROC surface using the ideal observer. *IEEE Trans. Med. Imag.* 2006 Aug; vol. 25(no. 8):979–986.
30. He X, Frey EC. An optimal three-class linear observer derived from decision theory. *IEEE Trans. Med. Imag.* 2007 Jan; vol. 26(no. 1):77–83.
31. He X, Frey EC. The meaning of the volume under the three-class ROC surface (VUS). *IEEE Trans. Med. Imag.* 2007 May; vol. 27(no. 5):577–588.
32. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982; vol. 143:29–36. [PubMed: 7063747]
33. Swets, JA.; Pickett, RM. *Evaluation of Diagnostic System: Methods from Signal Detection Theory*. New York: Academic; 1982.
34. Bamber D. Area above ordinal dominance graph and area below receiver operating characteristic graph. *J. Math. Psychol.* 1975; vol. 12:387–415.

Fig. 1.

(a) Three-class decision space. $\log LR_{13}$ and $\log LR_{32}$ are the optimal decision variables that span the decision space. (b) An example of the three-class ROC surface. TiF is the probability that class i is correctly classified.

Fig. 2.

(a) Definition of the three-class classification task model. (b) Sample images. Note that we deliberately used a very high SNR on the sample images for illustration purpose. The actually SNRs used in the experiments are much lower. Also note that the second column of images is defined as class 3 and the third column as class 2 for mathematical convenience.

Fig. 3. Octave-wide rotationally symmetric frequency channels and their corresponding spatial domain templates.

Fig. 4. Relationship between the 3-HT versus S_B as S_B increases. Note that the scales of both the x and y axes are different in the two figures. (a) Value of 3-HT as a function of S_B over the range 0.001 to 9. (b) Value of 3-HT as a function of S_B over the range 10 to 200.

Fig. 5. Relationship between the VUS versus S_B as S_B increases. (a) Value of VUS as a function of S_B over the range 0.001 to 9. (b) Value of VUS as a function of S_B over the range 10 to 200. Note that the scale of the x axis is different in the two plots.

Fig. 6.
VUS versus 3-HT plot for Experiment 1.

Fig. 7.
VUS versus 3-HT plot for Experiment 2. Note that the two plots have different scales in x axis (3-HT).

TABLE I

Parameters for Experiment 2

S_A	S_B
6.0, 3.0	0.001~200

TABLE II

Comparison of 3-HT and VUS for three data treatment

No	S_A	S_B	3-HT	VUS
i	6	10	3.66742	0.916149
ii	3	10	2.90474	0.763334
iii	3	15	5.89026	0.768114

Treatment	3-HT	VUS
i	3.66742	0.916149
ii	2.90474	0.763334
iii	5.89026	0.768114

TABLE III

Comparison of 2-HT and AUC for each pair of the three classes

	i. $S_A=6, S_B=10$			ii. $S_A=3, S_B=10$			iii. $S_A=3, S_B=15$		
	2-HT	AUC		2-HT	AUC		2-HT	AUC	
Classes 1 vs. 3	1.0228	0.92368		0.2778	0.771981		0.26775	0.76785	
Classes 1 vs. 2	4.1672	0.99806		3.2732	0.994745		6.6737	0.99987	
Classes 3 vs. 2	3.0617	0.99333		2.9846	0.992721		6.3117	0.99981	