AD_____

Award Number: W81XWH-04-1-0495

TITLE: Investigation of Three-Group Classifiers to Fully Automate Detection and Classification of Breast Lesions in an Intelligent CAD Mammography Workstation

PRINCIPAL INVESTIGATOR: Darrin C. Edwards, Ph.D.

CONTRACTING ORGANIZATION: University of Chicago
Chicago, IL 60637

REPORT DATE: May 2006

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-05-2006	Annual	1 May 2005 - 30 Apr 2006

4. TITLE AND SUBTITLE
Investigation of Three-Group Classifiers to Fully Automate Detection and Classification of Breast Lesions in an Intelligent CAD Mammography Workstation

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0495

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Darrin C. Edwards, Ph.D.

E-Mail: d-edwards@uchicago.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Chicago
Chicago, IL 60637

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT:

Our goal is to develop a fully automated classification scheme for a computer-aided diagnosis in mammography. Our Proposed scheme would classify computer detections into three groups: malignant lesions, benign lesions, and false-positive computer detections. During the past year, we have collected a database of 134 mammography cases with clustered microcalcification lesions. We have shown that three decision boundary lines used by three-group ideal observer are intricately related to one another. We have analyzed several recently proposed three-group classification methods in terms of the three-group ideals observer. Finally, we have developed principled theoretical motivations for various proposed three-group classification methods, given the selections of restricted or simplified three-group evaluation methods. A three-group classifier could potentially allow radiologists to detect more malignant breast lesions without increasing their false-positive biopsy rates.

15. SUBJECT TERMS
Computer-aided Diagnosis, X-ray Imaging

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	74	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

1 Introduction

Our goal is to develop a fully automated classification scheme for computer-aided diagnosis (CAD) in mammography. Traditional CAD classification schemes, and performance measurement tools such as receiver operating characteristic (ROC) analysis, are based on the premise that the observations are classified into two groups, most commonly malignant and benign. Such classification schemes are difficult to fully automate, as they analyze radiologist-identified lesions; this is because many false-positive (FP) detections produced by a computerized detection scheme cannot reasonably be classified as benign or malignant lesions. Our proposed scheme would classify computer detections into three groups: malignant lesions, benign lesions, and FP computer detections. This method presents considerable difficulties in terms of both signal detection theory and performance evaluation methods such as ROC analysis. Our efforts in this direction have thus generally been more theoretical than practical so far, but our results so far are promising.

2 Body

A wide variety of medical decision-making tasks, in particular tasks for which CAD has been proposed as an aid to the physician, can be formulated as "two-group classification" tasks. That is, the physician must use the information available about a patient (*e. g.*, a set of mammographic films of the patient, and the result of computer analysis of those images) to decide whether a patient belongs to a diseased, or abnormal, group or not (*e. g.*, whether a breast lesion suspicious enough to warrant further imaging procedures or biopsy is present or not).

ROC analysis has long been considered the most appropriate methodology for evaluating the performance of a two-group classifier or observer [1], particularly for medical decision-making tasks [2]. Furthermore, the optimal or "ideal" observer — that observer which achieves the best possible performance given a particular population of observational data — has also been well understood for quite some time [3]. In practice, the ideal observer requires knowledge of the probability density functions (PDFs) from which the observational data are drawn, and thus cannot be achieved in non-trivial tasks by human or automated observers. Nevertheless, successful methods for estimating ideal observer decision variables from a sample of observational data [4], and for plotting an ideal observer ROC curve from a sample of decision variable data [5], have been developed.

Although the form of the three-group ideal observer has also been known for some time [3], the development of a practical three-group classifier and a fully general extension of ROC analysis to three-group classification has proven quite difficult, primarily due to the tremendous increase in complexity encountered when one moves from two-group to three-group classification tasks. Briefly, characterizing the performance of a three-group classifier requires an ROC "hypersurface" with five degrees of freedom in a six-dimensional ROC space [6, 7] (by contrast, a two-group classifier is fully described by a simple curve in a two-dimensional ROC space). Despite these difficulties, our research efforts are focused on the development of a three-group classifier and performance evaluation methodology for breast lesion classification in a mammographic CAD system.

We strongly believe the development of such a three-group classifier to be of practical and not merely academic importance. In the past, two types of mammographic CAD schemes

have been investigated at the University of Chicago: one for automatically detecting mass lesions in mammograms [8–12], and one for classifying known lesions as malignant or benign [13–17]. Combining these two types of CAD scheme is inherently difficult, because the output of the detection scheme will necessarily include FP computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our approach has been to treat this problem explicitly as a three-group classification task. That is, the output of the detection scheme should be classified as malignant lesions, benign lesions, and non-lesions (FP computer detections), and the classifier to be estimated is the ideal observer decision function for this task. If successful, this approach would allow radiologists to identify more malignant lesions without increasing biopsy rates for patients without malignancy.

Our approved Statement of Work is as follows:

Task 1. Develop a three-group classifier for clustered microcalcifications in mammograms, Months 1-12.

- (a) Collect cases containing 180 malignant and 180 benign clusters of microcalcifications.
- (b) Determine truth state of imaged lesions by reviewing the images, radiologist reports, and pathology reports for these cases.
- (c) Obtain at least 180 FP computer detections from these cases using the existing detection scheme.
- (d) Train and test a three-group classifier on these lesions, using methodology we previously developed for mass lesions.

Task 2. Design and develop an interface for an intelligent workstation for CAD, Months 11-14.

- (a) Examine the most useful features of the interface of the existing intelligent CAD workstation for mammographic lesion detection.
- (b) Examine the most useful features of the interface of the existing CAD schemes in our laboratory for classifying manually detected lesions as malignant or benign.
- (c) Develop a simple interface drawing on the advantages of the existing detection and classification schemes, extended to the three-group classification task.
- (d) Test the interface with non-radiologist observers in our laboratory familiar with the goals of CAD and with interface design principles.

Task 3. Design and perform a pilot observer study measuring radiologists' performances using the three-group classification schemes and traditional two-group classification schemes, Months 15-24.

- (a) Recruit radiologists from our institution and neighboring institutions.

(b) Provide training to the radiologists in the use of the intelligent CAD workstation interfaces.

(c) Measure radiologist performance using the three-group intelligent workstation, and using the existing intelligent workstation for detecting lesions followed by manual selection of lesions to be analyzed by the existing schemes for two-group classification of lesions.

Task 4. Develop techniques to compare radiologists' performance in using the proposed three-group and traditional two-group classification schemes, Months 18-36.

(a) Develop methodology to extend two-group ROC analysis to tasks in which observations are classified into three groups.

(b) Develop methodology to determine the statistical significance of measured differences in performance between three-group classifiers.

(c) Use this methodology to analyze the observer data obtained in Task 3.

For Tasks 1(a) and 1(b), we have collected a database of 134 mammographic cases, four standard views per case; the majority of these cases contain malignant or benign clustered microcalcification lesions. The truth for the malignant microcalcification lesions is verified by pathology report, and that for the benign lesions by pathology report when biopsy was recommended, and by followup when that was recommended by the original radiologist. This is less than the number of malignant and benign lesions initially proposed for this project, but we will have the opportunity to supplement these with further such cases from the database of a colleague in our laboratories.

For Tasks 1(c) and 1(d), we initially encountered difficulties porting the computer code for the existing detection scheme from the legacy equipment for which it was written (IBM RISC 6000 machines, whose operating systems are no longer supported and whose hardware is too old to be considered reliable) to a modern PC workstation running a Linux operating system. These difficulties were traced to compiler incompatibilities between the two systems. A computer programmer in our laboratory with extensive experience with both systems and intimate familiarity with the internals of the detection scheme has investigated and eliminated the majority of these. It is anticipated that completion of Task 1 will require another quarter year of effort.

Our research accomplishments to date have focused largely on Task 4. Although the "methodology we previously developed for mass lesions" [18] was successful for estimating ideal observer decision *variables* based on lesion feature data, a practical classifier to make use of this decision variable data has not yet been implemented. As the difficulties in theoretically characterizing the behavior of such a three-group classifier are intimately related to evaluation of such a classifier's performance (*i. e.*, the development of a three-group extension to ROC analysis), such a reordering of the approved tasks seems logically justified.

We investigated in great detail the behavior of the three-group ideal observer. In particular, it is well-known that the three-group ideal observer makes decisions by partitioning a plane of two decision variables into three regions using three decision boundary lines [3]. We showed that the locations and orientations of these decision boundary lines are not arbitrary; given the slopes and y-intercepts, for example, of two of the lines, those of the third line are constrained to lie within a particular range of values [19]. (See Appendix A.) A detailed understanding of such properties of the three-group ideal observer will prove crucial to the

calculation of observer ROC operating points, and by extension to observer performance evaluation in general.

In our efforts to develop a three-group classifier and appropriate performance evaluation methodology, we have made every attempt to keep our analysis as general as possible despite the theoretical difficulties this entails. Other researchers have proposed three-group methodology by considering observers whose behavior is restricted in particular ways, or by considering only a subset of the possible performance characterization indices (the axes of ROC space), or both [20–24]. The inherent complexity of the three-group classification task makes direct comparison of different methods by different researchers difficult. To facilitate such a comparison, we analyzed the different methods in terms of the three-group ideal observer [25]. (See Appendix B.) In addition to providing us with valuable insight and experience in comparing different classifiers, which should ultimately prove directly relevant to the completion of Task 4, this work also enabled us to present to the observer performance and CAD research communities a useful framework within which comparison of superficially very different classifiers can readily be made. A poster presentation of the theoretical results of this and the preceding paragraph, as well as our research accomplishments during the first year of this award, was made at the 2005 US DOD Breast Cancer Research Program Era of Hope Meeting in Philadelphia, PA [26].

Most recently, we analyzed a simplified performance evaluation method (*i. e.*, an extension of ROC analysis to tasks with three groups) which considers only the three "sensitivities" of the observer — the three probabilities of correctly identifying an observation from one of the three respective groups. (This can, in general, be expected to yield an incomplete description of observer performance, which requires a set of six conditional classification probabilities [7].) This method was originally proposed by Mossman [22] for a pair of essentially *ad hoc* decision rules and arbitrary decision variables, and more recently advocated by He *et al.* [24] for a set of ideal observer decision variables and a decision rule shown [24,25,27] to be a special case of the ideal observer decision rule, and also shown [25,27] to be a special case of the decision rule proposed by Scurfield [21]. We were able to derive a more fundamental motivation for the decision rules described in those works, given the simplified performance description in terms of only the sensitivities, by applying previously successful Neyman-Pearson optimization methodology [3, 7] to this restricted performance evaluation strategy.

Simply put, assuming that one chooses to measure observer performance only in terms of the observer's sensitivities, we proved [28] that the optimal observer with respect to this metric is in fact the special case of the ideal observer proposed by He *et al.* [24]. (See Appendix C.) We then applied this analysis technique [29] to other decision strategies and performance evaluation strategies which we had previously analyzed in terms of the ideal observer decision rule [25]. (See Appendix D.) Given the difficulties inherent in a fully general description of three-class ideal observer behavior and performance evaluation, it is possible that a restricted or simplified model, similar to those proposed already by other researchers, may ultimately prove of greater practical value than the fully general theoretical model. We consider this work important, because it provides a principled theoretical framework in which to evaluate and compare such restricted and simplified models.

A detailed understanding of the properties of the general three-group ideal observer, and of the restricted and simplified models described above, will prove crucial to the calculation of observer ROC operating points, and by extension to observer performance evaluation in general. Since the initiation of funding for this project, the principal investigator and mentor have been holding regular meetings to discuss the theoretical challenges posed by this project

and to explore possible ways of overcoming those challenges.

3 Key Research Accomplishments

- Detailed investigation of the relationships among the decision boundary lines used by the three-group ideal observer (Appendix A)
- Analysis of several proposed three-group classification methods in the literature in terms of the three-group ideal observer (Appendix B)
- Development of principled theoretical motivation for proposed three-group classification methods given selection of restricted or simplified three-group evaluation methodology (Appendices C, D)

4 Reportable Outcomes

- Collection of database of 134 mammographic cases containing malignant and benign clustered microcalcification lesions, with truth determined by pathology (for biopsied lesions) or mammographic followup (benign lesions only)
- Porting of existing computerized scheme for detecting clustered microcalcifications in mammograms from legacy computer systems no longer in operation to workstations currently in use for this project
- D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.*, vol. 24, pp. 1566–1573, 2005.
- 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*, 2005, (accepted for publication 5/25/06).
- D. C. Edwards, C. E. Metz, R. M. Nishikawa, and M. L. Giger, "Investigation of three-group classifiers to fully automate detection and classification of breast lesions in computer-aided diagnosis for mammography," US DOD Breast Cancer Research Program Era of Hope Meeting, Philadelphia, PA, 2005.
- 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., SPIE, Bellingham, WA, 2006, pp. 61 460A1–61 460A7.
- D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.*, 2006, (submitted).

5 Conclusions

During the past year, with the assistance of colleagues in our laboratory, we have collected a database of 134 mammographic cases containing malignant and benign clustered microcalcification lesions, with truth determined by pathology (for biopsied lesions) or mammographic followup (benign lesions only), and we have ported the existing computerized scheme for detecting clustered microcalcifications in mammograms from legacy computer systems no longer in operation to workstations currently in use for this project.

We have continued to advance our theoretical understanding of the three-group ideal observer and methods of evaluating its performance. We showed that the three decision boundary lines used by the three-group ideal observer are not arbitrary, but are intricately related to one another. We analyzed several recently proposed three-group classification methods in terms of the three-group ideal observer. We reported on the important theoretical results we had devloped to date at the 2005 Breast Cancer Research Program Era of Hope Meeting. Finally, we developed principled theoretical motivations for various proposed three-group classification methods, given in each case the selection of a restricted or simplified three-group evaluation methodology.

Although our primary research accomplishments have been theoretical, they are crucial steps in the development of a practical three-group classifier and a fully general three-group performance evaluation methodology. Despite the considerable difficulties involved in such development, a CAD scheme incorporating a three-group classifier as we propose could potentially allow radiologists to detect more malignant breast lesions without increasing their FP biopsy rate. We believe this goal to be worth the necessary effort on our part.

References

[1] J. P. Egan, *Signal Detection Theory and ROC Analysis.* New York: Academic Press, 1975.

[2] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, no. 4, pp. 283–298, 1978.

[3] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I.* New York: John Wiley & Sons, 1968.

[4] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Trans. Med. Imag.*, vol. 20, pp. 886–899, 2001.

[5] C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.*, vol. 43, pp. 1–33, 1999.

[6] C. Ferri, J. Hernández-Orallo, and M. A. Salido, "Volume under the roc surface for multi-class problems: Exact computation and evaluation of approximations," Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València (Spain), Tech. Rep., 2003.

[7] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.*, vol. 23, pp. 891–895, 2004.

[8] U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.*, vol. 2, pp. 1–9, 1995.

[9] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Med. Phys.*, vol. 18, pp. 955–963, 1991.

[10] F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.*, vol. 28, pp. 473–481, 1993.

[11] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.*, vol. 21, pp. 445–452, 1994.

[12] M. A. Kupinski, "Computerized pattern classification in medical imaging," Ph.D. Thesis, The University of Chicago, Chicago, IL, 2000.

[13] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[14] Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.*, vol. 44, pp. 2579–2595, 1999.

[15] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness," *Acad. Radiol.*, vol. 7, pp. 1077–1084, 2000.

[16] Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.*, vol. 20, pp. 1285–1292, 2001.

[17] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis — Observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560–568, 2002.

[18] 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.*, vol. 31, pp. 81–90, 2004.

[19] D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.*, vol. 24, pp. 1566–1573, 2005.

[20] B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math. Psychol.*, vol. 40, pp. 253–269, 1996.

[21] ——, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math. Psychol.*, vol. 42, pp. 5–31, 1998.

[22] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, pp. 78–89, 1999.

[23] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, Eds., SPIE, Bellingham, WA, 2003, pp. 567–578.

[24] X. He, C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis — A decision theoretic approach under the ideal observer framework," *IEEE Trans. Med. Imag.*, vol. 25, pp. 571–581, 2006.

[25] 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*, 2005, (accepted for publication 5/25/06).

[26] D. C. Edwards, C. E. Metz, R. M. Nishikawa, and M. L. Giger, "Investigation of three-group classifiers to fully automate detection and classification of breast lesions in computer-aided diagnosis for mammography," US DOD Breast Cancer Research Program Era of Hope Meeting, Philadelphia, PA, 2005.

[27] D. C. Edwards and C. E. Metz, "Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, Eds., SPIE, Bellingham, WA, 2005, pp. 128–137.

[28] ——, "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., SPIE, Bellingham, WA, 2006, pp. 61 460A1–61 460A7.

[29] ——, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.*, 2006, (submitted).

A Restrictions on the Three-Class Ideal Observer's Decision Boundary Lines

Restrictions on the Three-Class Ideal Observer's Decision Boundary Lines

Darrin C. Edwards* and Charles E. Metz

***Abstract*—We are attempting to develop expressions for the coordinates of points on the three-class ideal observer's receiver operating characteristic (ROC) hypersurface as functions of the set of decision criteria used by the ideal observer. This is considerably more difficult than in the two-class classification task, because the conditional probabilities in question are not simply related to the cumulative distribution functions of the decision variables, and because the slopes and intercepts of the decision boundary lines are not independent; given the locations of two of the lines, the location of the third will be constrained depending on the other two. In this paper, we attempt to characterize those constraining relationships among the three-class ideal observer's decision boundary lines. As a result, we show that the relationship between the decision criteria and the misclassification probabilities is not one-to-one, as it is for the two-class ideal observer.**

***Index Terms*—Ideal observers, ROC analysis, three-class classification.**

I. Introduction

RECEIVER operating characteristic (ROC) analysis is the accepted methodology for analyzing the performance of a two-class classifier [1], in particular for medical decision-making tasks in which a patient is diagnosed as having or not having a particular condition based on features of a medical image [2]. In judging the performance of an observer measured via ROC analysis, the standard for comparison is the so-called ideal observer, that observer which outperforms any other possible observer given the statistical variability of the observational data being classified [1], [3]. Although the general form of the ideal observer in a classification task with three or more classes has been known for some time [3], the considerable complexities inherent to this model compared to the two-class classification task have hampered the development of extensions of ROC analysis which are both fully general and practically useful. (Several researchers have recently proposed restricted observer models or restricted evaluation methods [4]–[7].)

Despite these difficulties, research continues in this area because the advantages to be gained from a three-class classifier and appropriate evaluation methodology are considerable. In our own case, we seek to combine existing computer-aided diagnosis (CAD) schemes for detecting [8]–[12] mammographic mass lesions and classifying [13]–[17] them as malignant or benign. The combined scheme would serve as a fully automated classifier (the existing classifier requires initial manual identification of lesions by a radiologist), potentially allowing radiologists to reduce their false-positive biopsy rate without reducing their sensitivity for detection of malignancies. Simply concatenating the two types of scheme in a two-stage classifier would be inadequate, because the output of the detection scheme will necessarily include false-positive (FP) computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our initial efforts toward the goal of developing a true three-class classifier have been more theoretical than practical so far. We have shown that, just as the two-class ideal observer achieves the optimal two-class ROC curve for a given task, the N-class ideal observer achieves the optimal N-class ROC hypersurface [18]. (Note that the ideal observer is formally defined as that which minimizes the expected Bayes risk [3], and not in terms of classification performance, making this a nontrivial observation in both cases.) More soberingly, we found recently that an obvious generalization of the well-known performance metric, the area under the ROC curve (AUC), is not a useful performance metric in a classification task with three or more classes [19].

At present we are attempting to develop expressions for the coordinates of points on the three-class ideal observer's ROC hypersurface (the conditional probabilities for misclassifying observations [18], [20], [21]) as functions of the set of decision criteria used by the ideal observer. This is considerably more difficult than in the two-class classification task for two reasons. First, the conditional probabilities in question are not simply related to the cumulative distribution functions (cdfs) of the decision variables, but are integrals of those variables over domains determined by three decision boundary lines [3]. Second, the slopes and intercepts of the decision boundary lines are not independent; given the locations of two of the lines, we have found recently that the location of the third will be constrained depending on the other two.

In this paper, we attempt to characterize the constraining relationships just mentioned among the three-class ideal observer's

Manuscript received May 17, 2005; revised August 25, 2005. This work was supported by the U.S. Army Medical Research and Materiel Command under Grant W81XWH-04-1-0495 (D. C. Edwards, principal investigator). The Associate Editor responsible for coordinating the review of this paper and recommending its publication was E. Krupinski. *Asterisk indicates corresponding author.*

*D. C. Edwards is with the Department of Radiology, the University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637 USA (e-mail: d-edwards@uchicago.edu).

C. E. Metz is with the Department of Radiology, the University of Chicago, Chicago, IL 60637 USA.

Digital Object Identifier 10.1109/TMI.2005.859212

decision boundary lines. Although this paper is admittedly still removed from image analysis perse, we hope it may prove of interest to the CAD community and ultimately of relevance to a wide variety of medical image analysis tasks. In the next section we briefly review the structure of the three-class ideal observer and the notation we have been using to characterize it [18]. In Section III, we show that for a given location (slope and y-intercept) of the decision boundary line separating the first and third classes, the location of one of the remaining two lines is constrained in a particular way based on the location of the other.

These results are discussed in Section IV. Given the arbitrariness of the labels applied to the three classes (ie, which classes are considered first, second, or third), one would expect the selection of the fixed line in Section III to be similarly arbitrary, and indeed in Appendices A and B we show that corresponding and consistent results are obtained if one takes the location of the decision boundary line separating the second and third, or first and second, classes, respectively, to be given.

II. The Three-Class Ideal Observer

In [18], we showed that an N-class ideal observer makes decisions by partitioning a likelihood ratio decision variable space, where the boundaries of the partitions are given by hyperplanes

$$\begin{aligned}\text{decide}\quad d=\pi_i \text{ iff } &\sum_{k=1}^{N-1}(U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k)\mathrm{LR}_k \\ &\geq (U_{j|N}-U_{i|N})P(\mathbf{t}=\pi_N)\quad \{j<i\} \qquad (1)\\ \text{and } &\sum_{k=1}^{N-1}(U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k)\mathrm{LR}_k \\ &> (U_{j|N}-U_{i|N})P(\mathbf{t}=\pi_N)\quad \{j>i\}. \qquad (2)\end{aligned}$$

Here, $U_{i|j}$ is the utility of deciding an observation is from class π_i given that it is actually from class π_j; $P(\mathbf{t}=\pi_k)$ is the apriori probability that an observation is drawn from class π_k; and LR_k is the kth likelihood ratio, defined by the ratio $p(\vec{x}|\pi_k)/p(\vec{x}|\pi_N)$ of the probability density functions of the observational data (We use boldface type to denote random variables). The partitioning is determined by the parameters

$$\gamma_{ijk}\equiv (U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k) \qquad (3)$$

with i, j, and k varying from 1 to N, and $j\neq i$. Note that these parameters are not independent, however, because

$$\gamma_{ijk}=\gamma_{kjk}-\gamma_{kik}. \qquad (4)$$

We can impose the reasonable condition that the utility for correctly classifying an observation from a given class should be greater than any utility for incorrectly classifying an observation from the same class, i.e., $U_{i|i}>U_{j|i}\ \{i\neq j\}$. This gives, for $j\neq i$,

$$\gamma_{iji}>0 \qquad (5)$$

leaving $N(N-1)$ positive parameters (the rest are derivable from (4)).

Finally, note that the hyperplanes represented by (1) and (2) are unchanged if we multiply all of these equations by a single scalar, such as $1/(\Sigma_{i\neq j}\gamma_{iji})$. This leaves us with N^2-N-1 degrees of freedom, as expected.

The behavior of a three-class ideal observer is completely determined by the three decision boundary lines

$$\gamma_{121}\mathrm{LR}_1-\gamma_{212}\mathrm{LR}_2=\gamma_{313}-\gamma_{323} \qquad (6)$$

$$\gamma_{131}\mathrm{LR}_1+(\gamma_{232}-\gamma_{212})\mathrm{LR}_2=\gamma_{313} \qquad (7)$$

$$(\gamma_{131}-\gamma_{121})\mathrm{LR}_1+\gamma_{232}\mathrm{LR}_2=\gamma_{323} \qquad (8)$$

which we call, respectively, the "1-vs-2" line, the "1-vs-3" line, and the "2-vs-3" line. Note that if any two of these lines intersect, the third line must also share this intersection point. We also emphasize the simple interpretation, from (3), of each of the γ_{iji} parameters appearing in these decision boundary line equations as the difference in utilities between a "correct" and one particular "incorrect" decision (scaled by the apriori probability of the true class in question); and of each difference in the γ_{iji} parameters as a difference in utilities between two possible "incorrect" decisions [again scaled by the apriori probability of the true class in question; e.g., $\gamma_{313}-\gamma_{323}=(U_{2|3}-U_{1|3})P(\mathbf{t}=\pi_3)$].

From the conditions on the γ_{iji} parameters in (5), we can readily derive conditions on the decision boundaries themselves. If we denote the slope of the "i-vs-j" line by m_{ij}, its y-intercept by b_{ij}, and its x-intercept by χ_{ij}, we have

$$m_{12}=\frac{\gamma_{121}}{\gamma_{212}}>0 \qquad (9)$$

$$\chi_{13}=\frac{\gamma_{313}}{\gamma_{131}}>0 \qquad (10)$$

$$b_{23}=\frac{\gamma_{323}}{\gamma_{232}}>0. \qquad (11)$$

These are the three conditions stated in [22].

III. Restrictions Determined by the Parameters of the "1-vs.-3" Line

Constraints on the decision boundaries, in addition to those given in (9)–(11), can be obtained by considering the two cases $\gamma_{232}-\gamma_{212}>0$ and $\gamma_{232}-\gamma_{212}<0$. In the first case (ie, $\gamma_{232}>\gamma_{212}$, or $U_{1|2}>U_{3|2}$), we have

$$m_{13}=\frac{-\gamma_{131}}{\gamma_{232}-\gamma_{212}}<0 \qquad (12)$$

$$b_{13}=\frac{\gamma_{313}}{\gamma_{232}-\gamma_{212}}>0. \qquad (13)$$

We also have

$$\begin{aligned}m_{23}&=\frac{-(\gamma_{131}-\gamma_{121})}{\gamma_{232}}\\ &=\frac{(\gamma_{232}-\gamma_{212})m_{13}+\gamma_{212}m_{12}}{\gamma_{232}}\\ &=\left(1-\frac{\gamma_{212}}{\gamma_{232}}\right)m_{13}+\frac{\gamma_{212}}{\gamma_{232}}m_{12}. \qquad (14)\end{aligned}$$

This is a weighted sum of the slopes m_{12} and m_{13}, where the weights are positive and sum to one. Since we must have $m_{13}<m_{12}$ from (9) and (12), it must therefore be the case that

$$m_{13}\leq m_{23}\leq m_{12}. \qquad (15)$$

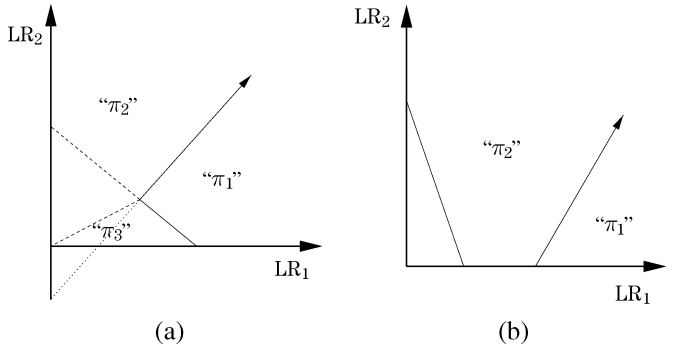

Fig. 1. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} > 0$ (implying $m_{13} < 0$ and $b_{13} > 0$) and $b_{12} < 0$. In (a), $\chi_{12} < \chi_{13}$, and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $b_{23} > 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $\chi_{12} \geq \chi_{13}$ and the "2-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "1-vs-3" lines shown); observations above this line will be decided "π_2," and those below this line will be decided "π_3."

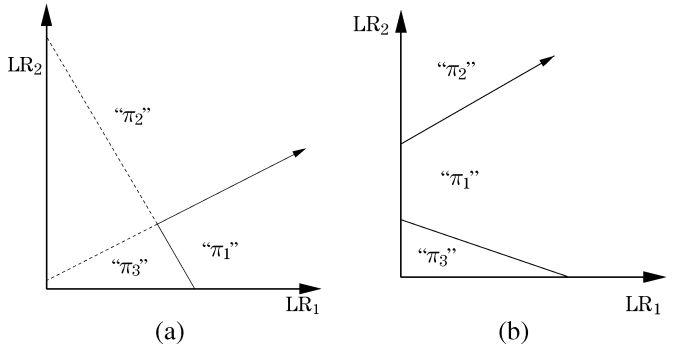

Fig. 2. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} > 0$ (implying $m_{13} < 0$ and $b_{13} > 0$) and $b_{12} \geq 0$. In (a), $b_{12} < b_{13}$, and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $b_{12} \geq b_{13}$ and the "2-vs-3" line can lie anywhere between the "1-vs-2" and "1-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_1" regardless of the position of this line.

Furthermore

$$\begin{aligned} b_{23} &= \frac{\gamma_{323}}{\gamma_{232}} \\ &= \frac{\gamma_{313} - (\gamma_{313} - \gamma_{323})}{\gamma_{232}} \\ &= \frac{(\gamma_{232} - \gamma_{212})b_{13} + \gamma_{212}b_{12}}{\gamma_{232}} \\ &= \left(1 - \frac{\gamma_{212}}{\gamma_{232}}\right) b_{13} + \frac{\gamma_{212}}{\gamma_{232}} b_{12}. \end{aligned} \tag{16}$$

This is a weighted sum of the y-intercepts b_{12} and b_{13}, where the weights are positive and sum to one; thus, in addition to (15), we have the condition

$$\min(b_{12}, b_{13}) \leq b_{23} \leq \max(b_{12}, b_{13}). \tag{17}$$

If $b_{12} < 0$, then (17) immediately reduces to $b_{12} \leq b_{23} \leq b_{13}$ (by (13), we are considering a special case in which $b_{13} > 0$). This is illustrated in Fig. 1 for the slightly different situations $\chi_{12} < \chi_{13}$ and $\chi_{12} \geq \chi_{13}$. If, on the other hand, $b_{12} \geq 0$, then (15) and (17) together imply two possible situations, depending on whether $b_{12} < b_{13}$ or $b_{12} \geq b_{13}$. These possibilities are illustrated in Fig. 2.

We now consider the case $\gamma_{232} - \gamma_{212} < 0$ (ie, $\gamma_{232} < \gamma_{212}$, or $U_{1|2} < U_{3|2}$), which yields

$$m_{13} = \frac{-\gamma_{131}}{\gamma_{232} - \gamma_{212}} > 0 \tag{18}$$

$$b_{13} = \frac{\gamma_{313}}{\gamma_{232} - \gamma_{212}} < 0. \tag{19}$$

We now have

$$\begin{aligned} m_{12} &= \frac{\gamma_{121}}{\gamma_{212}} \\ &= \frac{\gamma_{131} - (\gamma_{131} - \gamma_{121})}{\gamma_{212}} \\ &= \frac{-(\gamma_{232} - \gamma_{212})m_{13} + \gamma_{232}m_{23}}{\gamma_{212}} \\ &= \left(1 - \frac{\gamma_{232}}{\gamma_{212}}\right) m_{13} + \frac{\gamma_{232}}{\gamma_{212}} m_{23}. \end{aligned} \tag{20}$$

This is again a weighted sum in which the weights are positive and sum to one, giving

$$\min(m_{13}, m_{23}) \leq m_{12} \leq \max(m_{13}, m_{23}). \tag{21}$$

Furthermore

$$\begin{aligned} b_{12} &= \frac{\gamma_{313} - \gamma_{323}}{-\gamma_{212}} \\ &= \frac{-\gamma_{313} + \gamma_{323}}{\gamma_{212}} \\ &= \frac{-(\gamma_{232} - \gamma_{212})b_{13} + \gamma_{232}b_{23}}{\gamma_{212}} \\ &= \left(1 - \frac{\gamma_{232}}{\gamma_{212}}\right) b_{13} + \frac{\gamma_{232}}{\gamma_{212}} b_{23}. \end{aligned} \tag{22}$$

This is a weighted sum of the y-intercepts b_{13} and b_{23}, where the weights are positive and sum to one; thus, in addition to (21), we have the condition

$$b_{13} \leq b_{12} \leq b_{23} \tag{23}$$

since $b_{13} < b_{23}$ by (11) and (19).

If $m_{23} < 0$, then (21) immediately reduces to $m_{23} \leq m_{12} \leq m_{13}$ (by (18), we are considering a special case in which $m_{13} > 0$). This is illustrated in Fig. 3 for the slightly different situations $\chi_{13} < \chi_{23}$ and $\chi_{13} \geq \chi_{23}$. If, on the other hand, $m_{23} \geq 0$, then (21) and (23) together imply two possible situations, depending on whether $m_{23} < m_{13}$ or $m_{23} \geq m_{13}$. These possibilities are illustrated in Fig. 4.

One may of course ask what happens when $\gamma_{232} - \gamma_{212} = 0$ (ie, $\gamma_{232} = \gamma_{212}$, or $U_{1|2} = U_{3|2}$). In this case, both m_{13} and b_{13} are infinite. Furthermore

$$\begin{aligned} m_{23} &= \frac{-(\gamma_{131} - \gamma_{121})}{\gamma_{232}} \\ &= \frac{-\gamma_{131}}{\gamma_{232}} + \frac{\gamma_{121}}{\gamma_{212}} \\ &= \frac{-\gamma_{131}}{\gamma_{232}} + m_{12} \\ &\leq m_{12} \end{aligned} \tag{24}$$

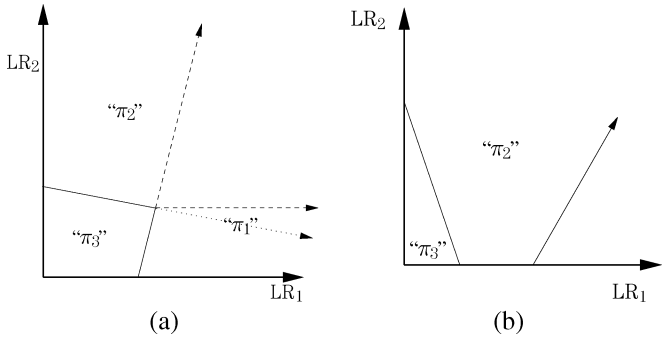

Fig. 3. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} < 0$ (implying $m_{13} > 0$ and $b_{13} < 0$) and $m_{23} < 0$. In (a), $\chi_{13} < \chi_{23}$, and the "1-vs-2" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $m_{12} > 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_1." In (b), $\chi_{13} \geq \chi_{23}$ and the "1-vs-2" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-3" and "2-vs-3" lines shown); observations above this line will be decided "π_2", and those below this line will be decided "π_1."

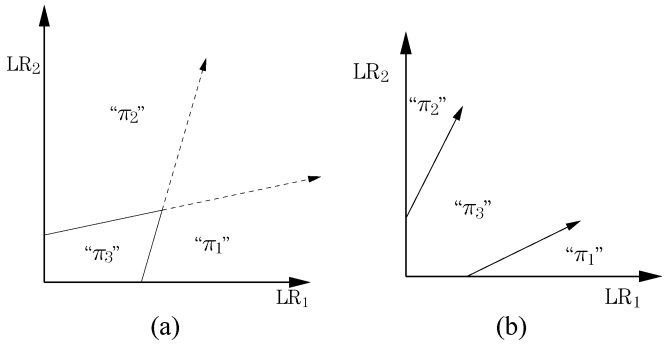

Fig. 4. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} < 0$ (implying $m_{13} > 0$ and $b_{13} < 0$) and $m_{23} \geq 0$. In (a), $m_{23} < m_{13}$, and the "1-vs-2" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2", and those below this line will be decided "π_1". In (b), $m_{23} \geq m_{13}$, and the "1-vs-2" line can lie anywhere between the "1-vs-3" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_3" regardless of the position of this line.

and

$$\begin{aligned} b_{12} &= \frac{\gamma_{323} - \gamma_{313}}{\gamma_{212}} \\ &= \frac{\gamma_{323}}{\gamma_{232}} + \frac{-\gamma_{313}}{\gamma_{212}} \\ &= b_{23} + \frac{-\gamma_{313}}{\gamma_{212}} \\ &\leq b_{23}. \end{aligned} \tag{25}$$

Together, (24) and (25) can be considered *either* a special case of the inequalities (15) and (17), if we take $m_{13} = -\infty$ and $b_{13} = +\infty$; *or* of the inequalities (21) and (23), if we take $m_{13} = +\infty$ and $b_{13} = -\infty$. This situation, for the slightly different cases $b_{12} < 0$ and $b_{12} \geq 0$, is illustrated in Fig. 5.

In this section, the possible values of the quantity $\gamma_{232} - \gamma_{212}$ were considered in order to determine properties of the ideal observer decision boundary lines. It may be argued that the choice of a parameter from the "1-vs-3" line, i.e., one of the three available lines, must be an arbitrary one. In fact, we may consider taking another parameter (or combination of parameters) from (6)–(8), and using it to determine conditions on the properties

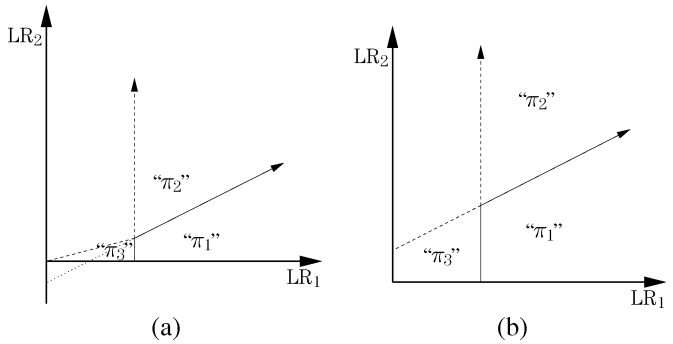

Fig. 5. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} = 0$ (implying $m_{13} = \mp\infty$ and $b_{13} = \pm\infty$). In (a), $b_{12} < 0$ and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $b_{23} > 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $b_{12} \geq 0$ and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3."

of the decision boundary lines as above. Given that all possible values of the quantity $\gamma_{232} - \gamma_{212}$ were considered, it is expected that no new conditions should be determinable (let alone conditions inconsistent with those already determined). In fact, this can readily be shown to be the case; however, due to the repetitive nature of the derivations involved, these are relegated to Appendices A and B.

IV. Discussion and Conclusion

The repetitive nature of the algebraic manipulations given in the preceding section and the Appendices should not be allowed to distract from the fundamental point being made: given the locations of two of the decision boundary lines, the location of the third is not completely arbitrary. That is, aside from the obvious [given (6)–(8)] constraint that the lines must share a common intersection point, it can also be shown that the slope of the third line is constrained by the slopes of the first two.

The significance of this result may be difficult to appreciate at first glance. It is perhaps best illustrated by comparison with the two-class classifier, for which the ROC operating point coordinates [e.g., the true-positive fraction (TPF) and false-positive fraction (FPF)] are determined by a single decision criterion γ, which is free to vary without restriction throughout its domain of definition. For the two-class ideal observer, in particular, an observation is decided "positive" (assigned to the class π_1) if $\mathrm{LR}_1 > \gamma$, where γ can take on any nonnegative value. Furthermore, the FPF and TPF are related in a very simple way to the cdfs of $\mathbf{LR}_1$, and are thus monotonic in the decision criterion γ. For the three-class ideal observer, this straightforward relationship is lost; indeed, Figs. 2(b), 4(b), 7(b), 9(b), 12(b), and 14(b) show that for certain values of four of the five decision criteria γ_{iji}, the misclassification probabilities (ie, the ROC operating point coordinates) can be independent of the fifth decision criterion.

More succinctly, the relationship between the decision criteria and the misclassification probabilities is *not* one-to-one, as it is for the two-class ideal observer. A correct formulation of the misclassification probabilities as functions of the decision criteria—necessary for an explicit calculation of the ideal

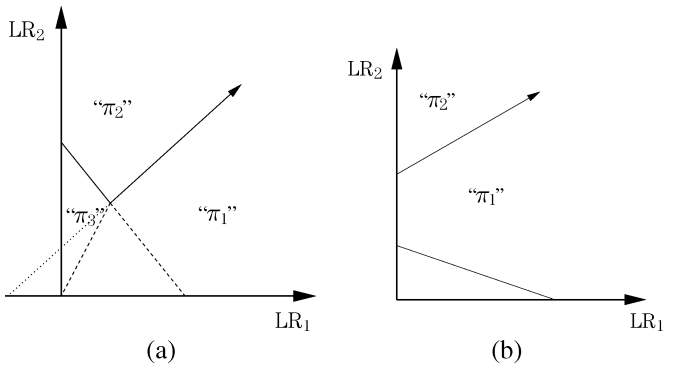

Fig. 6. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} > 0$ (implying $1/m_{23} < 0$ and $\chi_{23} > 0$) and $\chi_{12} < 0$. In (a), $b_{12} < b_{23}$, and the "1-vs-3" line can lie anywhere between the two dashed lines shown (the region between the left dashed and dotted lines is excluded because $\chi_{13} > 0$); observations in the unlabeled region to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3." In (b), $b_{12} \geq b_{23}$ and the "1-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "2-vs-3" lines shown); observations to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3."

observer's ROC hypersurface given the decision variable probability density functions—will require careful consideration of this issue. Although we have shown previously that the hypervolume under the ROC hypersurface is not a useful performance metric in general [19], it is still the case that the ROC hypersurface in terms of the set of misclassification probabilities (six in the three-class classification task) is a complete description of observer performance. We expect that a useful performance metric, assuming one exists, will be derived in some fashion from the ROC hypersurface. It is thus important to develop a complete understanding of the rather complicated relationships among the quantities involved, and we hope that this paper will prove of some use toward this goal.

Appendix A
Restrictions Determined by the Parameters of the "2-vs.-3" Line

Consider the quantity $\gamma_{131} - \gamma_{121}$ from (8). In particular, when $\gamma_{131} - \gamma_{121} > 0$ (ie, $\gamma_{131} > \gamma_{121}$, or $U_{2|1} > U_{3|1}$), we have

$$\frac{1}{m_{23}} = \frac{-\gamma_{232}}{\gamma_{131} - \gamma_{121}} < 0 \tag{26}$$

$$\chi_{23} = \frac{\gamma_{323}}{\gamma_{131} - \gamma_{121}} > 0. \tag{27}$$

Through reasoning similar to that of Section III, we also have

$$\frac{1}{m_{23}} \leq \frac{1}{m_{13}} \leq \frac{1}{m_{12}} \tag{28}$$

and

$$\min(\chi_{12}, \chi_{23}) \leq \chi_{13} \leq \max(\chi_{12}, \chi_{23}). \tag{29}$$

If $\chi_{12} < 0$, then (29) immediately reduces to $\chi_{12} \leq \chi_{13} \leq \chi_{23}$ (by (27), we are considering a special case in which $\chi_{23} > 0$). This is illustrated in Fig. 6 for the slightly different situations $b_{12} < b_{23}$ and $b_{12} \geq b_{23}$. If, on the other hand, $\chi_{12} \geq 0$, then (28) and (29) together imply two possible situations, depending on whether $\chi_{12} < \chi_{23}$ or $\chi_{12} \geq \chi_{23}$. These possibilities are illustrated in Fig. 7.

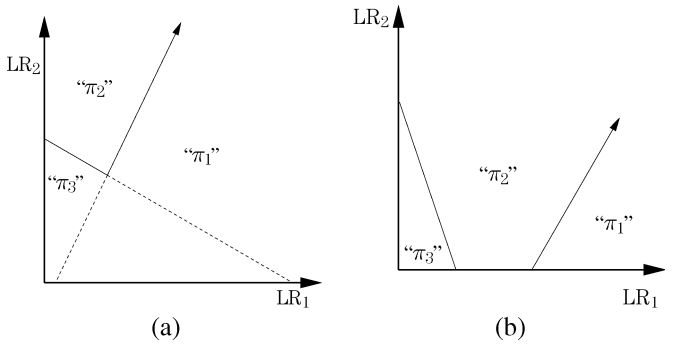

Fig. 7. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} > 0$ (implying $1/m_{23} < 0$ and $\chi_{23} > 0$) and $\chi_{12} \geq 0$. In (a), $\chi_{12} < \chi_{23}$, and the "1-vs-3" line can lie anywhere in the unlabeled region; observations to the left of this line will be decided "π_1," and those to the right of this line will be decided "π_3." In (b), $\chi_{12} \geq \chi_{23}$ and the "1-vs-3" line can lie anywhere between the "1-vs-2" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_2" regardless of the position of this line.

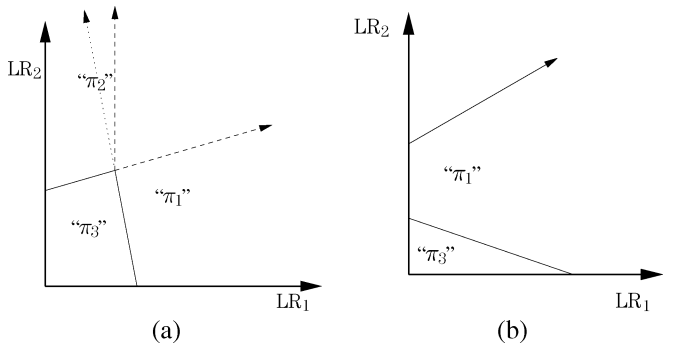

Fig. 8. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} < 0$ (implying $1/m_{23} > 0$ and $\chi_{23} < 0$) and $1/m_{13} < 0$. In (a), $b_{23} < b_{13}$, and the "1-vs-2" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $m_{12} > 0$ and, therefore, $1/m_{12} \geq 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_1." In (b), $b_{23} \geq b_{13}$ and the "1-vs-2" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-3" and "2-vs-3" lines shown); observations above this line will be decided "π_2", and those below this line will be decided "π_1."

If $\gamma_{131} - \gamma_{121} < 0$ (ie, $\gamma_{131} < \gamma_{121}$, or $U_{2|1} < U_{3|1}$), we have

$$\frac{1}{m_{23}} = \frac{-\gamma_{232}}{\gamma_{131} - \gamma_{121}} > 0 \tag{30}$$

$$\chi_{23} = \frac{\gamma_{323}}{\gamma_{131} - \gamma_{121}} < 0. \tag{31}$$

One can also show

$$\min\left(\frac{1}{m_{13}}, \frac{1}{m_{23}}\right) \leq \frac{1}{m_{12}} \leq \max\left(\frac{1}{m_{13}}, \frac{1}{m_{23}}\right) \tag{32}$$

and

$$\chi_{23} \leq \chi_{12} \leq \chi_{13}. \tag{33}$$

If $1/m_{13} < 0$, then (32) immediately reduces to $1/m_{13} \leq 1/m_{12} \leq 1/m_{23}$ (by (30), we are considering a special case in which $1/m_{23} > 0$). This is illustrated in Fig. 8 for the slightly different situations $b_{23} < b_{13}$ and $b_{23} \geq b_{13}$. If, on the other

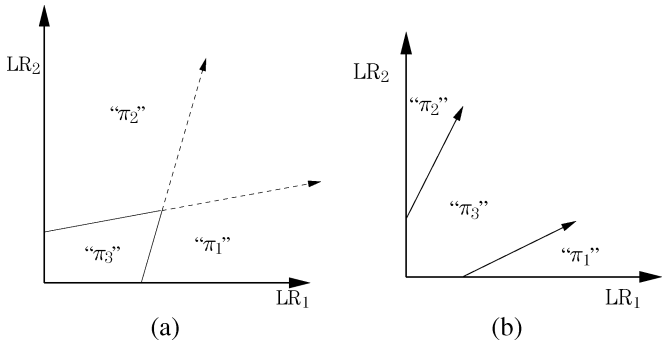

Fig. 9. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} < 0$ (implying $1/m_{23} > 0$ and $\chi_{23} < 0$) and $1/m_{13} \geq 0$. In (a), $1/m_{13} < 1/m_{23}$, and the "1-vs-2" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_1." In (b), $1/m_{13} \geq 1/m_{23}$ and the "1-vs-2" line can lie anywhere between the "1-vs-3" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_3" regardless of the position of this line.

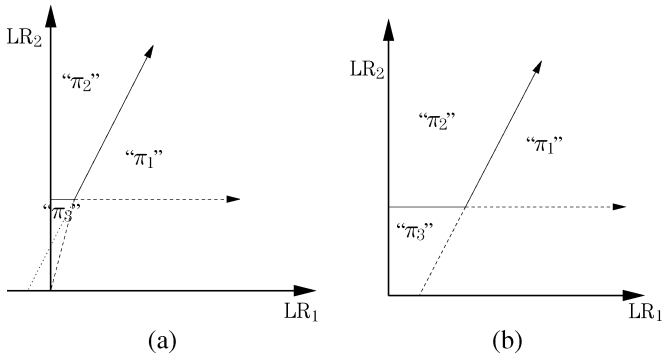

Fig. 10. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} = 0$ (implying $1/m_{23} = \mp\infty$ and $\chi_{23} = \pm\infty$). In (a), $\chi_{12} < 0$, and the "1-vs-3" line can lie anywhere between the two dashed lines shown (the region between the leftmost dashed and dotted lines is excluded because $\chi_{13} > 0$); observations in the unlabeled region to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3." In (b), $\chi_{12} \geq 0$ and the "1-vs-3" line can lie anywhere in the unlabeled region; observations to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3."

hand, $1/m_{13} \geq 0$, then (32) and (33) together imply two possible situations, depending on whether $1/m_{13} < 1/m_{23}$ or $1/m_{13} \geq 1/m_{23}$. These possibilities are illustrated in Fig. 9.

Finally, we consider the case $\gamma_{131} - \gamma_{121} = 0$ ($\gamma_{131} = \gamma_{121}$ or $U_{2|1} = U_{3|1}$), in which both $1/m_{23}$ and χ_{23} are infinite. We now have

$$\frac{1}{m_{13}} \leq \frac{1}{m_{12}} \tag{34}$$

and

$$\chi_{12} \leq \chi_{13}. \tag{35}$$

Together, (34) and (35) can be considered *either* a special case of the inequalities (28) and (29), if we take $1/m_{23} = -\infty$ and $\chi_{23} = +\infty$; *or* of the inequalities (32) and (33), if we take $1/m_{23} = +\infty$ and $\chi_{23} = -\infty$. This situation, for the slightly different cases $\chi_{12} < 0$ and $\chi_{12} \geq 0$, is illustrated in Fig. 10.

Notice that every figure in this appendix has one or more corresponding figures in Section III (depending on the possible values of the undetermined decision boundary parameter being illustrated in that figure). Specifically

$$\begin{aligned}
\text{Fig. } 6(\text{a}) &\Rightarrow \text{Figs. } 2(\text{a}),\ 3(\text{a}),\ 5(\text{b}) \\
\text{Fig. } 6(\text{b}) &\Rightarrow \text{Fig. } 2(\text{b}) \\
\text{Fig. } 7(\text{a}) &\Rightarrow \text{Figs. } 1(\text{a}),\ 3(\text{a}),\ 5(\text{a}) \\
\text{Fig. } 7(\text{b}) &\Rightarrow \text{Figs. } 1(\text{b}),\ 3(\text{b}),\ 5(\text{a}) \\
\text{Fig. } 8(\text{a}) &\Rightarrow \text{Figs. } 1(\text{a}),\ 2(\text{a}) \\
\text{Fig. } 8(\text{b}) &\Rightarrow \text{Fig. } 2(\text{b}) \\
\text{Fig. } 9(\text{a}) &\Rightarrow \text{Figs. } 4(\text{a}),\ 5(\text{a}),\ 5(\text{b}) \\
\text{Fig. } 9(\text{b}) &\Rightarrow \text{Fig. } 4(\text{b}) \\
\text{Fig. } 10(\text{a}) &\Rightarrow \text{Figs. } 2(\text{a}),\ 4(\text{a}),\ 5(\text{b}),\ 2(\text{b}) \\
\text{Fig. } 10(\text{b}) &\Rightarrow \text{Figs. } 1(\text{a}),\ 4(\text{a}),\ 5(\text{a}).
\end{aligned}$$

That is, none of the conditions derived in this section are inconsistent with those derived Section III. More importantly, note the symmetry between the corresponding equations and figures in Section III and this appendix, if one "swaps" the labels of classes π_1 and π_2, and additionally replaces m_{ij} with $1/m_{i'j'}$, χ_{ij} with $b_{i'j'}$, and b_{ij} with $\chi_{i'j'}$ ($i' = 1$ if $i = 2$, 2 if $i = 1$, and 3 if $i = 3$; similarly for j). Intuitively, if one "flips" the figures in one section about the $y = x$ line, one obtains the figures in the other section.

Appendix B
Restrictions Determined by the Parameters of the "1-vs.-2" Line

In this appendix, we consider the possible values of the quantity $\gamma_{313} - \gamma_{323}$. As in the preceding Appendix, we expect to obtain no conditions inconsistent with those already derived.

When $\gamma_{313} - \gamma_{323} > 0$ (ie, $\gamma_{313} > \gamma_{323}$, or $U_{2|3} > U_{1|3}$), we have

$$\frac{1}{b_{12}} = \frac{-\gamma_{212}}{\gamma_{313} - \gamma_{323}} < 0 \tag{36}$$

$$\frac{1}{\chi_{12}} = \frac{\gamma_{121}}{\gamma_{313} - \gamma_{323}} > 0. \tag{37}$$

Through reasoning similar to that of Section III, we also have

$$\frac{1}{b_{12}} \leq \frac{1}{b_{13}} \leq \frac{1}{b_{23}} \tag{38}$$

and

$$\min\left(\frac{1}{\chi_{23}}, \frac{1}{\chi_{12}}\right) \leq \frac{1}{\chi_{13}} \leq \max\left(\frac{1}{\chi_{23}}, \frac{1}{\chi_{12}}\right). \tag{39}$$

If $1/\chi_{23} \leq 0$, then (39) immediately reduces to $1/\chi_{23} \leq 1/\chi_{13} \leq 1/\chi_{12}$ (by (37), we are considering a special case in which $1/\chi_{12} > 0$). This is illustrated in Fig. 11 for the slightly different situations $m_{23} < m_{12}$ and $m_{23} \geq m_{12}$. If, on the other hand, $1/\chi_{23} > 0$, then (38) and (39) together imply two possible situations, depending on whether $1/\chi_{23} < 1/\chi_{12}$ or $1/\chi_{23} \geq 1/\chi_{12}$. These possibilities are illustrated in Fig. 12.

If $\gamma_{313} - \gamma_{323} < 0$ (ie, $\gamma_{313} < \gamma_{323}$, or $U_{2|3} < U_{1|3}$), we have

$$\frac{1}{b_{12}} = \frac{-\gamma_{212}}{\gamma_{313} - \gamma_{323}} > 0 \tag{40}$$

$$\frac{1}{\chi_{12}} = \frac{\gamma_{121}}{\gamma_{313} - \gamma_{323}} < 0. \tag{41}$$

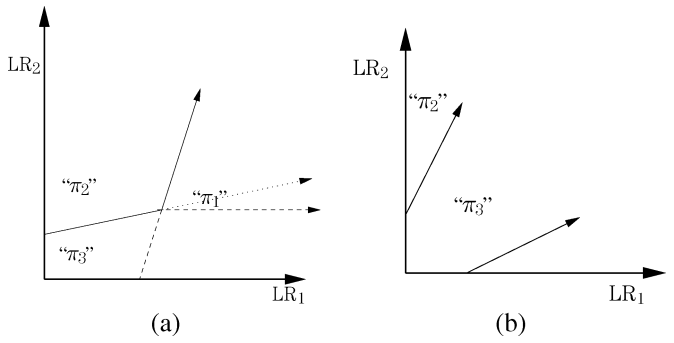

Fig. 11. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} > 0$ (implying $1/b_{12} < 0$ and $1/\chi_{12} > 0$) and $1/\chi_{23} \leq 0$. In (a), $m_{23} < m_{12}$, and the "1-vs-3" line can lie anywhere between the two dashed lines shown (the region between the horizontal dashed and dotted lines is excluded because $\chi_{13} > 0$ and, therefore, $1/\chi_{13} \geq 0$); observations in the unlabeled region to the left of this line will be decided "π_3", and those to the right of line will be decided "π_1." In (b), $m_{23} \geq m_{12}$, and the "1-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "2-vs-3" lines shown); observations to the left of this line will be decided "π_3," and those to the right of this line will be decided "π_1."

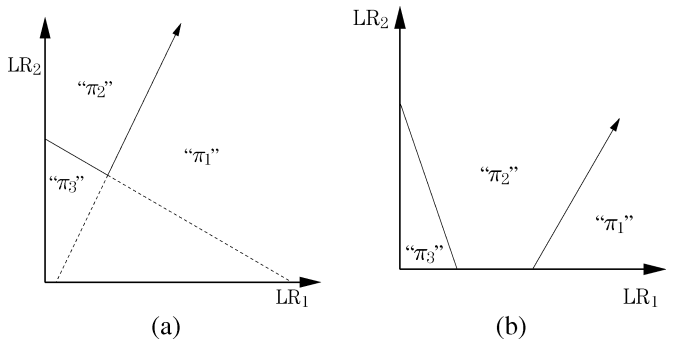

Fig. 12. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} > 0$ (implying $1/b_{12} < 0$ and $1/\chi_{12} > 0$) and $1/\chi_{23} > 0$. In (a), $1/\chi_{23} < 1/\chi_{12}$ and the "1-vs-3" line can lie anywhere in the unlabeled region; observations to the left of this line will be decided "π_3," and those to the right of this line will be decided "π_1." In (b), $1/\chi_{23} \geq 1/\chi_{12}$, and the "1-vs-3" line can lie anywhere between the "1-vs-2" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_2" regardless of the position of this line.

One can also show

$$\min\left(\frac{1}{b_{13}}, \frac{1}{b_{12}}\right) \leq \frac{1}{b_{23}} \leq \max\left(\frac{1}{b_{13}}, \frac{1}{b_{12}}\right) \tag{42}$$

and

$$\frac{1}{\chi_{12}} \leq \frac{1}{\chi_{23}} \leq \frac{1}{\chi_{13}}. \tag{43}$$

If $1/b_{13} \leq 0$, then (42) immediately reduces to $1/b_{13} \leq 1/b_{23} \leq 1/b_{12}$ (by (40), we are considering a special case in which $1/b_{12} > 0$). This is illustrated in Fig. 13 for the slightly different situations $m_{12} < m_{13}$ and $m_{12} \geq m_{13}$. If, on the other hand, $1/b_{13} > 0$, then (42) and (43) together imply two possible situations, depending on whether $1/b_{13} < 1/b_{12}$ or $1/b_{13} \geq 1/b_{12}$. These possibilities are illustrated in Fig. 14.

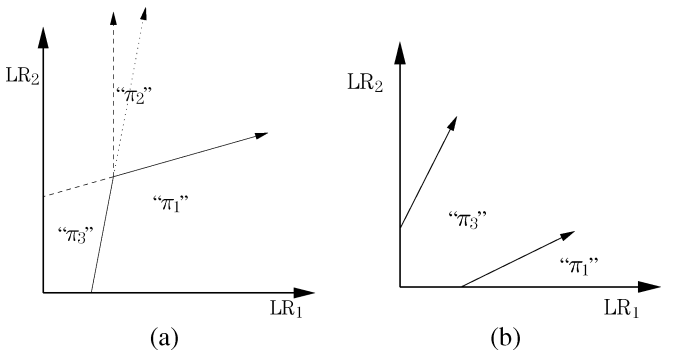

Fig. 13. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} < 0$ (implying $1/b_{12} > 0$ and $1/\chi_{12} < 0$) and $1/b_{13} \leq 0$. In (a), $m_{12} < m_{13}$, and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $b_{23} > 0$, and therefore $1/b_{23} \geq 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $m_{12} \geq m_{13}$, and the "2-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "1-vs-3" lines shown); observations above this line will be decided "π_2," and those below this line will be decided "π_3."

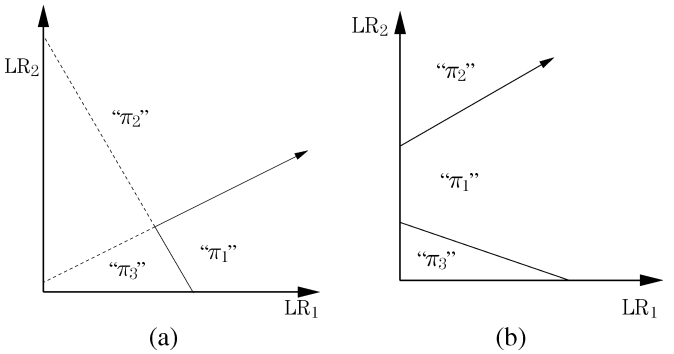

Fig. 14. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} < 0$ (implying $1/b_{12} > 0$ and $1/\chi_{12} < 0$) and $1/b_{13} > 0$. In (a), $1/b_{13} < 1/b_{12}$, and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3". In (b), $1/b_{13} \geq 1/b_{12}$, and the "2-vs-3" line can lie anywhere between the "1-vs-2" and "1-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_1" regardless of the position of this line.

Finally, we consider the case $\gamma_{323} - \gamma_{313} = 0$ (ie, $\gamma_{313} = \gamma_{323}$, or $U_{2|3} = U_{1|3}$), in which both $1/b_{12}$ and $1/\chi_{12}$ are infinite. We now have

$$\frac{1}{b_{13}} \leq \frac{1}{b_{23}} \tag{44}$$

and

$$\frac{1}{\chi_{23}} \leq \frac{1}{\chi_{13}}. \tag{45}$$

Together, (44) and (45) can be considered *either* a special case of the inequalities (38) and (39), if we take $1/b_{12} = -\infty$ and $1/\chi_{12} = +\infty$; *or* of the inequalities (42) and (43), if we take $1/b_{12} = +\infty$ and $1/\chi_{12} = -\infty$. This situation, for the slightly different cases $1/b_{13} \leq 0$ and $1/b_{13} > 0$, is illustrated in Fig. 15.

Notice that every figure in this appendix has one or more corresponding figures in Section III (depending on the possible

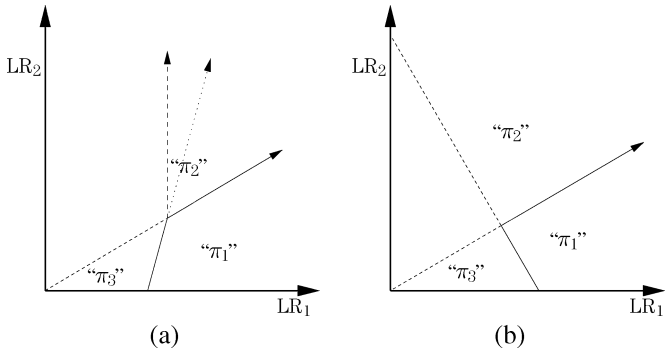

Fig. 15. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} = 0$ (implying $1/b_{12} = \mp\infty$ and $1/\chi_{12} = \pm\infty$). In (a), $1/b_{13} \leq 0$, and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $1/b_{23} \geq 0$); observations in the unlabeled region to above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $1/b_{13} > 0$, and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3."

values of the undetermined decision boundary parameter being illustrated in that figure). Specifically

$$\begin{aligned}
\text{Fig. } 11(\text{a}) &\Rightarrow \text{Figs. } 1(\text{a}),\ 4(\text{a}),\ 5(\text{a}) \\
\text{Fig. } 11(\text{b}) &\Rightarrow \text{Fig. } 4(\text{b}) \\
\text{Fig. } 12(\text{a}) &\Rightarrow \text{Figs. } 1(\text{a}),\ 3(\text{a}),\ 5(\text{a}) \\
\text{Fig. } 12(\text{b}) &\Rightarrow \text{Figs. } 1(\text{b}),\ 3(\text{b}),\ 5(\text{a}) \\
\text{Fig. } 13(\text{a}) &\Rightarrow \text{Figs. } 3(\text{a}),\ 4(\text{a}),\ 5(\text{b}) \\
\text{Fig. } 13(\text{b}) &\Rightarrow \text{Fig. } 4(\text{b}) \\
\text{Fig. } 14(\text{a}) &\Rightarrow \text{Fig. } 2(\text{a}) \\
\text{Fig. } 14(\text{b}) &\Rightarrow \text{Fig. } 2(\text{b}) \\
\text{Fig. } 15(\text{a}) &\Rightarrow \text{Figs. } 3(\text{a}),\ 4(\text{a}),\ 5(\text{b}) \\
\text{Fig. } 15(\text{b}) &\Rightarrow \text{Figs. } 2(\text{a}),\ 3(\text{a}),\ 4(\text{b}).
\end{aligned}$$

That is, none of the conditions derived in this appendix are inconsistent with those derived in Section III or Appendix A. More importantly, note the symmetry between the corresponding equations and figures in Sections III and this appendix, if one "swaps" the labels of classes π_2 and π_3, and additionally replaces m_{ij} with $1/\chi_{i'j'}$, χ_{ij} with $1/m_{i'j'}$, and b_{ij} with $1/b_{i'j'}$ ($i' = 1$ if $i = 1$, 2 if $i = 3$, and 3 if $i = 2$; similarly for j).

ACKNOWLEDGMENT

The authors thank the associate editor and anonymous reviewers for their suggestions to substantially improve the content and structure of this manuscript.

REFERENCES

[1] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic, 1975.

[2] C. E. Metz, "Basic principles of ROC analysis," *Sem. Nucl. Med.*, vol. VIII, no. 4, pp. 283–298, 1978.

[3] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*. New York: Wiley, 1968.

[4] B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math Psych.*, vol. 40, pp. 253–269, 1996.

[5] ——, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psych.*, vol. 42, pp. 5–31, 1998.

[6] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, pp. 78–89, 1999.

[7] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study," *Proc. SPIE Medical Imaging 2003: Image Processing*, vol. 5032, pp. 567–578, 2003.

[8] U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.*, vol. 2, pp. 1–9, 1995.

[9] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images," *Med. Phys.*, vol. 18, pp. 955–963, 1991.

[10] F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.*, vol. 28, pp. 473–481, 1993.

[11] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.*, vol. 21, pp. 445–452, 1994.

[12] M. A. Kupinski, "Computerized pattern classification in medical imaging," Ph.D. thesis, The Univ. Chicago, Chicago, IL, 2000.

[13] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[14] Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.*, vol. 44, pp. 2579–2595, 1999.

[15] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness," *Acad. Radiol.*, vol. 7, pp. 1077–1084, 2000.

[16] Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1285–1292, Dec. 2001.

[17] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: effectiveness of computer-aided diagnosis—observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560–568, 2002.

[18] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 891–895, Jul. 2004.

[19] 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.*, vol. 24, no. 3, pp. 293–299, Mar. 2005.

[20] A. Srinivasan, "Note on the Location of Optimal Classifiers in n-Dimensional ROC Space," Oxford Univ. Computing Lab., Oxford, U.K., Tech. Rep. PRG-TR-2-99, 1999.

[21] C. Ferri, J. Hernández-Orallo, and M. A. Salido, "Volume Under the roc Surface for Multi-Class Problems: Exact Computation and Evaluation of Approximations," Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València, València, Spain, Tec. Rep. 2003.

[22] C. E. Metz, "The optimal decision variable," Dept. Radiol., Univ. Chicago, unpublished lecture notes for the course Mathematics for Medical Physicists. 2000.

B Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule

Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule ⋆

Darrin C. Edwards† and Charles E. Metz

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

Abstract

We analyze recently proposed decision rules for three-class classification from the point of view of ideal observer decision theory. We consider three-class decision rules proposed by Scurfield, by Chan et al., and by Mossman. Scurfield's decision rule is shown to be a special case of the three-class ideal observer decision rule in three different situations. Chan et al. start with an ideal observer model and specify its decision-consequence utility structure in a way that causes two of the decision lines used by the ideal observer to overlap and the third line to become undefined. Finally, we show that, for a particular and obvious choice of ideal-observer-related decision variables, the Mossman decision rule cannot be a special case of the ideal observer decision rule. Despite the considerable difficulties presented by the three-class classification task, the three-class ideal observer provides a useful framework for analyzing a variety of three-class decision strategies.

Key words: ROC analysis, three-class classification, ideal observer decision rules

1 Introduction

We are attempting to develop a fully automated mass lesion classification scheme for computer-aided diagnosis (CAD) in mammography. This scheme will combine two schemes developed at the University of Chicago: one for automatically detecting mass lesions in mammograms (Bick, Giger, Schmidt,

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

† Corresponding author. Telephone: 773 834 5094; Fax: 773 702 0371
Email address: d-edwards@uchicago.edu (Darrin C. Edwards).

Nishikawa, Wolverton, and Doi, 1995; Yin, Giger, Doi, Metz, Vyborny, and Schmidt, 1991; Yin, Giger, Vyborny, Doi, and Schmidt, 1993; Yin, Giger, Doi, Vyborny, and Schmidt, 1994; Kupinski, 2000), and one for classifying known lesions as malignant or benign (Huo, Giger, Vyborny, Wolverton, Schmidt, and Doi, 1998; Huo, Giger, and Metz, 1999; Huo, Giger, Vyborny, Wolverton, and Metz, 2000; Huo, Giger, and Vyborny, 2001; Huo, Giger, Vyborny, and Metz, 2002). Combining these two types of CAD scheme is inherently difficult, because the output of the detection scheme will necessarily include false-positive (FP) computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our approach has been to treat this problem explicitly as a three-class classification task. That is, the outputs of the detection scheme should be classified as malignant lesions, benign lesions, and non-lesions (FP computer detections), and the classifier to be estimated is the ideal observer decision rule for this task. Such an approach presents considerable difficulties of its own. On the one hand, decision rules, in particular ideal observer decision rules, increase rapidly in complexity with the number of classes involved. On the other hand, a fully general performance evaluation method, such as a three-class extension of receiver operating characteristic (ROC) analysis, has yet to be developed. It should be mentioned that the simple model we have just described corresponds in the two-class classification task to ROC analysis performed "per detection;" that is, each "case" being classified corresponds to a small region of interest (ROI) in the image containing a single computer detection. Other formulations, such as ROC analysis "per image," ROC analysis "per patient" (for a set of images, such as the four mammographic views obtained in a typical screening setting), or free-response ROC (FROC) (Bunch, Hamilton, Sanderson, and Simmons, 1978; Chakraborty, 1989, 2002) analysis, are also possible, but their extension to tasks with three or more classes is beyond the scope of the present work.

The explicit form of the decision rule used by the ideal observer in a three-class classification task has been known for some time (Van Trees, 1968). For the reasons just stated, however, a practical and general method for estimating and evaluating observer performance has proven elusive. In particular, Scurfield (1996) defined the two-class information-based performance metric $D_{1:2} \equiv \log 2 - \mathrm{AUC} \log \mathrm{AUC} - (1 - \mathrm{AUC}) \log(1 - \mathrm{AUC})$ (where AUC is the area under the two-class ROC curve), and extended it to the three-class case for two different decision rules (Scurfield, 1996, 1998). Srinivasan (1999) investigated the optimality of discrete, multi-class ROC operating points, but not

continuous ROC hypersurfaces, under a cost function equivalent to the Bayes risk. Mossman (1999) evaluated the performance of a three-class classifier with a surface formed from the three correct classification probabilities. Hand and Till (2001) proposed the average of the areas under all $N(N-1)/2$ between-class ROC curves as a performance metric in an N-class classification task. Obuchowski, Applegate, Goske, Arheart, Myers, and Morrison (2001) elicited readers' estimates of the set of probabilities of each observation belonging to N classes, and then used conventional (two-class) ROC analysis to evaluate each of the $N(N-1)/2$ differences of these estimates for its ability to distinguish between the relevant pair of classes. Ferri, Hernández-Orallo, and Salido (2003) proposed a variety of algorithms for calculating the hypervolume under the convex hull obtained from a set of discrete ROC operating points; a modified version of the Hand and Till metric averaging the N areas under the ROC surfaces that measure the observer's ability to distinguish a given class from the remaining $N-1$; and a graphical "cobweb" representation of the observer's misclassification probabilities. Lachiche and Flach (2003) proposed iterative algorithms for finding the optimal among a discrete set of multi-class ROC operating points based on either percent correct or Bayes risk. Nakas and Yiannoutsos (2004) considered an observer using a decision rule similar to that of Scurfield (1996), and evaluated its performance statistically by extending methods proposed by Dreiseitl, Ohno-Machado, and Binder (2000). Patel and Markey (2005) applied a variety of proposed evaluation metrics, including the Hand and Till metric, the modified Hand and Till metric of Ferri, the "cobweb" graphical measure of Ferri, and the Mossman ROC surface, to radiologist assessment data of mammographic images from patients who subsequently underwent biopsy.

The works cited above demonstrate the difficulty in developing a fully general performance metric for classification tasks with more than two classes. Lacking such a performance metric in turn makes the development of observer decision rules for such tasks difficult, because they can at present be evaluated and compared only from a theoretical rather than an empirical perspective. Nevertheless, observer decision rule models for three-class classification tasks have been proposed relatively recently by several groups of researchers. In some cases, these models are motivated more by considerations of tractability than of complete generality. This is of course understandable given the inherent difficulties of three-class classification; however, we thought it might be of interest to analyze a number of recently proposed three-class decision rule models within an ideal observer decision rule framework.

In the next section, we review the three-class ideal observer decision rule. In the following three sections, we review recently proposed three-class decision rule models: one by Scurfield (1998), one by Chan, Sahiner, Hadjiiski, Petrick, and Zhou (2003), and one by Mossman (1999). In each case, the given decision rule is analyzed in terms of the ideal observer decision rule; where necessary

or expedient, assumptions are made about the observer's decision variables in order to facilitate this analysis. We emphasize that we do not attempt a review of the experimental methods or detailed analysis of proposed performance evaluation metrics in the works discussed; we are here interested only in the form of the decision rule which serves as the starting point for each work, and superficially in the proposed evaluation metrics inasmuch as they are related to those decision rules. (Because of the lack of a fully general performance metric, or figure of merit, for the three-class classification task, in particular apparent inconsistencies which are obtained from a straightforward generalization of the area under the ROC curve (Edwards, Metz, and Nishikawa, 2005), we do not attempt any validation or quantitative comparison of the proposed performance metrics.) The results of our analyses are briefly summarized in Sec. 6.

2 The Three-Class Ideal Observer

It can be shown (Van Trees, 1968; Edwards, Metz, and Kupinski, 2004b) that an N-class ideal observer makes decisions regarding statistically variable observations $\vec{\mathbf{x}}$ by partitioning a likelihood ratio decision variable space, where the boundaries of the partitions are given by hyperplanes:

$$\text{decide} \quad d = \pi_i \ \text{ iff}$$

$$\sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k) \mathrm{LR}_k \ \geq \ (U_{j|N} - U_{i|N}) P(\mathbf{t} = \pi_N) \quad \{j < i\} \quad (1)$$

$$\text{and}$$

$$\sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k) \mathrm{LR}_k \ > \ (U_{j|N} - U_{i|N}) P(\mathbf{t} = \pi_N) \quad \{j > i\}. \quad (2)$$

Here $U_{i|j}$ is the utility of deciding an observation is from class π_i given that it is actually from class π_j, and the $N-1$ likelihood ratios are defined as

$$\mathrm{LR}_k \equiv \frac{p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_k)}{p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_N)} \quad (3)$$

for $k < N$. We also define the actual class (the "truth") to which an observation belongs as $\mathbf{t}$, and the class to which it is assigned (the "decision") as $\mathbf{d}$, where $\mathbf{t}$ and $\mathbf{d}$ can take on any of the values $\pi_1, \ldots, \pi_i, \ldots, \pi_N$, the labels of the various classes. (We use boldface type to denote statistically variable quantities.) For simplicity, we will usually write π_k to denote the event $\mathbf{t} = \pi_k$, as in the *a priori* probability $P(\pi_k)$.

The partitioning of the decision variable space is determined by the parameters

$$\gamma_{ijk} \equiv (U_{i|k} - U_{j|k})P(\pi_k), \tag{4}$$

with i, j, and k varying from 1 to N, and $j \neq i$. Note that these parameters are not independent, however, because

$$\gamma_{ijk} = \gamma_{kjk} - \gamma_{kik}. \tag{5}$$

We can impose the reasonable condition that the utility for correctly classifying an observation from a given class should be greater than any utility for incorrectly classifying an observation from the same class, *i. e.*, $U_{i|i} > U_{j|i} \quad \{i \neq j\}$. This gives, for $j \neq i$,

$$\gamma_{iji} > 0, \tag{6}$$

leaving $N(N-1)$ parameters (the rest are derivable from (5)).

Finally, note that the hyperplanes represented by (1) and (2) are unchanged if we multiply all of these relations by a single scalar, such as $1/(\sum_{i\neq j} \gamma_{iji})$. This leaves us with $N^2 - N - 1$ degrees of freedom, as expected, and effectively imposes the condition

$$\sum_{i\neq j} \gamma_{iji} = 1. \tag{7}$$

The behavior of a three-class ideal observer is completely determined by the three decision boundary lines

$$\gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 = \gamma_{313} - \gamma_{323} \tag{8}$$

$$\gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 = \gamma_{313} \tag{9}$$

$$(\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 = \gamma_{323}, \tag{10}$$

which we call, respectively, the "1-*vs.*-2" line, the "1-*vs.*-3" line, and the "2-*vs.*-3" line. Note that if any two of these lines intersect, the third line must also share this intersection point. We also emphasize the simple interpretation, from (4), of each of the γ_{iji} parameters appearing in these decision boundary line equations as the difference in utilities between a "correct" and one particular "incorrect" decision (scaled by the *a priori* probability of the true class in question); and of each difference in the γ_{iji} parameters as a difference in utilities between two possible "incorrect" decisions (again scaled by the *a priori* probability of the true class in question).

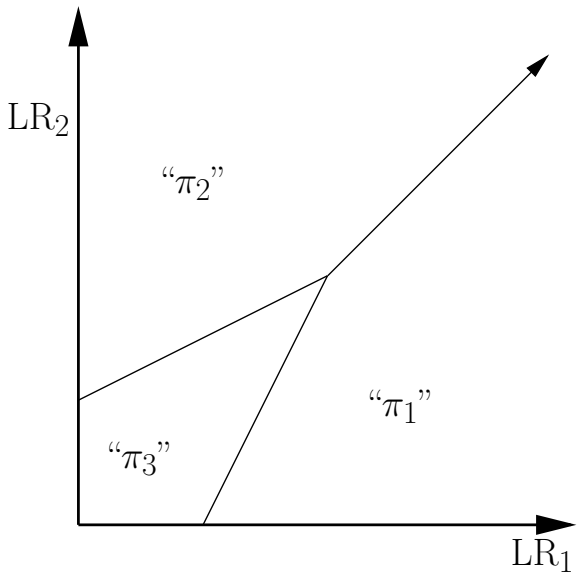

Fig. 1. Example three-class ideal observer decision rule, given the values of the decision parameters $\gamma_{121} = \gamma_{212} = 3/14$ and $\gamma_{131} = \gamma_{313} = \gamma_{232} = \gamma_{323} = 1/7$. Note that $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\mathbf{t} = \pi_i)$.

An example ideal observer decision rule for particular values of the utilities $U_{i|j}$, and hence of the parameters γ_{iji}, is shown in Fig. 1. Here we have chosen $\gamma_{121} = \gamma_{212} = 3/14$ and $\gamma_{131} = \gamma_{313} = \gamma_{232} = \gamma_{323} = 1/7$, yielding the decision boundary lines

$$\frac{3}{14}\mathrm{LR}_1 - \frac{3}{14}\mathrm{LR}_2 = 0 \quad \{\text{“1-}vs\text{.-2”}\} \tag{11}$$

$$\frac{1}{7}\mathrm{LR}_1 - \frac{1}{14}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{“1-}vs\text{.-3”}\} \tag{12}$$

$$-\frac{1}{14}\mathrm{LR}_1 + \frac{1}{7}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{“2-}vs\text{.-3”}\}. \tag{13}$$

These simplify to the equations $\mathrm{LR}_2 = \mathrm{LR}_1$, $\mathrm{LR}_2 = 2\mathrm{LR}_1 - 2$, and $\mathrm{LR}_2 = \mathrm{LR}_1/2 + 1$, respectively.

3 The Scurfield Decision Rule

Scurfield investigated a decision rule applied to two-dimensional statistically variable data ($\vec{\mathbf{y}} \equiv (\mathbf{y}_1, \mathbf{y}_2)$) drawn from three classes (Scurfield, 1998). The application domain was human observer performance modeling for acoustical psychophysics experiments. (In prior work, Scurfield investigated a decision rule for three-class classification of univariate data (Scurfield, 1996). We will not review that prior work here, because at present we are interested in relating given observer models to the general three-class ideal observer model for

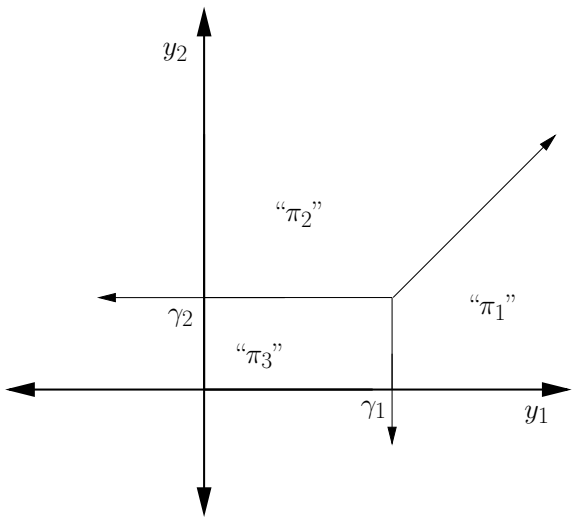

Fig. 2. Decision rule investigated by Scurfield, for the decision parameters γ_1 and γ_2.

multivariate observational data, which — except in degenerate cases — will yield two-dimensional decision variable data by (3).) In Scurfield's work, no assumptions are made about the decision variables $\mathbf{y}_1$ and $\mathbf{y}_2$; in particular, these decision variables are not assumed to be related in any way to an ideal observer model. This is entirely appropriate given the nature of the problem domain Scurfield investigated — *i. e.*, human observer performance modeling. It can readily be shown, however, that if one chooses to make such assumptions, special cases of the Scurfield model are in fact special cases of an ideal observer decision rule.

The Scurfield decision rule is dependent on two decision parameters, which we will call γ_1 and γ_2. The decision rule can be written as

$$\text{decide} \quad d = \pi_1 \quad \text{iff} \quad y_1 - y_2 \geq \gamma_1 - \gamma_2 \quad \text{and} \quad y_1 \geq \gamma_1; \tag{14}$$

$$\text{decide} \quad d = \pi_2 \quad \text{iff} \quad y_1 - y_2 < \gamma_1 - \gamma_2 \quad \text{and} \quad y_2 \geq \gamma_2; \tag{15}$$

$$\text{decide} \quad d = \pi_3 \quad \text{iff} \quad y_1 < \gamma_1 \quad \text{and} \quad y_2 < \gamma_2. \tag{16}$$

This decision rule is illustrated in Fig. 2.

From these relations, one can define the decision boundary lines

$$y_1 - y_2 = \gamma_1 - \gamma_2 \quad \{\text{"1-}vs\text{.-2"}\} \tag{17}$$

$$y_1 = \gamma_1 \quad \{\text{"1-}vs\text{.-3"}\} \tag{18}$$

$$y_2 = \gamma_2 \quad \{\text{"2-}vs\text{.-3"}\}. \tag{19}$$

If we choose $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$ for some set of observational data $\vec{\mathbf{x}}$, we have

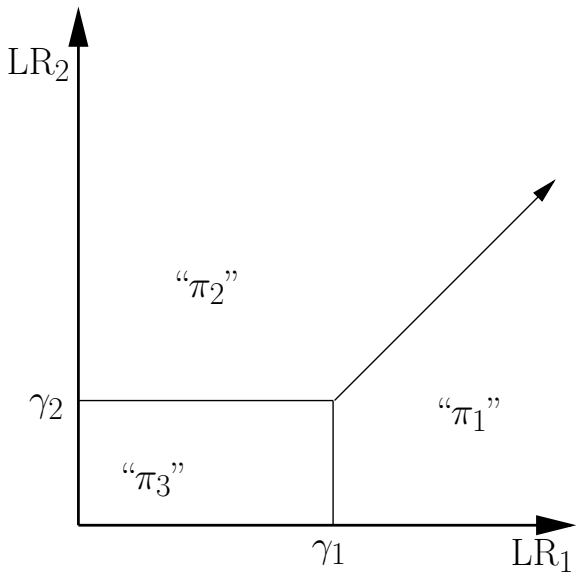

Fig. 3. A special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{212} = \gamma_{131} = \gamma_{232} = 1/(\gamma_1 + \gamma_2 + 4)$, $\gamma_{313} = \gamma_1/(\gamma_1 + \gamma_2 + 4)$, and $\gamma_{323} = \gamma_2/(\gamma_1 + \gamma_2 + 4)$. The parameters γ_1 and γ_2 are positive but otherwise arbitrary; this decision rule is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$.

$$\frac{1}{\gamma_0}\mathrm{LR}_1 - \frac{1}{\gamma_0}\mathrm{LR}_2 = \frac{\gamma_1 - \gamma_2}{\gamma_0} \quad \{\text{“1-}vs\text{.-2”}\} \tag{20}$$

$$\frac{1}{\gamma_0}\mathrm{LR}_1 = \frac{\gamma_1}{\gamma_0} \quad \{\text{“1-}vs\text{.-3”}\} \tag{21}$$

$$\frac{1}{\gamma_0}\mathrm{LR}_2 = \frac{\gamma_2}{\gamma_0} \quad \{\text{“2-}vs\text{.-3”}\}, \tag{22}$$

where $\gamma_0 \equiv \gamma_1 + \gamma_2 + 4$ (to impose consistency with (7)). Note the similarity in form between these equations and (8)–(10). If we require γ_1 and γ_2 to be positive, the correspondence is exact, and this special case of (8)–(10) is illustrated in Fig. 3. (In fact, the intersection of the ideal observer decision boundary lines can lie in any quadrant. However, given a set of decision boundary lines with slopes as depicted in Fig. 2, the occurrence of the intersection point in any quadrant other than the first would result in an ideal observer operating point for which no observations were assigned to class π_3. This “degenerate” case will not be considered here.) As an aside, it is of some interest to note that if $\gamma_1 = \gamma_2 = 1$, the decision boundary line equations reduce to $\mathrm{LR}_1 = \mathrm{LR}_2$, yielding $p(\vec{x}|\pi_1) = p(\vec{x}|\pi_2)$; $\mathrm{LR}_1 = 1$, yielding $p(\vec{x}|\pi_1) = p(\vec{x}|\pi_3)$; and $\mathrm{LR}_2 = 1$, yielding $p(\vec{x}|\pi_2) = p(\vec{x}|\pi_3)$. That is, the decision boundary lines correspond, in the observational data space, to the loci of intersection of the observational data probability density functions. (This is illustrated in Figs. 2B and 2C of Scurfield (1998).)

A second correspondence between Scurfield's decision rule and the ideal observer decision rule can be obtained by taking $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$, with γ_1 and γ_2 now unrestricted. Substituting this definition in (17)–(19), we obtain

$$\log(\mathrm{LR}_1) - \log(\mathrm{LR}_2) = \gamma_1 - \gamma_2 \quad \{\text{“1-}vs\text{.-2”}\} \tag{23}$$
$$\log(\mathrm{LR}_1) = \gamma_1 \quad \{\text{“1-}vs\text{.-3”}\} \tag{24}$$
$$\log(\mathrm{LR}_2) = \gamma_2 \quad \{\text{“2-}vs\text{.-3”}\}. \tag{25}$$

Taking exponentials on each side of these equations then gives

$$\frac{\mathrm{LR}_1}{\mathrm{LR}_2} = e^{\gamma_1 - \gamma_2} \quad \{\text{“1-}vs\text{.-2”}\} \tag{26}$$
$$\mathrm{LR}_1 = e^{\gamma_1} \quad \{\text{“1-}vs\text{.-3”}\} \tag{27}$$
$$\mathrm{LR}_2 = e^{\gamma_2} \quad \{\text{“2-}vs\text{.-3”}\}; \tag{28}$$

we can then rearrange terms and divide the equations by a constant factor to obtain

$$\frac{e^{-\gamma_1}}{\gamma_0}\mathrm{LR}_1 - \frac{e^{-\gamma_2}}{\gamma_0}\mathrm{LR}_2 = 0 \quad \{\text{“1-}vs\text{.-2”}\} \tag{29}$$
$$\frac{e^{-\gamma_1}}{\gamma_0}\mathrm{LR}_1 = \frac{1}{\gamma_0} \quad \{\text{“1-}vs\text{.-3”}\} \tag{30}$$
$$\frac{e^{-\gamma_2}}{\gamma_0}\mathrm{LR}_2 = \frac{1}{\gamma_0} \quad \{\text{“2-}vs\text{.-3”}\}, \tag{31}$$

where $\gamma_0 \equiv 2(e^{-\gamma_1} + e^{-\gamma_2} + 1)$. By inspection, this is again a special case of (8)–(10), which is illustrated in Fig. 4. (This special case is currently the subject of independent analysis by He, Metz, Tsui, Links, and Frey (2006).) As an aside, we note that if $\gamma_1 = \gamma_2 = 0$, the resulting decision boundary lines again correspond, in the observational data space, to the loci of intersection of the observational data probability density functions, as was pointed out in the text following (20)–(22).

Finally, if we take $\mathbf{y}_1 \equiv P(\pi_1|\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv P(\pi_2|\vec{\mathbf{x}})$, and require $0 < \gamma_1 < 1$ and $0 < \gamma_2 < 1$, we obtain

$$P(\pi_1|\vec{x}) - P(\pi_2|\vec{x}) = \gamma_1 - \gamma_2 \quad \{\text{“1-}vs\text{.-2”}\} \tag{32}$$
$$P(\pi_1|\vec{x}) = \gamma_1 \quad \{\text{“1-}vs\text{.-3”}\} \tag{33}$$
$$P(\pi_2|\vec{x}) = \gamma_2 \quad \{\text{“2-}vs\text{.-3”}\}, \tag{34}$$

as illustrated in Fig. 5.

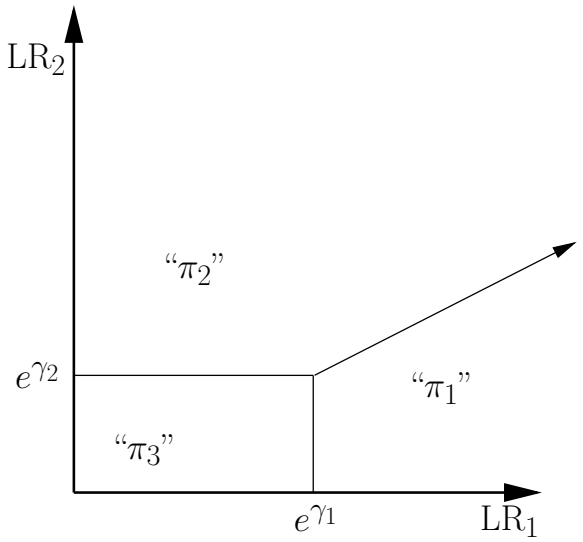

Fig. 4. A special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{131} = e^{-\gamma_1}/\gamma_0$, $\gamma_{212} = \gamma_{232} = e^{-\gamma_1}/\gamma_0$, $\gamma_{313} = \gamma_{323} = 1/\gamma_0$, and $\gamma_0 \equiv 2(e^{-\gamma_1} + e^{-\gamma_2} + 1)$. The parameters γ_1 and γ_2 are arbitrary; this decision rule is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$.

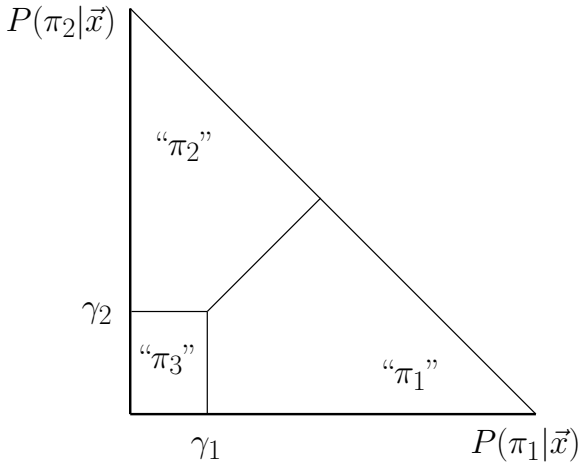

Fig. 5. A special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv P(\pi_1|\vec{x})$ and $\mathbf{y}_2 \equiv P(\pi_2|\vec{x})$.

Note that (3) can be written as

$$\mathrm{LR}_i = \frac{P(\pi_i|\vec{x})p(\vec{x})/P(\pi_i)}{p(\vec{x}|\pi_3)} \qquad \{i : 1 \le i \le 2\}$$

$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i P(\pi_i)}{p(\vec{x})/p(\vec{x}|\pi_3)}$$

$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i[P(\pi_i)/P(\pi_3)]}{1 + \mathrm{LR}_1[P(\pi_1)/P(\pi_3)] + \mathrm{LR}_2[P(\pi_2)/P(\pi_3)]}. \tag{35}$$

This allows us to rewrite (32)–(34) as

$$\frac{1-(\gamma_1-\gamma_2)}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 - \frac{1+(\gamma_1-\gamma_2)}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_1-\gamma_2}{\gamma_0} \tag{36}$$

$$\frac{1-\gamma_1}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 - \frac{\gamma_1}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_1}{\gamma_0} \tag{37}$$

$$-\frac{\gamma_2}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 + \frac{1-\gamma_2}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_2}{\gamma_0}, \tag{38}$$

respectively, where $\gamma_0 \equiv (2-2\gamma_1+\gamma_2)P(\pi_1)/P(\pi_3)+(2+\gamma_1-2\gamma_2)P(\pi_2)/P(\pi_3)+\gamma_1+\gamma_2$. This is again a special case of (8)–(10), as the quantities $1-(\gamma_1-\gamma_2)$, $1+(\gamma_1-\gamma_2)$, $1-\gamma_1$, and $1-\gamma_2$ are all positive given $0<\gamma_1<1$ and $0<\gamma_2<1$.

Scurfield (1998) points out that the observer which maximizes P_C, the "percent correct" or probability of a correct response, is a special case of the ideal observer (*i. e.*, a single operating point achievable by the ideal observer for the given task). This observer follows the Scurfield decision rule model with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$, and decision parameters given by $e^{\gamma_1} = P(\pi_3)/P(\pi_1)$ and $e^{\gamma_2} = P(\pi_3)/P(\pi_2)$. It is interesting to note that the Scurfield decision rule model can in fact be used to describe ideal observer performance for an even wider class of operating points, as shown in this section.

To evaluate the performance of an observer using the decision rule in (17)–(19), Scurfield plots a set of six surfaces in three-dimensional ROC spaces, giving $P(\mathbf{d}=\pi_2|\mathbf{t}=\alpha(\pi_2))$ as a function of $P(\mathbf{d}=\pi_1|\mathbf{t}=\alpha(\pi_1))$ and $P(\mathbf{d}=\pi_3|\mathbf{t}=\alpha(\pi_3))$. Here α is one of the six possible permutations of three symbols. Scurfield gives a probabilistic interpretation for this evaluation methodology: the volume under each surface is the probability of a particular outcome in a three-alternative forced choice experiment, and thus the six volumes must sum to one. This constraint means that at most five of the surfaces are independent. However, given the number of conditional probabilities $P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j)$ involved, one can show that only four such surfaces are required to completely specify the tradeoffs among the observer's conditional classification probabilities. Without loss of generality, we consider plotting each of $P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_1)$, $P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_3)$, $P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_1)$, and $P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_2)$ as functions of $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_2)$ and $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_3)$. (As with Scurfield's plots, these are well defined because Scurfield's decision rule has two degrees of freedom, namely the parameters γ_1 and γ_2.)

Now consider one of Scurfield's plots, for example that which gives $P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_2)$ as a function of $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_1)$ and $P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_3)$. Because these are conditional probabilities, we have

$$P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_1) = 1 - P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_1) - P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_1) \tag{39}$$

$$P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_2) = 1 - P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_2) - P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_2) \tag{40}$$

$$P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3) = 1 - P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_3) - P(\mathbf{d} = \pi_2|\mathbf{t} = \pi_3). \quad (41)$$

Each of the conditional probabilities on the right hand side of these equations can be written as functions of $P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_2)$ and $P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_3)$ in our formulation; thus the surface given in this plot is determined parametrically by the set of four surfaces we have given. Similar remarks hold for the other five surfaces used by Scurfield. In general, for an N-class classification task using a Scurfield-type decision rule with $N - 1$ degrees of freedom (the generalization to N classes of (17)–(19)), one can show that a set of $(N - 1)^2$ hypersurfaces with $N - 1$ degrees of freedom in N-dimensional ROC spaces is necessary to fully characterize the observer's performance, although the interpretation of those hypersurfaces is not necessarily as straightforward or elegant as that provided for the $N! - 1$ hypersurfaces used by Scurfield.

4 The Chan Decision Rule

Chan et al. are investigating three-class classifiers for computer-aided diagnosis (Chan et al., 2003). Their work is motivated by reasoning similar in principle to that which we independently arrived at when we began to consider this problem. In particular, they consider a clinical situation in which observations must be classified as malignant, benign, or normal. The goal of their work is not just the psychophysical measurement of the performance of an existing (*e. g.*, human) observer, but the optimization of the performance of a system (containing components with parameters subject to experimental control, *e. g.* an artificial neural network) to aid a radiologist or clinician. Thus they are free, at least in theory, to start explicitly from an ideal observer model in constructing their decision rule.

In order to reduce the complexity of the ideal observer decision rule to manageable proportions, Chan et al. impose restrictions on the utilities used by their observer. In their formulation, the class we are labeling π_1 is the benign class; π_2, the normal class; and the malignant class is π_3. They further assume that the possible values of any utility $U_{i|j}$ are restricted to the interval $[0, 1]$. They then set $U_{1|1} = U_{2|2} = U_{3|3} = 1$ (*i. e.*, correctly identifying any case has maximal utility). Furthermore, they require $U_{2|1} = U_{1|2} = 1$ and $U_{1|3} = U_{2|3} = 0$ (*i. e.*, misidentifying a benign case as normal, or vice versa, has no significant cost reducing the utility of such a decision from the maximum, but misclassifying an actually malignant case as benign or normal has the minimum possible utility). Finally, $U_{3|1}$ and $U_{3|2}$ are assumed to have arbitrary values on the open interval $(0, 1)$ (*i. e.*, misclassifying an actually non-malignant case as malignant will have some cost reducing the utility of such a decision from the maximum, but such a misclassification is in some sense "better" than missing an actual malignancy). It is important to note

that these assumptions are arguably relevant to a reasonable model of a clinical situation, and are thus of interest beyond their superficial advantage in reducing the degrees of freedom involved in the observer's decision rule. We will, however, only consider the latter issue in the remainder of this section.

Substituting the values of the utilities given above into (4), we obtain decision boundary lines of the form

$$0\,\mathrm{LR}_1 + 0\,\mathrm{LR}_2 = 0 \qquad \{\text{“1-}vs\text{.-2”}\} \quad (42)$$

$$\frac{(1-U_{3|1})P(\pi_1)}{\gamma_0}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\pi_2)}{\gamma_0}\mathrm{LR}_2 = \frac{P(\pi_3)}{\gamma_0} \qquad \{\text{“1-}vs\text{.-3”}\} \quad (43)$$

$$\frac{(1-U_{3|1})P(\pi_1)}{\gamma_0}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\pi_2)}{\gamma_0}\mathrm{LR}_2 = \frac{P(\pi_3)}{\gamma_0} \qquad \{\text{“2-}vs\text{.-3”}\} \quad (44)$$

where $\gamma_0 \equiv 1 + P(\pi_3) - U_{3|1}P(\pi_1) - U_{3|2}P(\pi_2)$. Note that, as Chan et al. point out, the "1-*vs.*-2" line is in fact undefined for this choice of utilities, while the "1-*vs.*-3" and "2-*vs.*-3" lines are identical. This is a general consequence of (8)–(10); if any two of these equations yield identical lines, the third line must be undefined. (Note that, strictly speaking, the utility structure employed by Chan et al. is excluded from our formulation by the requirement stated in (6). However, this issue — *i. e.*, whether the ideal observer's performance should be considered to include such limiting cases — is largely a definitional, rather than a fundamental, issue, because (6) could just as readily have been formulated as a non-negativity constraint, rather than a strict inequality as we have chosen.)

The decision rule considered by Chan et al. is illustrated in Fig. 6. It can be argued that, in a sense, the output of this classifier belongs to only two classes, malignant and non-malignant; in particular, because (42) is undefined, this observer will never unequivocally decide $\mathbf{d} = \pi_1$ (benign) or $\mathbf{d} = \pi_2$ (normal). In fact, if $U_{3|1} = U_{3|2}$, the observer's performance is identical with that of a two-class ideal observer which distinguishes between the malignant and non-malignant (benign plus normal) classes. However, in the more general case in which $U_{3|1} \neq U_{3|2}$, the observer considered by Chan et al. is able to achieve ROC operating points not accessible by the two-class ideal observer. (That is, the three-class ideal observer can achieve points below the two-class ideal observer's ROC curve in a two-class ROC space, or, equivalently, points off the curve representing the two-class ideal observer's performance plotted in a three-class ROC space.) Intuitively, their observer makes decisions based on the three distribution functions of the observational data, even though the observer's output consists of only two possible responses.

Chan et al. evaluate the performance of their observer by plotting $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3)$ as a function of $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_1)$ and $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_2)$. Note that this single two-dimensional surface is sufficient to completely characterize the

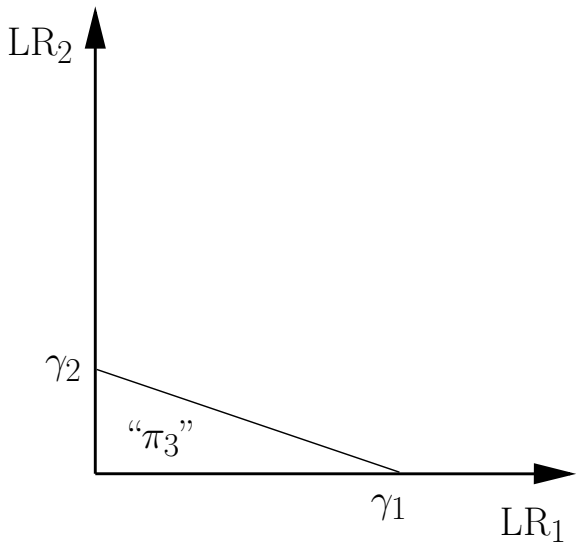

Fig. 6. The decision rule investigated by Chan et al., which is a special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{212} = 0$, $\gamma_{131} = (1 - U_{3|1})P(\pi_1)/\gamma_0$, $\gamma_{232} = (1 - U_{3|2})P(\pi_2)/\gamma_0$, and $\gamma_{313} = \gamma_{323} = P(\pi_3)/\gamma_0$; here $\gamma_0 \equiv 1 + P(\pi_3) - U_{3|1}P(\pi_1) - U_{3|2}P(\pi_2)$. Observations in the unlabeled region are decided "not π_3", *i. e.*, either "π_1" or "π_2". The intercepts γ_1 and γ_2 are $P(\pi_3)/[(1 - U_{3|1})P(\pi_1)]$ and $P(\pi_3)/[(1 - U_{3|2})P(\pi_2)]$, respectively.

tradeoffs among the conditional classification probabilities of their observer. This is because, as just stated, the observer's output consists of only two possible responses, and thus we have only six classification probabilities $P(\mathbf{d} = \pi_i|\mathbf{t} = \pi_j)$ rather than the nine expected in a three-class classification task. These six conditional probabilities are still constrained by three equations, however:

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_1) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_1) = 1 \quad (45)$$

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_2) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_2) = 1 \quad (46)$$

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_3) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3) = 1, \quad (47)$$

where the expression $\mathbf{d} = \tilde{\pi}_3$ indicates that the observer decides that the observation does not belong to class π_3. These constraint equations allow us to eliminate three of the six conditional probabilities, leaving a single ROC surface with two degrees of freedom in a three-dimensional ROC space.

5 The Mossman Decision Rule

Mossman investigates (Mossman, 1999) a decision rule applied to a set of three decision variables $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$, subject to the constraint

$$\mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 = 1, \tag{48}$$

as well as $0 \leq \mathbf{y}_i \leq 1 \quad \{1 \leq i \leq 3\}$. This is consistent with the constraint on the *a posteriori* class probabilities, $P(\pi_1|\vec{\mathbf{x}}) + P(\pi_2|\vec{\mathbf{x}}) + P(\pi_3|\vec{\mathbf{x}}) = 1$; these quantities are known to be directly related to the likelihood ratio ideal observer decision variables (Kupinski, Edwards, Giger, and Metz, 2001; Edwards, Lan, Metz, Giger, and Nishikawa, 2004a). Mossman does not explicitly require, however, that the decision variables in (48) be the *a posteriori* class probabilities (*e. g.*, they may be noisy estimates of these quantities).

The decision rule considered by Mossman, which depends on two decision parameters γ_1 and γ_2, is

$$\begin{aligned}
&\text{decide} \quad d = \pi_1 \quad \text{iff} \quad y_2 - y_1 \leq \gamma_2 \quad \text{and} \quad y_3 \leq \gamma_1; && (49)\\
&\text{decide} \quad d = \pi_2 \quad \text{iff} \quad y_2 - y_1 > \gamma_2 \quad \text{and} \quad y_3 \leq \gamma_1; && (50)\\
&\text{decide} \quad d = \pi_3 \quad \text{iff} \quad y_3 > \gamma_1. && (51)
\end{aligned}$$

where $0 \leq \gamma_1 \leq 1$ and $-1 \leq \gamma_2 \leq 1$. From these relations, and given the relation $y_3 = 1 - y_1 - y_2$ from (48), one can define the decision boundary lines

$$\begin{aligned}
&y_1 - y_2 = -\gamma_2 \quad \{\text{“1-}vs\text{.-2”}\} && (52)\\
&y_1 + y_2 = 1 - \gamma_1 \quad \{\text{“1-}vs\text{.-3”}\} && (53)\\
&y_1 + y_2 = 1 - \gamma_1 \quad \{\text{“2-}vs\text{.-3”}\}. && (54)
\end{aligned}$$

This decision rule is illustrated in Fig. 7. Note that, similar to the Chan et al. decision rule, the “1-*vs.*-3” and “2-*vs.*-3” decision boundary lines are identical.

We now consider a special case of the Mossman decision rule in which $\mathbf{y}_1 = P(\pi_1|\vec{\mathbf{x}})$, $\mathbf{y}_2 = P(\pi_2|\vec{\mathbf{x}})$, and $\mathbf{y}_3 = P(\pi_3|\vec{\mathbf{x}})$ for some observational data vector $\vec{\mathbf{x}}$. As in Sec. 3, we make the substitution in (35); this allows us to rewrite (52)–(54) as

$$\begin{aligned}
&(1+\gamma_2)\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 - (1-\gamma_2)\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = -\gamma_2 \quad \{\text{“1-}vs\text{.-2”}\} && (55)\\
&\gamma_1\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 + \gamma_1\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = 1 - \gamma_1 \quad \{\text{“1-}vs\text{.-3”}\} && (56)
\end{aligned}$$

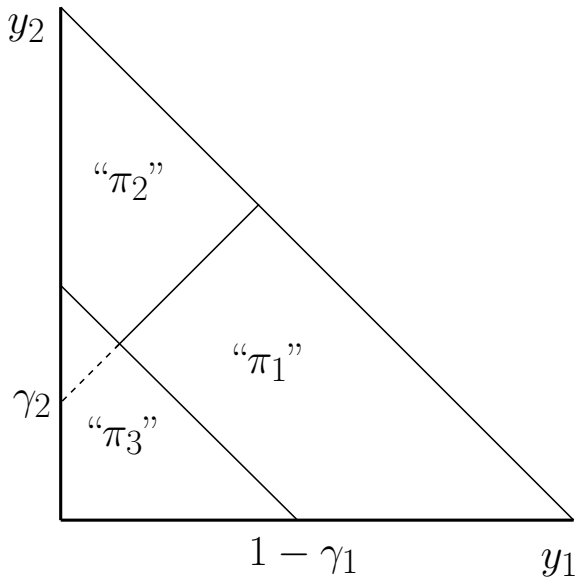

Fig. 7. Decision rule investigated by Mossman, for the decision parameters γ_1 and γ_2, shown in the *a posteriori* class probability space.

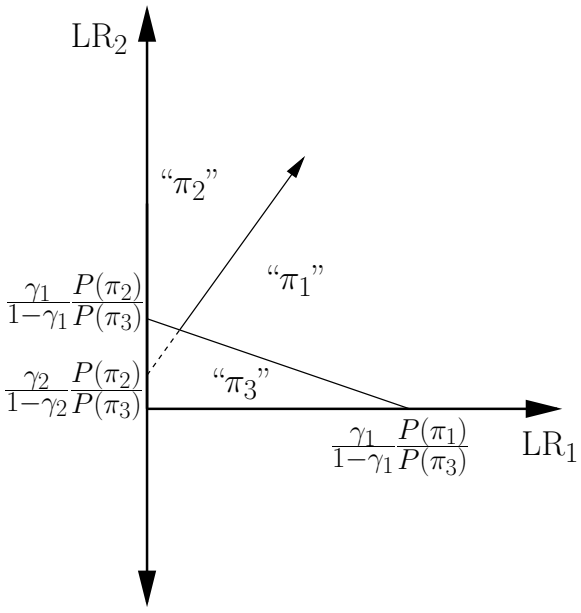

Fig. 8. Decision rule investigated by Mossman, for the decision parameters γ_1 and γ_2, shown in likelihood ratio space.

$$\gamma_1 \frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 \qquad + \gamma_1 \frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = 1-\gamma_1 \quad \{\text{“2-}\textit{vs}\text{.-3”}\}, \tag{57}$$

This version of the decision rule is illustrated in Fig. 8.

Although the Mossman decision rule for this choice of decision variables appears similar in form to the ideal observer decision rule, recall from Sec. 4 that if two of the decision boundary line equations are identical, the third

must yield a line identical to the first two or be undefined. Another way to see this is to note that the coefficients of (10) are differences of the corresponding coefficients of (8) and (9). If the coefficients of (9) and (10) are identical, it must be the case that the coefficients of (8) are all zero. For the Mossman decision rule, this would require $1+\gamma_2 = 0$, $1-\gamma_2 = 0$, and $\gamma_2 = 0$ simultaneously, which is clearly impossible.

It follows that, for this particular choice of decision variables (related in a straightforward way to the ideal observer's decision variables), the decision rule considered by Mossman cannot represent possible ideal observer performance for any choice of the utilities $U_{i|j}$ in (1) and (2). (One can construct probability density functions such that the Mossman observer's behavior for a particular choice of decision criteria (γ_1 and γ_2 in (49)–(51)) corresponds to ideal observer behavior at a particular operating point. However, we do not at present have any reason to believe that this result can be generalized to arbitrary probability density functions or to arbitrary choices of decision criteria for a given choice of probability density functions.)

Mossman proposed that the ROC surface obtained by plotting $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3)$ as a function of $P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_1)$ and $P(\mathbf{d} = \pi_2|\mathbf{t} = \pi_2)$ be used to evaluate the performance of the observer. Although this surface is clearly well-defined (the Mossman decision rule has two degrees of freedom, namely the parameters γ_1 and γ_2), it follows from the discussion at the end of Sec. 3 that four such surfaces in three-dimensional ROC spaces are needed to completely characterize the tradeoffs among the observer's conditional classification probabilities.

6 Discussion and Conclusions

We examined three decision rules proposed recently for three-class classification tasks by different researchers. The basis for our evaluation was ideal observer decision theory, primarily because our own interest in the three-class classification task is its possible application to CAD. A major goal in the development of a computerized scheme for CAD is the optimization of the performance of that scheme, in order to provide the maximum benefit to clinicians and thus to their patients. It should thus be kept clearly in mind that the ideal observer framework may not be as relevant, for example, to work which is motivated by purely psychophysical considerations (Scurfield, 1996, 1998; Mossman, 1999) — *i. e.*, where the goal is to estimate of the properties of an existing observer.

That being said, the three-class classification task is difficult enough that it is perhaps worth making any attempt to analyze, from a single point of view, the

work of the relatively few researchers investigating this problem, even in cases where that point of view is not necessarily relevant to the underlying motivations for that work. We feel the insights we have gained from the analysis of various decision rules presented here should provide at least some justification for that claim.

In particular, Scurfield points out (Scurfield, 1998) that his proposed decision rule is in fact an ideal observer decision rule for a single ideal observer operating point, namely the observer which maximizes the probability of any correct response (or "percent correct" or P_C). We were able to show that, under various assumptions, a larger set of such correspondences between the Scurfield observer and the ideal observer exists.

Chan et al. are working on the application of three-class classification to CAD, and thus explicitly take the ideal observer as the starting point in the development of their decision rule (Chan et al., 2003). Although this rendered our analysis of that decision rule in terms of ideal observer decision theory largely trivial, their decision rule merits attention as an example of a situation in which the ideal observer is indeed making use of information from the three classes of observations (*i. e.*, its behavior is demonstrably different from that of a two-class ideal observer), while only producing two different responses for those observations. In two-class classification, the only corresponding examples are trivial: either the observer always calls observations positive (achieving an operating point of $(\mathrm{FPF} = 1, \mathrm{TPF} = 1)$, where FPF is the false-positive fraction and TPF the true-positive fraction) or always calls them negative $(\mathrm{FPF} = 0, \mathrm{TPF} = 0)$.

Finally, we showed that, given a particular and obvious choice of ideal-observer-related decision variables, the decision rule proposed by Mossman (Mossman, 1999) does not correspond to ideal observer behavior for any possible values of the observer's utilities. However, we note that the structure of the Mossman decision rule — a simple sequence of thresholds on single decision variables — may indeed serve as a reasonable model for human observer performance in certain situations, *e. g.*, differential diagnosis. That such a decision rule fails to be an ideal observer decision rule may be considered surprising, given the properties the Mossman decision rule shares with that of Chan et al. — in particular, the identity of two out of the three decision boundary lines. The reasons why one decision rule can be said to correspond to ideal observer behavior, while a rule similar in structure does not when used with a particular and obvious choice of decision variables, are connected to fundamental constraints on the ideal observer's behavior; given the inherent complexities of the three-class classification task, it is easy for such subtleties to be overwhelmed by other details. A close comparison of two possible three-class classification decision rules can thus provide an immediate and intuitive understanding of such properties, even though a complete and fully general solution to the

three-class classification problem remains elusive.

Acknowledgments

The authors thank Vit Drga for bringing Brian Scurfield's work to their attention, and thank both Vit Drga and Brian Scurfield for helpful conversations concerning that work. The authors thank Heang-Ping Chan and Berkman Sahiner for helpful conversations concerning their work.

References

Bick, U., Giger, M. L., Schmidt, R. A., Nishikawa, R. M., Wolverton, D. E., Doi, K., 1995. Automated segmentation of digitized mammograms. Acad. Radiol. 2, 1–9.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., Simmons, A. H., 1978. A free response approach to the measurement and characterization of radiographic-observer performance. J. Appl. Photogr. Eng. 4, 166–172.

Chakraborty, D. P., 1989. Maximum likelihood analysis of free-response operating characteristic (FROC) data. Med. Phys. 16, 561–568.

Chakraborty, D. P., 2002. Statistical power in observer-performance studies: Comparison of the receiver operating characteristic and free-response methods in tasks involving localization. Acad. Radiol. 9, 147–156.

Chan, H.-P., Sahiner, B., Hadjiiski, L. M., Petrick, N., Zhou, C., 2003. Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study. In: Milan Sonka, J. Michael Fitzpatrick (Eds.), Proc. SPIE Vol. 5032 Medical Imaging 2003: Image Processing. SPIE, Bellingham, WA, pp. 567–578.

Dreiseitl, S., Ohno-Machado, L., Binder, M., 2000. Comparing three-class diagnostic tests by three-way ROC analysis. Med. Decis. Making 20, 323–331.

Edwards, D. C., Lan, L., Metz, C. E., Giger, M. L., Nishikawa, R. M., 2004a. Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions. Med. Phys. 31, 81–90.

Edwards, D. C., Metz, C. E., Kupinski, M. A., 2004b. Ideal observers and optimal ROC hypersurfaces in N-class classification. IEEE Trans. Med. Imag. 23, 891–895.

Edwards, D. C., Metz, C. E., Nishikawa, R. M., 2005. The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks. IEEE Trans. Med. Imag. 24, 293–299.

Ferri, C., Hernández-Orallo, J., Salido, M. A., 2003. Volume under the roc surface for multi-class problems: Exact computation and evaluation of ap-

proximations. Tech. rep., Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València (Spain).
Hand, D. J., Till, R. J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45, 171–186.
He, X., Metz, C. E., Tsui, B. M. W., Links, J. M., Frey, E. C., 2006. Three-class ROC analysis — A decision theoretic approach under the ideal observer framework. IEEE Trans. Med. Imag. 25, 571–581.
Huo, Z., Giger, M. L., Metz, C. E., 1999. Effect of dominant features on neural network performance in the classification of mammographic lesions. Phys. Med. Biol. 44, 2579–2595.
Huo, Z., Giger, M. L., Vyborny, C. J., 2001. Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis. IEEE Trans. Med. Imag. 20, 1285–1292.
Huo, Z., Giger, M. L., Vyborny, C. J., Metz, C. E., 2002. Breast cancer: Effectiveness of computer-aided diagnosis — Observer study with independent database of mammograms. Radiology 224, 560–568.
Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., Metz, C. E., 2000. Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness. Acad. Radiol. 7, 1077–1084.
Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., Schmidt, R. A., Doi, K., 1998. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad. Radiol. 5, 155–168.
Kupinski, M. A., 2000. Computerized pattern classification in medical imaging. Ph.D. thesis, The University of Chicago, Chicago, IL.
Kupinski, M. A., Edwards, D. C., Giger, M. L., Metz, C. E., 2001. Ideal observer approximation using Bayesian classification neural networks. IEEE Trans. Med. Imag. 20, 886–899.
Lachiche, N., Flach, P., January 2003. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). AAAI Press, Washington, D.C., pp. 416–423.
Mossman, D., 1999. Three-way ROCs. Med. Decis. Making 19, 78–89.
Nakas, C. T., Yiannoutsos, C. T., 2004. Ordered multiple-class roc analysis with continuous measurements. Statist. Med. 23, 3437–3449.
Obuchowski, N. A., Applegate, K. E., Goske, M. J., Arheart, K. L., Myers, M. T., Morrison, S., 2001. The 'differential diagnosis' for multiple diseases: Comparison with the binary-truth state experiment in two empirical studies. Acad. Radiol. 8, 947–954.
Patel, A. C., Markey, M. K., 2005. Comparison of three-class classification performance metrics: a case study in breast cancer CAD. In: Miguel P. Eckstein, Yulei Jiang (Eds.), Proc. SPIE Vol. 5749 Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment. SPIE, Bellingham, WA, pp. 581–589.
Scurfield, B. K., 1996. Multiple-event forced-choice tasks in the theory of signal

596 detectability. J. Math. Psychol. 40, 253–269.
597 Scurfield, B. K., 1998. Generalization of the theory of signal detectability to
598 n-event m-dimensional forced-choice tasks. J. Math. Psychol. 42, 5–31.
599 Srinivasan, A., 1999. Note on the location of optimal classifiers in n-
600 dimensional ROC space. Tech. Rep. PRG-TR-2-99, Oxford University Com-
601 puting Laboratory, Wolfson Building, Parks Road, Oxford.
602 Van Trees, H. L., 1968. Detection, Estimation and Modulation Theory: Part
603 I. John Wiley & Sons, New York.
604 Yin, F.-F., Giger, M. L., Doi, K., Metz, C. E., Vyborny, C. J., Schmidt, R. A.,
605 1991. Computerized detection of masses in digital mammograms: Analysis
606 of bilateral subtraction images. Med. Phys. 18, 955–963.
607 Yin, F.-F., Giger, M. L., Doi, K., Vyborny, C. J., Schmidt, R. A., 1994. Com-
608 puterized detection of masses in digital mammograms: Automated align-
609 ment of breast images and its effect on bilateral-subtraction technique. Med.
610 Phys. 21, 445–452.
611 Yin, F.-F., Giger, M. L., Vyborny, C. J., Doi, K., Schmidt, R. A., 1993. Com-
612 parison of bilateral-subtraction and single-image processing techniques in
613 the computerized detection of mammographic masses. Invest. Radiol. 28,
614 473–481.

C Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities

Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities

Darrin C. Edwards* and Charles E. Metz

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

ABSTRACT

We have shown in previous work that an ideal observer in a classification task with N classes achieves the optimal receiver operating characteristic (ROC) hypersurface in a Neyman-Pearson sense. That is, the hypersurface obtained by taking one of the ideal observer's misclassification probabilities as a function of the other N^2-N-1 misclassification probabilities is never above the corresponding hypersurface obtained by any other observer. Due to the inherent complexity of evaluating observer performance in an N-class classification task with $N > 2$, some researchers have suggested a generally incomplete but more tractable evaluation in terms of a hypersurface plotting only the N "sensitivities" (the probabilities of correctly classifying observations in the various classes). An N-class observer generally has up to N^2-N-1 degrees of freedom, so a given sensitivity will still vary when the other $N-1$ are held fixed; a well-defined hypersurface can be constructed by considering only the maximum possible value of one sensitivity for each achievable value of the other $N-1$. We show that optimal performance in terms of this generally incomplete performance descriptor, in a Neyman-Pearson sense, is still achieved by the N-class ideal observer. That is, the hypersurface obtained by taking the maximal value of one of the ideal observer's correct classification probabilities as a function of the other $N-1$ is never below the corresponding hypersurface obtained by any other observer.

Keywords: ROC analysis, three-class classification, ideal observer decision rules

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 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 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 risk[4]) 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 date 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] More recently, we showed that complicated constraining relationships exist among the decision criteria themselves for the ideal observer.[8] These constraining relationships appear to imply that it is highly unlikely that analytical expressions for the conditional error probabilities in terms of the decision criteria can be developed which are as simple to interpret as those for the two-class task.[6]

Despite the difficulties just described, the potential benefits to be gained from a practical performance evaluation methodology for classification tasks with three classes have motivated a number of research groups to propose such methods. These practical methods reduce the number of degrees of freedom required to describe the observer's performance, either by implicitly leaving the remaining degrees of freedom out of the analysis, or

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

by explicitly imposing restrictions on the form of the observer's decision rule or on the set of decision criteria used by the observer.

Scurfield evaluated an observer which used a specified decision rule with only two degrees of freedom (as opposed to the five decision criteria used by the general three-class ideal observer) by plotting a set of six (two-dimensional) surfaces in three-dimensional ROC spaces.[9] Mossman proposed plotting the surface formed only from the set of three "sensitivities" (conditional probabilities of correctly classifying observations) for an observer with two degrees of freedom, and applied this method to an observer with a specified decision rule.[10] Chan *et al.* began with an ideal observer model, and reduced the number of decision criteria from five to two by imposing explicit assumptions on the observer's decision utilities; the observer's performance was then plotted as a surface in a three-dimensional ROC space, the axes of which are the probabilities of deciding an observation to be malignant conditional on each of the three actual class memberships.[11] He *et al.* investigated an ideal observer model in which the decision rule is restricted to a form similar to that proposed by Scurfield; the nature of the restrictions is such that performance evaluation in terms of only the three sensitivities provides a complete description of this observer's performance.[12]

A common theme among these remarkably diverse methods is the idea of an "ROC surface," *i.e.*, a surface with two degrees of freedom in a three-dimensional ROC space. An appealing feature of such a construct is its visualizability: it can be plotted as readily as any elevation map, for example, in stark contrast to the fully general three-class classification task involving a hypersurface with five degrees of freedom in a six-dimensional ROC space as mentioned above. While it is true that not all of the proposed methods described in the preceding paragraph involve a "sensitivity" ROC surface, the general division of an N-class observer's conditional decision probabilities into a set of N sensitivities and a set of $N^2 - N$ misclassification rates[5] makes this particular construct a natural candidate for further analysis.

On the other hand, it can be argued that measurement of performance in terms of only N conditional classification rates must be an incomplete description of observer performance in a classification task with more than two classes, which requires $N^2 - N$ such classification rates as stated above. Acknowledging this incompleteness, we would like to ask whether there is any sense in which such an incomplete performance metric is at least well-defined. In particular, is there any observer decision rule, dependent on only $N - 1$ (rather than $N^2 - N - 1$) decision criteria, for which the observer's sensitivity ROC hypersurface is always above the corresponding hypersurface obtained for any other observer? If so, what form does this decision rule take?

In the next section, we show that the three-class observer which optimizes performance only in terms of the sensitivity surface is in fact the three-class ideal observer, with its decision utilities constrained in a particular way (reducing its degrees of freedom from five to two as necessary). Additionally, the form of the constraints on the ideal observer's behavior are identical to those considered by He *et al.*.[12] In Sec. 3, we extend this result to the general case of an N-class observer, showing that the observer which attains the optimal sensitivity hypersurface is a restricted form of the N-class ideal observer, and in particular a straightforward generalization of the three-class observer considered by He *et al.*[12] to N classes. Our conclusions are stated in Sec. 4.

2. THREE-CLASS OBSERVERS

We have shown[5] that the N-class ideal observer — that observer which minimizes Bayes risk — also achieves optimal performance in an ROC sense, by virtue of satisfying the Neyman-Pearson criterion. This was the same argument used by Van Trees[4] to show that the two-class ideal observer achieves the optimal ROC curve for a given two-class classification task. This technique of satisfying the Neyman-Pearson criterion, essentially an application of an integral form of the method of Lagrange multipliers,[13] is straightforward (conceptually, if not notationally) and flexible, and we apply it in this section to answer the question of what observer optimizes performance in terms of only the three observer sensitivities.

We denote by P_{ij} the conditional probability of a given observer deciding an observation is drawn from the ith class, conditional on it actually being drawn from the jth class. Thus, the three sensitivities are P_{11}, P_{22}, and P_{33}. Decisions are assumed to be made based on statistically variable observational data; in particular,

$$P_{ij} \equiv \int_{Z_i} p(\vec{x}|\pi_j)\, d^m\vec{x}, \tag{1}$$

where Z_i is the region for which observations $\vec{\mathbf{x}}$ (of dimension m) are decided to belong to the class labeled π_i ($1 \leq i \leq 3$).

Without loss of generality, we seek to maximize P_{33} subject to the constraints $P_{11} = \alpha_{11}$ and $P_{22} = \alpha_{22}$ where $0 \leq \alpha_{11} \leq 1$ and $0 \leq \alpha_{22} \leq 1$. We define the function

$$F \equiv P_{33} + \lambda_{11}(P_{11} - \alpha_{11}), + \lambda_{22}(P_{22} - \alpha_{22}) \tag{2}$$

where λ_{11} and λ_{22} are the so-called Lagrange multipliers. Note that if we can find a decision rule (a partitioning of the domain of $\vec{\mathbf{x}}$ into Z_1, Z_2, and Z_3) that maximizes F for arbitrary values of λ_{11} and λ_{22}, then this will be equivalent to maximizing P_{33} at the point at which the constrain equations are satisfied (*i.e.*, at the point $P_{11} = \alpha_{11}, P_{22} = \alpha_{22}$).

We first rewrite F by applying rules for conditional probabilities:

$$\begin{aligned} F &= -\lambda_{11}\alpha_{11} - \lambda_{22}\alpha_{22} + (1 - P_{13} - P_{23}) + \lambda_{11}(1 - P_{21} - P_{31}) + \lambda_{22}(1 - P_{12} - P_{32}) \\ &= 1 + \lambda_{11}(1 - \alpha_{11}) + \lambda_{22}(1 - \alpha_{22}) - \{\lambda_{22}P_{12} + P_{13} + \lambda_{11}P_{21} + P_{23} + \lambda_{11}P_{31} + \lambda_{22}P_{32}\} \\ &= 1 + \lambda_{11}(1 - \alpha_{11}) + \lambda_{22}(1 - \alpha_{22}) - \left\{ \int_{Z_1} \lambda_{22} p(\vec{x}|\pi_2) + p(\vec{x}|\pi_3)\, d^m\vec{x} \right. \\ &\quad \left. + \int_{Z_2} \lambda_{11} p(\vec{x}|\pi_1) + p(\vec{x}|\pi_3)\, d^m\vec{x} + \int_{Z_3} \lambda_{11} p(\vec{x}|\pi_1) + \lambda_{22} p(\vec{x}|\pi_2)\, d^m\vec{x} \right\}. \end{aligned} \tag{3}$$

For a given set of values of the parameters λ_{11} and λ_{22}, F is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in Eq. 3) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad \lambda_{22} p(\vec{x}|\pi_2) < \lambda_{11} p(\vec{x}|\pi_1) \quad \text{and} \quad p(\vec{x}|\pi_3) < \lambda_{11} p(\vec{x}|\pi_1) \tag{4}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad \lambda_{11} p(\vec{x}|\pi_1) \leq \lambda_{22} p(\vec{x}|\pi_2) \quad \text{and} \quad p(\vec{x}|\pi_3) < \lambda_{22} p(\vec{x}|\pi_2) \tag{5}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad \lambda_{11} p(\vec{x}|\pi_1) \leq p(\vec{x}|\pi_3) \quad \text{and} \quad \lambda_{22} p(\vec{x}|\pi_2) \leq p(\vec{x}|\pi_3). \tag{6}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 > 0 \quad \text{and} \quad \lambda_{11}\text{LR}_1 > 1 \tag{7}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 \leq 0 \quad \text{and} \quad \lambda_{22}\text{LR}_2 > 1 \tag{8}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad \lambda_{11}\text{LR}_1 \leq 1 \quad \text{and} \quad \lambda_{22}\text{LR}_2 \leq 1, \tag{9}$$

where $\text{LR}_i \equiv p(\vec{x}|\pi_i)/p(\vec{x}|\pi_3)$ are the likelihood ratio decision variables used by the ideal observer.[4,5] The decision boundary lines which partition the $(\text{LR}_1, \text{LR}_2)$ decision plane into the regions Z_1, Z_2, and Z_3 are thus

$$\lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 = 0 \tag{10}$$

$$\lambda_{11}\text{LR}_1 = 1 \tag{11}$$

$$\lambda_{22}\text{LR}_2 = 1. \tag{12}$$

Note that Eq. 12 is just the difference between Eqs. 10 and 11. If we require λ_{11} and λ_{22} to be positive, the decision rule is an ideal observer decision rule.[5] Since neither the decision variables nor the form of the decision rule depend on the particular choices of α_{11} and α_{22}, we can conclude that the three-class sensitivity ROC surface, obtained by allowing λ_{11} and λ_{22} to take on all possible positive values, is optimal for the observer defined in Eqs. 10–12, in the sense that no other observer can achieve a higher sensitivity surface (*i.e.*, a surface with a greater value of P_{33} at a given value of (P_{11}, P_{22})). The optimal observer for this performance metric is seen to be the three-class ideal observer, with its decision criteria constrained so that the line separating classes π_1 and π_3 is vertical, the line separating classes π_2 and π_3 is horizontal, and the line separating classes π_1 and

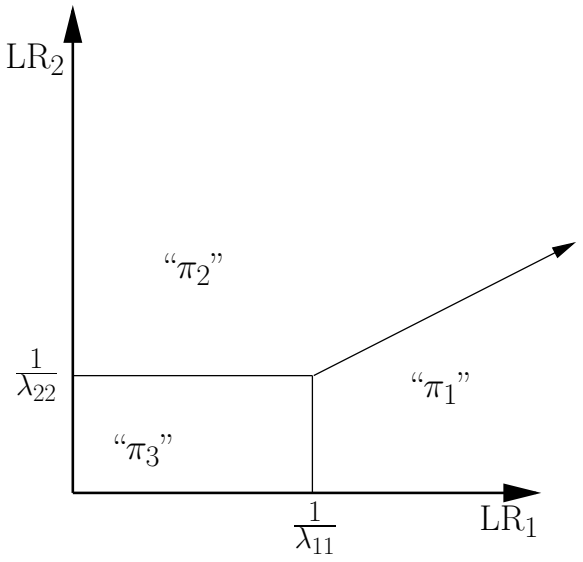

Figure 1. The decision rule which is found to be optimal in the sense of maximizing the ROC surface composed of only the observer sensitivities. The decision variables are the likelihood ratios used by the general three-class ideal observer, and the number of decision criteria is reduced from five (for the general three-class ideal observer) to two.

π_2 passes through the origin with slope $\lambda_{11}/\lambda_{22}$ (and thus intersects the other two lines as required). Note that the number of free decision criteria has been reduced from five (for the general three-class ideal observer) to two (as expected for a surface in a three-dimensional ROC space).

This decision rule is shown in Fig. 1. It is interesting to note that this observer is identical to the special case of the ideal observer evaluated by He *et al.*,[12] which we have shown[14,15] to be a special case of the decision rule proposed by Scurfield.[9]

3. *N*-CLASS OBSERVERS

The results of the preceding section can be generalized to tasks with N classes for any $N > 2$. We now have a set of N^2 conditional classification probabilities P_{ij}, with N sensitivities P_{ii}. Equation 1 remains unchanged, except that there are of course now N regions Z_i into which the domain of $\vec{x}$ is partitioned (*i.e.*, classes into which the observations are classified), and the observations are drawn from N distributions of the form $p(\vec{x}|\pi_j)$.

Without loss of generality, we seek to maximize P_{NN} subject to the constraints $P_{ii} = \alpha_{ii}$ for $1 \leq i \leq N-1$, where $0 \leq \alpha_{ii} \leq 1$. We define the function

$$F \equiv P_{NN} + \sum_{i=1}^{N-1} \lambda_{ii}(P_{ii} - \alpha_{ii}), \tag{13}$$

where the λ_{ii} are the Lagrange multipliers. Note that if we can find a decision rule (a partitioning of the domain of $\vec{\mathbf{x}}$ into Z_i $\{1 \leq i \leq N\}$) that maximizes F for arbitrary values of the λ_{ii}, then this will be equivalent to maximizing P_{NN} at the point at which the constrain equations are satisfied (*i.e.*, at the point $P_{ii} = \alpha_{ii}$ $\{1 \leq i \leq N-1\}$).

As in the preceding section, we rewrite F by applying rules for conditional probabilities to obtain:

$$F = -\sum_{i=1}^{N-1} \lambda_{ii}\alpha_{ii} + \left(1 - \sum_{i=1}^{N-1} P_{iN}\right) + \sum_{i=1}^{N-1} \lambda_{ii} \left(1 - \sum_{\substack{j=1 \\ j \neq i}}^{N} P_{ji}\right)$$

$$
\begin{aligned}
&= 1 + \sum_{i=1}^{N-1} \lambda_{ii}(1-\alpha_{ii}) - \left\{ \left[\sum_{i=1}^{N-1} \left(\sum_{\substack{j=1 \\ j\neq i}}^{N} \lambda_{jj} P_{ij} \right) + P_{iN} \right] + \left[\sum_{i=1}^{N-1} \lambda_{ii} P_{Ni} \right] \right\} \\
&= 1 + \sum_{i=2}^{N} \lambda_{ii}(1-\alpha_{ii}) \\
&\quad - \left\{ \sum_{i=1}^{N-1} \int_{Z_i} \left[\sum_{\substack{j=1 \\ j\neq i}}^{N} \lambda_{jj} p(\vec{x}|\pi_j) \right] + p(\vec{x}|\pi_N)\, d^m\vec{x} + \int_{Z_N} \sum_{i=1}^{N-1} \lambda_{ii} p(\vec{x}|\pi_i)\, d^m\vec{x} \right\}. \qquad (14)
\end{aligned}
$$

For a given set of values of the parameters λ_{ii} $\{1 \leq i \leq N-1\}$, F is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning choosing the regions Z_i such that a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in Eq. 14) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$
\begin{aligned}
\text{decide} \quad \pi_i\{i<N\} \quad \text{iff} \quad & \lambda_{jj} p(\vec{x}|\pi_j) < \lambda_{ii} p(\vec{x}|\pi_i) \qquad \{i<j<N\} \\
\text{and} \quad & p(\vec{x}|\pi_N) < \lambda_{ii} p(\vec{x}|\pi_i) \\
\text{and} \quad & \lambda_{jj} p(\vec{x}|\pi_j) \leq \lambda_{ii} p(\vec{x}|\pi_i) \qquad \{j<i<N\} \qquad (15) \\
\text{decide} \quad \pi_N \quad \text{iff} \quad & \lambda_{jj} p(\vec{x}|\pi_j) \leq p(\vec{x}|\pi_N) \qquad \{j<N\}. \qquad (16)
\end{aligned}
$$

We can divide these relations by $p(\vec{x}|\pi_N)$ to obtain

$$
\begin{aligned}
\text{decide} \quad \pi_i\{i<N\} \quad \text{iff} \quad & \lambda_{ii}\mathrm{LR}_i - \lambda_{jj}\mathrm{LR}_j > 0 \qquad \{i<j<N\} \\
& \text{and} \lambda_{ii}\mathrm{LR}_i > 1 \\
& \text{and} \lambda_{jj}\mathrm{LR}_j - \lambda_{ii}\mathrm{LR}_i \leq 0 \qquad \{j<i<N\} \qquad (17) \\
\text{decide} \quad \pi_N \quad \text{iff} \quad & \lambda_{jj}\mathrm{LR}_j \leq 1 \qquad \{j<N\}, \qquad (18)
\end{aligned}
$$

where $\mathrm{LR}_i \equiv p(\vec{x}|\pi_i)/p(\vec{x}|\pi_N)$ are the likelihood ratio decision variables used by the ideal observer.[4,5] The decision boundary hyperplanes which partition the $\vec{\mathrm{LR}} \equiv (\mathrm{LR}_1, \ldots, \mathrm{LR}_{N-1})$ decision space into the regions Z_i are thus

$$
\begin{aligned}
\lambda_{ii}\mathrm{LR}_i - \lambda_{jj}\mathrm{LR}_j &= 0 \qquad \{i<j<N\} \qquad (19) \\
\lambda_{ii}\mathrm{LR}_i &= 1 \qquad \{i<N\}. \qquad (20)
\end{aligned}
$$

Note that any of these equations, for example that defining part of the boundary between classes π_j and π_k , can be expressed as the difference of two other such equations (in this example, those defining boundaries between classes π_i and π_j, and between classes pi_i and π_k). If we require the λ_{ii} to be positive, the resulting decision rule is an ideal observer decision rule.[5] Since neither the decision variables nor the form of the decision rule depend on the particular choices of α_{ii}, we can conclude that the N-class sensitivity ROC hypersurface, obtained by allowing the λ_{ii} to take on all possible positive values, is optimal for the observer defined in Eqs. 19 and 20, in the sense that no other observer can achieve a higher sensitivity hypersurface (*i.e.*, one with a greater value of P_{NN} at a given value of $(P_{11}, \ldots, P_{(N-1)(N-1)})$). The optimal observer for this performance metric is seen to be the N-class ideal observer, with its decision criteria constrained so that the boundary separating classes π_i and π_N is a hyperplane defined by $\mathrm{LR}_i = 1/\lambda_{ii}$, while the boundary separating classes π_i and π_j is a hyperplane defined by $\lambda_{ii}\mathrm{LR}_i = \lambda_{jj}\mathrm{LR}_j$.

Although an intuitive geometric understanding of this decision rule is more elusive than in the three-class case, it is at least evident that the boundaries intersect as expected; that is, the boundary separating classes π_i and π_j intersects the boundary separating classes π_i and π_k, and also intersects the boundary separating

classes π_j and π_k. Note also that the number of free decision criteria has been reduced from $N^2 - N - 1$ (for the general N-class ideal observer) to $N - 1$ (as expected for a hypersurface in an N-dimensional ROC space). More importantly, comparison of Eqs. 19 and 20 with Eqs. 10–12 reveals this N-class observer to be an obvious extension from three to N classes of the observer described in the preceding section.

4. CONCLUSIONS

A fully general performance evaluation methodology for the three-class classification task has yet to be developed, a frustrating state of affairs given the great success and wide application of ROC analysis to two-class classification tasks. A primary reason for the difficulty in developing a fully general extension of ROC analysis to the three-class classification task is the rapid increase in the number of performance measurement variables and decision criteria necessary to characterize observer (in particular, ideal observer) performance. Specifically, the number of sensitivities or misclassification rates needed increases from two to six (and to $N^2 - N$ in the general case), while the number of decision criteria increases from a single decision variable threshold to a set of five mutually constrained[8] criteria (and to $N^2 - N - 1$ in the general case). In short, the complexity of the problem increases not linearly with the number of classes, but quadratically.

The motivation for the numerous proposed methods, outlined in Sec. 1, for evaluating the performance of a three-class classifier in terms of two-dimensional surfaces in three-dimensional ROC spaces (rather than the five-dimensional hypersurfaces in six-dimensional ROC spaces required by the theory) is thus quite clear. We currently lack a theoretical framework with which to judge the appropriateness of any of the proposed methods to any particular classification task. However, even if one chooses to adopt a performance evaluation metric known to provide an incomplete description of observer performance, it is still reasonable to ask what observer, if any, will achieve optimal performance with respect to that metric.

We have addressed that question in regard to measurement of an observer's performance in terms of only its sensitivities (the probabilities of correctly classifying the three, or in general N, classes of observations). Theoretically, this is clearly an incomplete measure of performance (another set of three, or in general $N^2 - 2N$, misclassification rates are necessary). Conceding this point, we consider it a nontrivial observation, derived in the preceding sections, that the observer which optimizes this limited performance metric is not one unrelated to the general ideal observer, nor an arcane special case of the ideal observer, but a special case of the ideal observer which is in a subjective sense quite simple, and which has been independently evaluated from very different perspectives by other researchers.[9,12] We find these results at once reassuring and encouraging, and hope that research into this thorny problem will continue to bear unexpected fruit.

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. D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.* **24**, pp. 1566–1573, 2005.
9. B. K. Scurfield, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psychol.* **42**, pp. 5–31, 1998.
10. D. Mossman, "Three-way ROCs," *Med. Decis. Making* **19**, pp. 78–89, 1999.
11. H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 567–578, (SPIE, Bellingham, WA), 2003.
12. X. He, C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis — I. A decision theoretic approach," *IEEE Trans. Med. Imag.*, 2005. (in review).
13. S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations: Second Edition*, Harcourt Brace Jovanovich, San Diego, CA, 1986.
14. D. C. Edwards and C. E. Metz, "Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, eds., pp. 128–137, (SPIE, Bellingham, WA), 2005.
15. 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.*, 2005. (in review).

D Optimization of restricted ROC surfaces in three-class classification tasks

Optimization of restricted ROC surfaces in three-class classification tasks

Darrin C. Edwards* and Charles E. Metz

Abstract

We have shown previously that an N-class ideal observer achieves the optimal receiver operating characteristic (ROC) hypersurface in a Neyman-Pearson sense. Due to the inherent complexity of evaluating observer performance even in a three-class classification task, some researchers have suggested a generally incomplete but more tractable evaluation in terms of a surface plotting only the three "sensitivities." More generally, one can evaluate observer performance with a single sensitivity or misclassification probability as a function of two linear combinations of sensitivities or misclassification probabilities. We consider four such formulations including the "sensitivity" surface. In each case we show that the optimal observer with respect to the given evaluation method is a special case of the ideal observer, with certain constraints placed on the ideal observer's decision utilities. Furthermore, we show that if these utility constraints are imposed on a general expression for expected utility, this quantity is found to depend only on those sensitivities and misclassification probabilities used to construct the ROC surface in question. That is, for the observer which maximizes performance with respect to the given restricted ROC surface, that ROC surface provides a complete description of the observer's performance in an expected-utility sense.

Index Terms

ROC analysis, three-class classification, ideal observer decision rules

This work was supported by the US Army Medical Research and Materiel Command under Grant W81XWH-04-1-0495 (D. C. Edwards, principal investigator).

*D. C. Edwards is with the Department of Radiology, the University of Chicago, Chicago, IL, USA (e-mail: d-edwards@uchicago.edu).

C. E. Metz is with the Department of Radiology, the University of Chicago, Chicago, IL, USA.

Optimization of restricted ROC surfaces in three-class classification tasks

I. 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 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 extension of ROC analysis 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 risk [4]) 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 date 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]. More recently, we showed that complicated constraining relationships exist among the decision criteria themselves for the ideal observer [8]. These constraining relationships appear to imply that it is highly unlikely that analytical expressions for the conditional error probabilities in terms of the decision criteria can be developed which are as simple to interpret as those for the two-class task [6].

Despite the difficulties just described, the potential benefits to be gained from a practical performance evaluation methodology for classification tasks with three classes have motivated a number of research groups to propose such methods. These practical methods reduce the number of degrees of freedom required to describe the observer's performance, either by implicitly leaving the remaining degrees of freedom out of the analysis, or by explicitly imposing restrictions

on the form of the observer's decision rule or on the set of decision criteria used by the observer.

Scurfield evaluated an observer which used a specified decision rule with only two degrees of freedom (in general a three-class observer can have up to five degrees of freedom) by plotting a set of six (two-dimensional) surfaces in three-dimensional ROC spaces [9]. Mossman proposed plotting the surface formed only from the set of three "sensitivities" (conditional probabilities of correctly classifying observations) for an observer with two degrees of freedom, and applied this method to an observer with a specified decision rule [10]. Chan *et al.* began with an ideal observer model, and reduced the number of decision criteria from five to two by imposing explicit assumptions on the observer's decision utilities; a description of the observer's performance (which they also showed to be complete) was then plotted as a surface in a three-dimensional ROC space, the axes of which are the probabilities of deciding an observation to be malignant conditional on each of the three actual class memberships [11]. He *et al.* investigated a special case of the ideal observer model which is also a special case of the decision rule proposed by Scurfield; they showed that due to the assumptions of their model, performance evaluation in terms of only the three sensitivities provides a complete description of this observer's performance [12].

A common theme among these remarkably diverse methods is the idea of an "ROC surface," *i.e.*, a surface with two degrees of freedom in a three-dimensional ROC space. An appealing feature of such a construct is its visualizability: it can be plotted as readily as any elevation map, for example, in stark contrast to the fully general three-class classification task involving a hypersurface with five degrees of freedom in a six-dimensional ROC space as mentioned above.

On the other hand, it can be argued that measurement of three-class classification performance in terms of only three conditional classification rates will yield an incomplete description of observer performance. (A complete description should require six such conditional classification rates as stated above.) Acknowledging this possible incompleteness, we would like to ask whether there is any sense in which such a restricted performance evaluation method is at least well-defined. In particular, suppose we elect to measure performance in terms of an ROC surface given by a single sensitivity or conditional error rate as a function of two different linear combinations of other sensitivities or conditional error rates). We then ask, is there any observer decision rule, dependent on only two (rather than five) decision criteria, for which the specified ROC surface is never below (when the surface's dependent variable is a sensitivity) or never above (when the

surface's dependent variable is a conditional error rate) the corresponding surface obtained for any other observer? If so, what form does this decision rule take?

In the remainder of this work, four different observer decision strategies proposed recently in the literature are analyzed with regard to the questions just posed. Each strategy considered is a special case of the three-class ideal observer, which classifies observations by maximizing the expected utility of its decisions. For each special case considered here, the expected utility is constrained to depend on only three (rather than six) conditional classification rates. We show, in each case, that the observer which maximizes performance, in a Neyman-Pearson sense [4], [5], is in fact the proposed special case of the ideal observer.

In Sec. II, we consider the decision rule proposed by Chan *et al.* [11]; in Sec. III, that proposed by He *et al.* [12], which is itself a special case (in which the decision variables used are the logarithms of the likelihood ratios of the data being classified) of the decision rule proposed by Scurfield [9]; and, in Secs. IV and V, two other special cases of the Scurfield decision rule, in which the decision variables are, respectively, the likelihood ratios and the *a posteriori* class membership probabilities of the data being classified. Finally, we summarize these results and present some brief conclusions in Sec. VI.

II. The Chan et al. Observer

The expected utility of the decisions made by an observer in an N-class classification task can be expressed as [5]

$$\begin{aligned} E\{\mathbf{U}\} &= \sum_{i=1}^{N}\sum_{j=1}^{N} U_{i|j} P(\mathbf{d}=\pi_i, \mathbf{t}=\pi_j) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} U_{i|j} P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j) P(\mathbf{t}=\pi_j), \end{aligned} \tag{1}$$

where the labels π_1 through π_N identify the classes to which observations belong; the number $U_{i|j}$ is defined as the utility of deciding an observation belongs to class π_i given that it is actually drawn from class π_j; and the random variables $\mathbf{t}$ and $\mathbf{d}$ indicate the true class to which a randomly drawn observation belongs and the observer's decision for classifying that observation, respectively. For notational simplicity, we will write the conditional classification rate $P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j)$ as P_{ij}, and the *a priori* class membership probability $P(\mathbf{t}=\pi_i)$ as $P(\pi_i)$.

For a three-class classification task, the expected utility can be written explicitly as

$$\begin{aligned} E\{\mathbf{U}\} = \quad & [U_{1|1}P_{11} + U_{2|1}P_{21} + U_{3|1}P_{31}]P(\pi_1) \\ + \quad & [U_{1|2}P_{12} + U_{2|2}P_{22} + U_{3|2}P_{32}]P(\pi_2) \\ + \quad & [U_{1|3}P_{13} + U_{2|3}P_{23} + U_{3|3}P_{33}]P(\pi_3). \end{aligned} \tag{2}$$

Note that the nine conditional classification rates P_{ij} appearing in this expression are not independent; for example, given the definition of conditional probability, it must be the case that $P_{11} + P_{21} + P_{31} = 1$. Thus within any pair of square brackets, one of the three conditional classification rates can be eliminated, leaving an expression which depends in general on six conditional classification rates.

Chan *et al.* consider a classification task in which class π_1 represents "benign," class π_2 "normal," and class π_3 "malignant" observations (*e. g.*, for structures evident in a medical image) [11]. They simplify the expression in (2) by restricting all values of utility to lie between 0 and 1; by setting the "correct decision" utilities $U_{1|1}$, $U_{2|2}$, and $U_{3|3}$ to be 1; the "missed malignancy" utilities $U_{1|3}$ and $U_{2|3}$ to be 0; and the utilities for incorrect decisions not involving malignancies $U_{1|2}$ and $U_{2|1}$ to be 1. The remaining "false-positive" utilities $U_{3|1}$ and $U_{3|2}$ are free to vary in the range $[0, 1]$.

With these assumptions, the expression for expected utility is reduced to

$$\begin{aligned} E\{\mathbf{U}_{\text{Chan}}\} = \quad & [P_{11} + P_{21} + U_{3|1}P_{31}]P(\pi_1) \\ + \quad & [P_{12} + P_{22} + U_{3|2}P_{32}]P(\pi_2) \\ + \quad & P_{33}P(\pi_3). \end{aligned} \tag{3}$$

This can in turn be simplified further using the definition of conditional probability to yield

$$\begin{aligned} E\{\mathbf{U}_{\text{Chan}}\} = \quad & [1 - P_{31} + U_{3|1}P_{31}]P(\pi_1) \\ + \quad & [1 - P_{32} + U_{3|2}P_{32}]P(\pi_2) \\ + \quad & P_{33}P(\pi_3); \end{aligned} \tag{4}$$

as Chan *et al.* point out [11], this expression depends on three rather than six conditional classification rates, namely $P_{3|1}$, $P_{3|2}$, and $P_{3|3}$. These three rates are used to construct the ROC space in which they analyze the performance of their observer. That observer in turn is the

special case of the ideal observer obtained by imposing the above constraints on the decision utilities $U_{i|j}$.

The three-class ideal observer makes decisions by partitioning a likelihood ratio decision variable plane into three regions with three intersecting lines [4], [5]. The likelihood ratios can be taken to be $\mathbf{LR}_1 \equiv p(\vec{\mathbf{x}}|\pi_1)/p(\vec{\mathbf{x}}|\pi_3)$ and $\mathbf{LR}_2 \equiv p(\vec{\mathbf{x}}|\pi_2)/p(\vec{\mathbf{x}}|\pi_3)$, ratios of the conditional probability density functions of the observational data $\vec{\mathbf{x}}$ taken as functions of that random observational data. (We use boldface type to denote statistically variable quantities.) In the notation we advocate [8], the equations for the three decision boundary lines are

$$\gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 = \gamma_{313} - \gamma_{323} \tag{5}$$

$$\gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 = \gamma_{313} \tag{6}$$

$$(\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 = \gamma_{323}, \tag{7}$$

which we call, respectively, the "1-*vs.*-2" line, the "1-*vs.*-3" line, and the "2-*vs.*-3" line. Here $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\pi_i)$. Although we have found it useful to assume these quantities to be strictly positive, this is not a fundamental requirement, and Chan *et al.* indeed allow some of them (*e. g.*, γ_{121}) to be zero (consistent with the constraints they place on the $U_{i|j}$ as described above). They obtain the resulting ideal observer decision lines

$$0\mathrm{LR}_1 - 0\mathrm{LR}_2 = 0 \qquad \{\text{"1-}vs\text{.-2"}\} \tag{8}$$

$$(1 - U_{3|1})P(\pi_1)\mathrm{LR}_1 + (1 - U_{3|2})P(\pi_2)\mathrm{LR}_2 = P(\pi_3) \qquad \{\text{"1-}vs\text{.-3"}\} \tag{9}$$

$$(1 - U_{3|1})P(\pi_1)\mathrm{LR}_1 + (1 - U_{3|2})P(\pi_2)\mathrm{LR}_2 = P(\pi_3) \qquad \{\text{"2-}vs\text{.-3"}\}, \tag{10}$$

which actually correspond to a single line (as the first is undefined and the remaining two are degenerate). This decision strategy is illustrated in Fig. 1.

In summary, Chan *et al.* begin with an ideal observer model, impose particular constraints on the decision utilities in that model, and then determine, based on those constraints, both the resulting form of the special case of the ideal observer and the conditional classification rates appropriate to measuring its performance. We now wish to pose a question from a different point of view: suppose one chooses to measure arbitrary (*i. e.*, not necessarily ideal) observer performance only in terms of the conditional classification rates P_{33}, P_{31}, and P_{32}, ignoring the other rates. For any observer, we can construct an ROC surface with P_{33} as a function of P_{31}

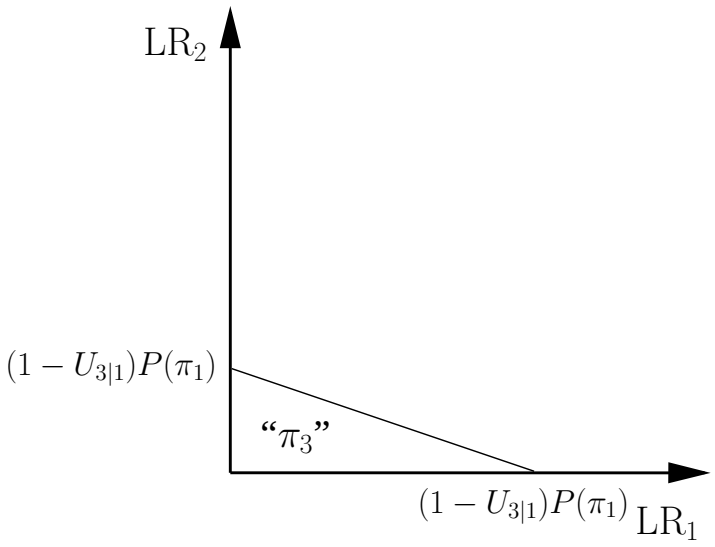

Fig. 1. The decision strategy investigated by Chan *et al.*, which is a special case of the ideal observer decision strategy. Observations in the unlabeled region are decided “not π_3,” *i. e.*, either “π_1” or “π_2”.

and P_{32}. (For an observer with more than two degrees of freedom in its decision strategy, one can simply define the surface to be the maximum value of P_{33} achievable at any given (P_{31}, P_{32}) pair.) What observer, if any, will achieve optimal performance with respect to this surface?

A convenient method for defining “optimal performance” here is in terms of the Neyman-Pearson criterion [4], [5]; the technique of satisfying the Neyman-Pearson criterion is essentially an application of an integral form of the method of Lagrange multipliers [13]. We seek to maximize P_{33} at a particular point $(P_{31} = \alpha_{31}, P_{32} = \alpha_{32})$ in the domain of the given ROC space. Another way of stating this is to consider P_{33}, P_{31}, and P_{32} as functionals of the observer's decision rule; we seek to maximize P_{33} subject to the constraints $P_{31} = \alpha_{31}$ and $P_{32} = \alpha_{32}$. To find this maximum, we define a function

$$F_{\text{Chan}} \equiv P_{33} + \lambda_{31}(P_{31} - \alpha_{31}) + \lambda_{32}(P_{32} - \alpha_{32}), \tag{11}$$

where λ_{31} and λ_{32} are free parameters (the so-called Lagrange multipliers). Note that maximizing F_{Chan} at the particular point $(P_{31} = \alpha_{31}, P_{32} = \alpha_{32})$ is equivalent to maximizing P_{33} at that point; if the maxima for arbitrary points (P_{31}, P_{32}) are achieved by a single decision rule independent of α_{31} and α_{32}, the resulting surface will be the desired optimal surface.

As stated in the material leading up to (5)–(7), the decisions here are assumed to be made

based on statistically variable observational data. Explicitly,

$$P_{ij} \equiv \int_{Z_i} p(\vec{x}|\pi_j)\, d^m\vec{x}, \tag{12}$$

where Z_i is the region for which observations $\vec{\mathbf{x}}$ (of dimension m) are decided to belong to the class labeled π_i ($1 \leq i \leq 3$). The expression for F_{Chan} can then be simplified as follows:

$$\begin{aligned} F_{\text{Chan}} &= 1 - P_{13} - P_{23} + \lambda_{31}P_{31} - \lambda_{31}\alpha_{31} + \lambda_{32}P_{32} - \lambda_{32}\alpha_{32} \\ &= 1 - \lambda_{31}\alpha_{31} - \lambda_{32}\alpha_{32} - \{P_{13} + P_{23} - \lambda_{31}P_{31} - \lambda_{32}P_{32}\} \\ &= 1 - \lambda_{31}\alpha_{31} - \lambda_{32}\alpha_{32} - \Big\{\int_{Z_1} p(\vec{x}|\pi_3)\, d^m\vec{x} + \int_{Z_2} p(\vec{x}|\pi_3)\, d^m\vec{x} \\ &\quad + \int_{Z_3} -\lambda_{31}p(\vec{x}|\pi_1) - \lambda_{32}p(\vec{x}|\pi_2)\, d^m\vec{x}\Big\}. \end{aligned} \tag{13}$$

F_{Chan} is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in (13)) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad p(\vec{x}|\pi_3) < p(\vec{x}|\pi_3) \quad \text{and} \quad p(\vec{x}|\pi_3) < -\lambda_{31}p(\vec{x}|\pi_1) - \lambda_{32}p(\vec{x}|\pi_2) \tag{14}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad p(\vec{x}|\pi_3) \leq p(\vec{x}|\pi_3) \quad \text{and} \quad p(\vec{x}|\pi_3) < -\lambda_{31}p(\vec{x}|\pi_1) - \lambda_{32}p(\vec{x}|\pi_2) \tag{15}$$

$$\begin{aligned} \text{decide} \quad \pi_3 \quad \text{iff} \quad & -\lambda_{31}p(\vec{x}|\pi_1) - \lambda_{32}p(\vec{x}|\pi_2) \leq p(\vec{x}|\pi_3) \\ & \text{and} \quad -\lambda_{31}p(\vec{x}|\pi_1) - \lambda_{32}p(\vec{x}|\pi_2) \leq p(\vec{x}|\pi_3). \end{aligned} \tag{16}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad 0\text{LR}_1 - 0\text{LR}_2 > 0 \quad \text{and} \quad -\lambda_{31}\text{LR}_1 - \lambda_{32}\text{LR}_2 > 1 \tag{17}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad 0\text{LR}_1 - 0\text{LR}_2 \leq 0 \quad \text{and} \quad -\lambda_{31}\text{LR}_1 - \lambda_{32}\text{LR}_2 > 1 \tag{18}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad -\lambda_{31}\text{LR}_1 - \lambda_{32}\text{LR}_2 \leq 1 \quad \text{and} \quad -\lambda_{31}\text{LR}_1 - \lambda_{32}\text{LR}_2 \leq 1. \tag{19}$$

(We assume without loss of generality that $p(\vec{x}|\pi_3) > 0$, because the task reduces to a two-class problem for values of $\vec{x}$ such that $p(\vec{x}|\pi_3) = 0$.) The boundary lines which partition the

$(\mathbf{LR}_1, \mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are thus

$$0\mathrm{LR}_1 - 0\mathrm{LR}_2 = 0 \qquad \{\text{“1-}\textit{vs.}\text{-2”}\} \tag{20}$$

$$-\lambda_{31}\mathrm{LR}_1 - \lambda_{32}\mathrm{LR}_2 = 1 \qquad \{\text{“1-}\textit{vs.}\text{-3”}\} \tag{21}$$

$$-\lambda_{31}\mathrm{LR}_1 - \lambda_{32}\mathrm{LR}_2 = 1 \qquad \{\text{“2-}\textit{vs.}\text{-3”}\}. \tag{22}$$

If we require λ_{31} and λ_{32} to be nonpositive, and then define the quantities $U_{3|1}$ and $U_{3|2}$ such that $-\lambda_{31} = (1 - U_{3|1})P(\pi_1)/P(\pi_3)$ and $-\lambda_{32} = (1 - U_{3|2})P(\pi_2)/P(\pi_3)$, the resulting decision strategy is found to be identical to that stated in (8)–(10). The special case of the ideal observer proposed by Chan *et al.*, whose performance depends only on the conditional classification rates P_{33}, P_{31}, and P_{32} by (4), is indeed the observer which obtains optimal performance with respect to this set of conditional classification rates.

III. The He et al. Observer

He *et al.* also begin with an ideal observer model and thus with the expression for expected utility given in (2); the classification task of interest to them is to distinguish two types of abnormal cardiac ejection from normal cardiac behavior in nuclear medicine studies [12]. They simplify this expression by requiring that the two possible incorrect classifications of observations actually from a given class be equal. That is, $U_{2|1} = U_{3|1}$, $U_{1|2} = U_{3|2}$, and $U_{1|3} = U_{2|3}$. The expression for expected utility is thereby reduced to

$$\begin{aligned} E\{\mathbf{U}_{\mathrm{He}}\} = \quad & [U_{1|1}P_{11} + U_{2|1}(P_{21} + P_{31})]P(\pi_1) \\ + \quad & [U_{2|2}P_{22} + U_{1|2}(P_{12} + P_{32})]P(\pi_2) \\ + \quad & [U_{3|3}P_{33} + U_{1|3}(P_{13} + P_{23})]P(\pi_3). \end{aligned} \tag{23}$$

This can in turn be simplified further using the definition of conditional probability to yield

$$\begin{aligned} E\{\mathbf{U}_{\mathrm{He}}\} = \quad & [U_{2|1} + (U_{1|1} - U_{2|1})P_{11}]P(\pi_1) \\ + \quad & [U_{1|2} + (U_{2|2} - U_{1|2})P_{22}]P(\pi_2) \\ + \quad & [U_{1|3} + (U_{3|3} - U_{1|3})P_{33}]P(\pi_3); \end{aligned} \tag{24}$$

as He *et al.* point out [12], this expression depends on only the three “sensitivities” P_{11}, P_{22}, and P_{33}, rather than six conditional classification rates. The three sensitivities are used to construct the

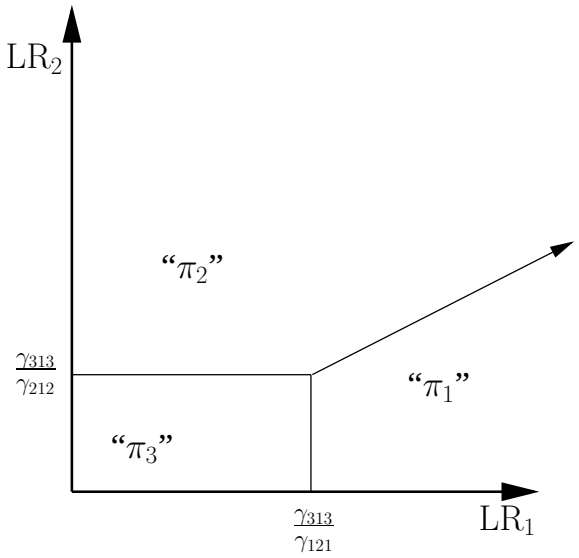

Fig. 2. The decision strategy investigated by He *et al.*, which is a special case of the ideal observer decision strategy, and which can also be shown to be a special case of the Scurfield observer with decision variables equal to the logarithms of the likelihood ratios of the observational data.

ROC space (equivalent to that proposed by Mossman [10]) in which they analyze the performance of their observer. That observer in turn is the special case of the ideal observer obtained by imposing the above constraints on the decision utilities $U_{i|j}$.

Applying the stated constrains on the utilities to the ideal observer decision boundary lines given in (5)–(7) yields

$$\gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 = 0 \tag{25}$$

$$\gamma_{121}\mathrm{LR}_1 = \gamma_{313} \tag{26}$$

$$\gamma_{212}\mathrm{LR}_2 = \gamma_{313}. \tag{27}$$

This decision strategy is illustrated in Fig. 2. We have recently shown [14] that this decision strategy is a special case of that proposed by Scurfield [9] when the decision variables used by the Scurfield observer are the logarithms of the likelihood ratios of the observational data.

We now consider evaluating the performance of an arbitrary observer in the ROC space constructed only from the observer's sensitivities (*i.e.*, P_{11}, P_{22}, and P_{33}). Without loss of generality, we can define such an observer's ROC surface as P_{33} considered as a function of P_{11} and P_{22}; to find the optimal observer with respect to this restricted performance evaluation

method, we apply the Neyman-Pearson criterion to maximize P_{33} subject to the constraints $(P_{11} = \alpha_{11}, P_{22} = \alpha_{22})$. We define the function

$$F_{\text{He}} \equiv P_{33} + \lambda_{11}(P_{11} - \alpha_{11}) + \lambda_{22}(P_{22} - \alpha_{22}), \tag{28}$$

where λ_{11} and λ_{22} are again the Lagrange multipliers.

Using (12), this can be simplified to yield

$$\begin{aligned} F_{\text{He}} &= 1 - P_{13} - P_{23} + \lambda_{11}(1 - P_{21} - P_{31}) - \lambda_{11}\alpha_{11} + \lambda_{22}(1 - P_{12} - P_{32}) - \lambda_{22}\alpha_{22} \\ &= 1 - \lambda_{11}\alpha_{11} - \lambda_{22}\alpha_{22} - \{P_{13} + P_{23} + \lambda_{11}(P_{21} + P_{31}) + \lambda_{22}(P_{12} + P_{32})\} \\ &= 1 - \lambda_{11}\alpha_{11} - \lambda_{22}\alpha_{22} - \Big\{\int_{Z_1} \lambda_{22}p(\vec{x}|\pi_2) + p(\vec{x}|\pi_3)\, d^m\vec{x} \\ &\quad + \int_{Z_2} \lambda_{11}p(\vec{x}|\pi_1) + p(\vec{x}|\pi_3)\, d^m\vec{x} + \int_{Z_3} \lambda_{11}p(\vec{x}|\pi_1) + \lambda_{22}p(\vec{x}|\pi_2)\, d^m\vec{x}\Big\}. \end{aligned} \tag{29}$$

F_{He} is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in (29)) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad \lambda_{22}p(\vec{x}|\pi_2) < \lambda_{11}p(\vec{x}|\pi_1) \quad \text{and} \quad p(\vec{x}|\pi_3) < \lambda_{11}p(\vec{x}|\pi_1) \tag{30}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad \lambda_{11}p(\vec{x}|\pi_1) \leq \lambda_{22}p(\vec{x}|\pi_2) \quad \text{and} \quad p(\vec{x}|\pi_3) < \lambda_{22}p(\vec{x}|\pi_2) \tag{31}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad \lambda_{11}p(\vec{x}|\pi_1) \leq p(\vec{x}|\pi_3) \quad \text{and} \quad \lambda_{22}p(\vec{x}|\pi_2) \leq p(\vec{x}|\pi_3). \tag{32}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 > 0 \quad \text{and} \quad \lambda_{11}\text{LR}_1 > 1 \tag{33}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 \leq 0 \quad \text{and} \quad \lambda_{22}\text{LR}_2 > 1 \tag{34}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad \lambda_{11}\text{LR}_1 \leq 1 \quad \text{and} \quad \lambda_{22}\text{LR}_2 \leq 1. \tag{35}$$

The boundary lines which partition the $(\mathbf{LR}_1, \mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are thus

$$\lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 = 0 \qquad \{\text{“1-}\textit{vs.}\text{-2”}\} \tag{36}$$

$$\lambda_{11}\text{LR}_1 = 1 \qquad \{\text{“1-}\textit{vs.}\text{-3”}\} \tag{37}$$

$$\lambda_{22}\text{LR}_2 = 1 \qquad \{\text{“2-}\textit{vs.}\text{-3”}\}. \tag{38}$$

If we require λ_{11} and λ_{22} to be positive, and define the quantities $\gamma_{121} \equiv \lambda_{11}\gamma_{313}$ and $\gamma_{212} \equiv \lambda_{22}\gamma_{313}$ for some arbitrary positive γ_{313}, then the resulting decision strategy is found to be identical to that stated in (25)–(27). The special case of the ideal observer proposed by He *et al.*, whose performance depends only on the conditional classification rates P_{11}, P_{22}, and P_{33} by (24), is indeed the observer which obtains optimal performance with respect to this set of conditional classification rates.

IV. The Scurfield Observer (Likelihood Ratio)

In the preceding two sections, we considered decision strategies that have been proposed by other researchers as special cases of the three-class ideal observer decision strategy. That is, particular constraints were explicitly imposed in the work cited on the decision utilities used by the ideal observer. The remaining two decision strategies we consider in the present work are special cases of a decision strategy proposed by Scurfield [9] which was not claimed to be generally related to the ideal observer; specifically, Scurfield specified the decision boundary lines used by the observer, but made no assumptions concerning the observer's two decision variables.

We showed recently [14] that if particular forms of the observer's decision variables related to the likelihood ratios of the observational data are chosen, then the resulting decision strategies can be shown to be special cases of the ideal observer decision strategy. One such special case is the observer analyzed by He *et al.* [12], discussed in Sec. III, in which the decision variables used by the Scurfield observer are the logarithms of the likelihood ratios. Two other such special cases are the Scurfield observer with the likelihood ratios themselves as decision variables, which we consider in this section; and that with the *a posteriori* class membership probabilities used as decision variables, considered in Sec. V. A minor difference from the preceding two sections is that we must determine the the implicit constraints on the ideal observer's utilities from the known form of the decision rule, rather than the other way around.

The general Scurfield observer makes decisions by partitioning a decision variable plane $(\mathbf{y}_1, \mathbf{y}_2)$ into three regions *via* the decision boundary lines

$$y_1 - y_2 = \gamma_1 - \gamma_2 \tag{39}$$

$$y_1 = \gamma_1 \tag{40}$$

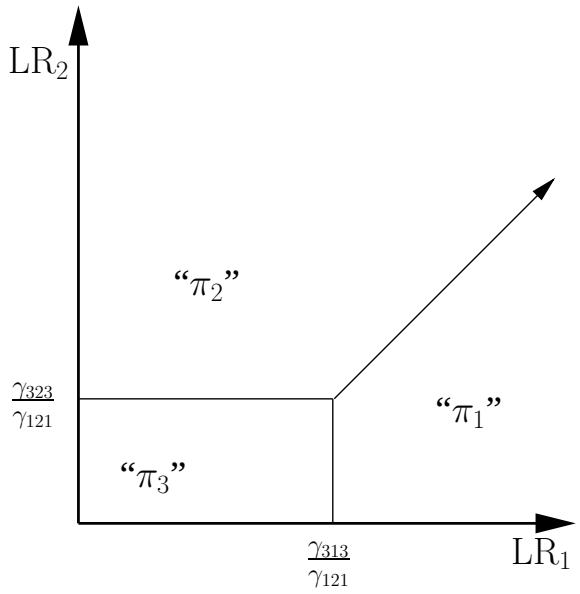

Fig. 3. A special case of the decision strategy investigated by Scurfield, in which the decision variables used are the likelihood ratios $(\mathbf{LR}_1, \mathbf{LR}_2)$ of the observational data.

$$y_2 \;=\; \gamma_2, \tag{41}$$

where γ_1 and γ_2 are parameters upon which the observer's performance depends (roughly equivalent to the decision criterion of a two-class classifier). When the decision variables are themselves the likelihood ratios $(\mathbf{LR}_1, \mathbf{LR}_2)$, this becomes in our notation

$$\mathrm{LR}_1 - \mathrm{LR}_2 \;=\; \frac{\gamma_{313} - \gamma_{323}}{\gamma_{121}} \tag{42}$$

$$\mathrm{LR}_1 \;=\; \frac{\gamma_{313}}{\gamma_{121}} \tag{43}$$

$$\mathrm{LR}_2 \;=\; \frac{\gamma_{323}}{\gamma_{121}}. \tag{44}$$

(Compare (39)–(41) with (5)–(7), and note that in order for the "1-*vs.*-2" line to have unit slope, it must be the case that $\gamma_{121} = \gamma_{212}$.) This decision strategy is illustrated in Fig. 3.

The relations $\gamma_{121} = \gamma_{131}$ and $\gamma_{212} = \gamma_{232}$ evident from the above equations immediately give the constraints on the decision utilities $U_{2|1} = U_{3|1}$ and $U_{1|2} = U_{3|2}$. Furthermore, the relation $\gamma_{121} = \gamma_{212}$ gives $(U_{1|1} - U_{2|1})P(\pi_1) = (U_{2|2} - U_{1|2})P(\pi_2)$. (Recall from Sec. II that $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\pi_i)$.) This allows us to simplify the expression for expected utility in (2)

to yield

$$\begin{aligned} E\{\mathbf{U}_{\text{Scurfield:LR}}\} = & \quad [U_{1|1}P_{11} + U_{2|1}(P_{21} + P_{31})]P(\pi_1) \\ & + \quad [U_{2|2}P_{22} + U_{1|2}(P_{12} + P_{32})]P(\pi_2) \\ & + \quad [U_{1|3}P_{13} + U_{2|3}P_{23} + U_{3|3}P_{33}]P(\pi_3). \end{aligned} \tag{45}$$

This can in turn be simplified further using the definition of conditional probability to yield

$$\begin{aligned} E\{\mathbf{U}_{\text{Scurfield:LR}}\} = & \quad [U_{1|1}P_{11} + U_{2|1}(1 - P_{11})]P(\pi_1) \\ & + \quad [U_{2|2}P_{22} + U_{1|2}(1 - P_{22})]P(\pi_2) \\ & + \quad [U_{1|3}P_{13} + U_{2|3}P_{23} + U_{3|3}(1 - P_{13} - P_{23})]P(\pi_3) \\ = & \quad [U_{2|1} + (U_{1|1} - U_{2|1})P_{11}]P(\pi_1) \\ & + \quad [U_{1|2} + (U_{2|2} - U_{1|2})P_{22}]P(\pi_2) \\ & + \quad [U_{3|3} + (U_{1|3} - U_{3|3})P_{13} + (U_{2|3} - U_{3|3})P_{23}] \\ = & \quad U_{2|1}P(\pi_1) + U_{1|2}P(\pi_2) + U_{3|3}P(\pi_3) \\ & + \quad (P_{11} + P_{22})(U_{1|1} - U_{2|1})P(\pi_1) \\ & + \quad [P_{13}(U_{1|3} - U_{3|3}) + P_{23}(U_{2|3} - U_{3|3})]P(\pi_3). \end{aligned} \tag{46}$$

This expression for the observer's expected utility depends on only three terms related to conditional classification rates: P_{13} and P_{23}, which may be regarded as the misclassification rates for observations actually drawn from class π_3; and $P_{11} + P_{22}$, which may be regarded as the "total sensitivity" for observations actually drawn from classes π_1 and π_2 (ignoring the *a priori* rates for such observations).

We now consider evaluating the performance of an arbitrary observer in an ROC-like space constructed from the quantities $P_{11} + P_{22}$, P_{13}, and P_{23}. We will define the ROC-like surface used to evaluate observer performance as the first quantity considered as a function of the two misclassification rates. To find the optimal observer with respect to this restricted performance evaluation method, we apply the Neyman-Pearson criterion to maximize $P_{11} + P_{22}$ subject to the constraints $(P_{13} = \alpha_{13}, P_{23} = \alpha_{23})$. We define the function

$$F_{\text{Scurfield:LR}} \equiv P_{11} + P_{22} + \lambda_{13}(P_{13} - \alpha_{13}) + \lambda_{23}(P_{23} - \alpha_{23}), \tag{47}$$

where λ_{13} and λ_{23} are the Lagrange multipliers.

Using (12), this can be simplified to yield

$$\begin{aligned} F_{\text{Scurfield:LR}} &= 1 - P_{21} - P_{31} + 1 - P_{12} - P_{32} + \lambda_{13}P_{13} - \lambda_{13}\alpha_{13} + \lambda_{23}P_{23} - \lambda_{23}\alpha_{23} \\ &= 2 - \lambda_{13}\alpha_{13} - \lambda_{23}\alpha_{23} - \{P_{21} + P_{31} + P_{12} + P_{32} - \lambda_{13}P_{13} - \lambda_{23}P_{23}\} \\ &= 2 - \lambda_{13}\alpha_{13} - \lambda_{23}\alpha_{23} - \Big\{ \int_{Z_1} p(\vec{x}|\pi_2) - \lambda_{13}p(\vec{x}|\pi_3)\, d^m\vec{x} \\ &\quad + \int_{Z_2} p(\vec{x}|\pi_1) - \lambda_{23}p(\vec{x}|\pi_3)\, d^m\vec{x} + \int_{Z_3} p(\vec{x}|\pi_1) + p(\vec{x}|\pi_2)\, d^m\vec{x} \Big\}. \end{aligned} \tag{48}$$

$F_{\text{Scurfield:LR}}$ is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in (48)) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{aligned} \text{decide} \quad \pi_1 \quad \text{iff} \quad & p(\vec{x}|\pi_2) - \lambda_{13}p(\vec{x}|\pi_3) < p(\vec{x}|\pi_1) - \lambda_{23}p(\vec{x}|\pi_3) \\ & \text{and} \quad -\lambda_{13}p(\vec{x}|\pi_3) < p(\vec{x}|\pi_1) \end{aligned} \tag{49}$$

$$\begin{aligned} \text{decide} \quad \pi_2 \quad \text{iff} \quad & p(\vec{x}|\pi_1) - \lambda_{23}p(\vec{x}|\pi_3) \le p(\vec{x}|\pi_2) - \lambda_{13}p(\vec{x}|\pi_3) \\ & \text{and} \quad -\lambda_{23}p(\vec{x}|\pi_3) < p(\vec{x}|\pi_2) \end{aligned} \tag{50}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad p(\vec{x}|\pi_1) \le -\lambda_{13}p(\vec{x}|\pi_3) \quad \text{and} \quad p(\vec{x}|\pi_2) \le -\lambda_{23}p(\vec{x}|\pi_3). \tag{51}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\text{decide} \quad \pi_1 \quad \text{iff} \quad \text{LR}_1 - \text{LR}_2 > -\lambda_{13} + \lambda_{23} \quad \text{and} \quad \text{LR}_1 > -\lambda_{13} \tag{52}$$

$$\text{decide} \quad \pi_2 \quad \text{iff} \quad \text{LR}_1 - \text{LR}_2 \le -\lambda_{13} + \lambda_{23} \quad \text{and} \quad \text{LR}_2 > -\lambda_{23} \tag{53}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad \text{LR}_1 \le -\lambda_{13} \quad \text{and} \quad \text{LR}_2 \le \lambda_{23}. \tag{54}$$

The boundary lines which partition the $(\mathbf{LR}_1, \mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are thus

$$\text{LR}_1 - \text{LR}_2 = -\lambda_{13} + \lambda_{23} \qquad \{\text{“1-}\textit{vs.}\text{-2”}\} \tag{55}$$

$$\text{LR}_1 = -\lambda_{13} \qquad \{\text{“1-}\textit{vs.}\text{-3”}\} \tag{56}$$

$$\text{LR}_2 = -\lambda_{23} \qquad \{\text{“2-}\textit{vs.}\text{-3”}\}. \tag{57}$$

If we require λ_{13} and λ_{23} to be negative, and define the quantities $\gamma_{313} \equiv -\lambda_{13}\gamma_{121}$ and $\gamma_{323} \equiv -\lambda_{23}\gamma_{121}$ for some arbitrary positive γ_{121}, then the resulting decision strategy is found to be identical to that stated in (42)–(44). This special case of the observer proposed by Scurfield, which we have shown to be a special case of the ideal observer [14], has a performance that depends only on the quantities $P_{11} + P_{22}$, P_{13}, and P_{23} by (46). This is indeed the observer which obtains optimal performance with respect to this set of quantities related to the conditional classification rates.

V. The Scurfield Observer (a posteriori Class Probability)

Equations (39)–(41) in Sec. IV give the equations for the decision boundary lines of the general Scurfield observer. If we now use two of the *a posteriori* class membership probabilities, such as $P(\pi_1|\vec{\mathbf{x}})$ and $P(\pi_2|\vec{\mathbf{x}})$, as the decision variables, the equations become

$$P(\pi_1|\vec{x}) - P(\pi_2|\vec{x}) = \gamma_1 - \gamma_2 \tag{58}$$

$$P(\pi_1|\vec{x}) = \gamma_1 \tag{59}$$

$$P(\pi_2|\vec{x}) = \gamma_2, \tag{60}$$

with $0 \leq \gamma_1 \leq 1$ and $0 \leq \gamma_2 \leq 1$. (Note that $P(\pi_3|\vec{x}) = 1 - P(\pi_1|\vec{x}) - P(\pi_2|\vec{x})$, meaning this third probability is not needed as an independent decision variable; the particular choice of which two probabilities to use is of course arbitrary.) This decision strategy, which we have shown recently to be a special case of the ideal observer decision strategy [14], is illustrated in Fig. 4.

We can reexpress the above equations in terms of likelihood ratios by exploiting the relation

$$\begin{aligned} P(\pi_i|\vec{x}) &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{p(\vec{x})} \\ &= \frac{\mathrm{LR}_i[P(\pi_i)/P(\pi_3)]}{1 + \mathrm{LR}_1[P(\pi_1)/P(\pi_3)] + \mathrm{LR}_2[P(\pi_2)/P(\pi_3)]}, \end{aligned} \tag{61}$$

where the second equation is obtained by dividing the numerator and denominator by $p(\vec{x}|\pi_3)P(\pi_3)$.

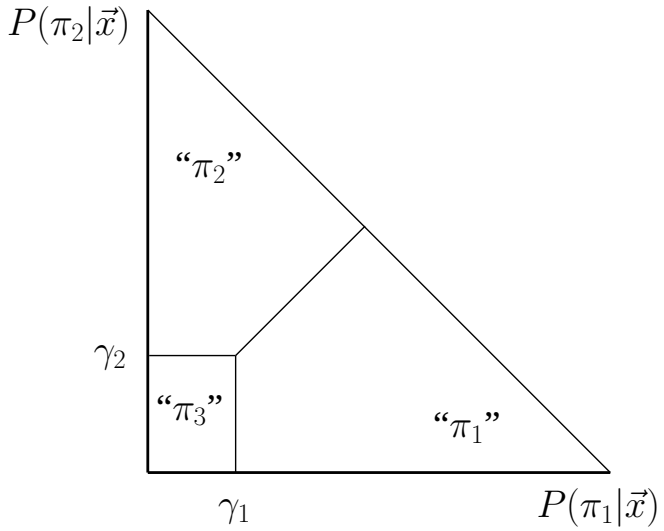

Fig. 4. A special case of the decision strategy investigated by Scurfield, in which the decision variables used are the *a posteriori* class membership probabilities $P(\pi_1|\vec{\mathbf{x}})$ and $P(\pi_2|\vec{\mathbf{x}})$ of the observational data.

The equations for the decision boundary lines become

$$\mathrm{LR}_1\frac{P(\pi_1)}{P(\pi_3)} - \mathrm{LR}_2\frac{P(\pi_2)}{P(\pi_3)} = (\gamma_1-\gamma_2)\left(1+\mathrm{LR}_1\frac{P(\pi_1)}{P(\pi_3)}+\mathrm{LR}_2\frac{P(\pi_2)}{P(\pi_3)}\right) \quad (62)$$

$$\mathrm{LR}_1\frac{P(\pi_1)}{P(\pi_3)} = \gamma_1\left(1+\mathrm{LR}_1\frac{P(\pi_1)}{P(\pi_3)}+\mathrm{LR}_2\frac{P(\pi_2)}{P(\pi_3)}\right) \quad (63)$$

$$\mathrm{LR}_2\frac{P(\pi_2)}{P(\pi_3)} = \gamma_2\left(1+\mathrm{LR}_1\frac{P(\pi_1)}{P(\pi_3)}+\mathrm{LR}_2\frac{P(\pi_2)}{P(\pi_3)}\right), \quad (64)$$

which can in turn be simplified to yield

$$[1-(\gamma_1-\gamma_2)]P(\pi_1)\mathrm{LR}_1 - [1+(\gamma_1-\gamma_2)]P(\pi_2)\mathrm{LR}_2 = (\gamma_1-\gamma_2)P(\pi_3) \quad (65)$$

$$(1-\gamma_1)P(\pi_1)\mathrm{LR}_1 - \gamma_1 P(\pi_2)\mathrm{LR}_2 = \gamma_1 P(\pi_3) \quad (66)$$

$$-\gamma_2 P(\pi_1)\mathrm{LR}_1 + (1-\gamma_2)P(\pi_2)\mathrm{LR}_2 = \gamma_2 P(\pi_3). \quad (67)$$

Although the above equations for the decision boundary lines are notably more complicated than those of the previous three sections, we can still relate the parameters γ_1 and γ_2 to the decision rule parameters of (5)–(7) to obtain constraints on the utilities $U_{i|j}$. For example,

comparison of (66) with (6) gives

$$\begin{aligned} \gamma_{232} - \gamma_{212} &= -\gamma_1 P(\pi_2) \\ U_{1|2} - U_{3|2} &= -\gamma_1, \qquad (68) \\ \gamma_{313} &= \gamma_1 P(\pi_3) \\ U_{3|3} - U_{1|3} &= \gamma_1. \qquad (69) \end{aligned}$$

This immediately gives the constraint

$$-(U_{1|2} - U_{3|2}) = U_{3|3} - U_{1|3}. \qquad (70)$$

Similarly, comparison of (67) and (7) gives

$$\begin{aligned} \gamma_{131} - \gamma_{121} &= -\gamma_2 P(\pi_1) \\ U_{2|1} - U_{3|1} &= -\gamma_2, \qquad (71) \\ \gamma_{323} &= \gamma_2 P(\pi_3) \\ U_{3|3} - U_{2|3} &= \gamma_2, \qquad (72) \end{aligned}$$

yielding the constraint

$$-(U_{2|1} - U_{3|1}) = U_{3|3} - U_{2|3}. \qquad (73)$$

Finally, we add the first two coefficient of (65) and then compare with (5) to obtain

$$\begin{aligned} [1 - (\gamma_1 - \gamma_2)] - [1 + (\gamma_1 - \gamma_2)] &= -2(\gamma_1 - \gamma_2) \\ (U_{1|1} - U_{2|1}) - (U_{2|2} - U_{1|2}) &= -2(U_{2|3} - U_{1|3}). \qquad (74) \end{aligned}$$

(On the right hand side of the above equation, we have made use of (69) and (72).) Note that the remaining terms in (65)–(67) involving γ_1 or γ_2 are simply differences of terms already considered, and would thus yield no further constraints on the utilities.

We can now impose constraints (70), (73), and (74) on the general expression (2) for expected

utility to obtain the expected utility for this observer:

$$\begin{aligned}
E\{\mathbf{U}_{\text{Scurfield:AP}}\} = \quad & [U_{1|1}P_{11} + U_{2|1}(1 - P_{11} - P_{31}) + U_{3|1}P_{31}]P(\pi_1) \\
+ \quad & [U_{1|2}(1 - P_{22} - P_{32}) + U_{2|2}P_{22} + U_{3|2}P_{32}]P(\pi_2) \\
+ \quad & [U_{1|3}P_{13} + U_{2|3}P_{23} + U_{3|3}(1 - P_{13} - P_{23})]P(\pi_3) \\
= \quad & [(U_{1|1} - U_{2|1})P_{11} - (U_{2|1} - U_{3|1})P_{31} + U_{2|1}]P(\pi_1) \\
+ \quad & [(U_{2|2} - U_{1|2})P_{22} - (U_{1|2} - U_{3|2})P_{32} + U_{1|2}]P(\pi_2) \\
+ \quad & [-(U_{3|3} - U_{1|3})P_{13} - (U_{3|3} - U_{2|3})P_{23} + U_{3|3}]P(\pi_3) \\
= \quad & \{(U_{1|1} - U_{2|1})P_{11} + (U_{3|3} - U_{2|3})P_{31} + U_{2|1}\}P(\pi_1) \\
+ \quad & \{[(U_{1|1} - U_{2|1}) + 2(U_{2|3} - U_{1|3})]P_{22}(U_{3|3} - U_{1|3})P_{32} + U_{1|2}\}P(\pi_2) \\
+ \quad & \{-(U_{3|3} - U_{1|3})P_{13} - (U_{3|3} - U_{2|3})P_{23} + U_{3|3}]P(\pi_3) \\
= \quad & U_{2|1}P(\pi_1) + U_{1|2}P(\pi_2) + U_{3|3}P(\pi_3) \\
+ \quad & (U_{1|1} - U_{2|1})[P(\pi_1)P_{11} + P(\pi_2)P_{22}] \\
+ \quad & (U_{3|3} - U_{1|3})[P(\pi_2)P_{32} + 2P(\pi_2)P_{22} - P(\pi_3)P_{13}] \\
+ \quad & (U_{3|3} - U_{2|3})[P(\pi_1)P_{31} - 2P(\pi_2)P_{22} - P(\pi_3)P_{23}]. \qquad (75)
\end{aligned}$$

As was the case for the decision strategies of the preceding three sections, the expected utility of this observer (and thus its performance, as it too is a special case of the ideal observer) depends on only three quantities related to conditional classification rates (but not the observer's decision utilities), namely the quantities in square brackets in (75).

The first quantity, being a weighted sum of "sensitivities" with positive weights, is immediately seen to be quite suitable for the dependent variable of an ROC surface — a higher value of this quantity is clearly preferable to a lower one. (Indeed, $P(\pi_1)P_{11} + P(\pi_2)P_{22}$ has an intuitive interpretation as the probability of a randomly drawn observation being both (i) from either class π_1 or π_2 and also (ii) correctly classified as such. Compare the corresponding quantity $P_{11} + P_{22}$ from Sec. IV, which is technically not even a probability.) The second two quantities in square brackets in (75) discourage any such straightforward interpretation, but this is perhaps to be expected: the pleasantly symmetric form of the Scurfield decision rule of (39)–(41) in this case holds in the $(P(\pi_1|\vec{\mathbf{x}}), P(\pi_2|\vec{\mathbf{x}}))$ decision variable plane; due to the complexity of the

transformation in (61), this symmetry will be lost in the likelihood ratio decision variable plane, and the expression for expected utility will be correspondingly opaque.

In any case, we now consider evaluating the performance of an arbitrary observer in an ROC-like space constructed from the quantities $P(\pi_1)P_{11}+P(\pi_2)P_{22}$, $P(\pi_2)P_{32}+2P(\pi_2)P_{22}-P(\pi_3)P_{13}$, and $P(\pi_1)P_{31}-2P(\pi_2)P_{22}-P(\pi_3)P_{23}$. We will define the ROC-like surface used to evaluate observer performance as the first quantity considered as a function of the second two. To find the optimal observer with respect to this restricted performance evaluation method, we apply the Neyman-Pearson criterion to maximize $P(\pi_1)P_{11}+P(\pi_2)P_{22}$ subject to the constraints $P(\pi_2)P_{32}+2P(\pi_2)P_{22}-P(\pi_3)P_{13}=\alpha_1, P(\pi_1)P_{31}-2P(\pi_2)P_{22}-P(\pi_3)P_{23}=\alpha_2)$. We define the function

$$\begin{aligned} F_{\text{Scurfield:AP}} \equiv\ & P(\pi_1)P_{11}+P(\pi_2)P_{22} \\ & +\lambda_1[P(\pi_2)P_{32}+2P(\pi_2)P_{22}-P(\pi_3)P_{13}-\alpha_1] \\ & +\lambda_2[P(\pi_1)P_{31}-2P(\pi_2)P_{22}-P(\pi_3)P_{23}-\alpha_2], \end{aligned} \tag{76}$$

where λ_1 and λ_2 are the Lagrange multipliers.

Using (12), this can be simplified to yield

$$\begin{aligned} F_{\text{Scurfield:AP}} =\ & -\lambda_1\alpha_1-\lambda_2\alpha_2+P(\pi_1)\int_{Z_1} p(\vec{x}|\pi_1)\,d^m\vec{x}+P(\pi_2)\int_{Z_2} p(\vec{x}|\pi_2)\,d^m\vec{x} \\ & +\lambda_1\left[P(\pi_2)\int_{Z_3} p(\vec{x}|\pi_2)\,d^m\vec{x}+2P(\pi_2)\int_{Z_2} p(\vec{x}|\pi_2)\,d^m\vec{x}\right. \\ & \left.-P(\pi_3)\int_{Z_1} p(\vec{x}|\pi_3)\,d^m\vec{x}\right] \\ & +\lambda_2\left[P(\pi_1)\int_{Z_3} p(\vec{x}|\pi_1)\,d^m\vec{x}-2P(\pi_2)\int_{Z_2} p(\vec{x}|\pi_2)\,d^m\vec{x}\right. \\ & \left.-P(\pi_3)\int_{Z_2} p(\vec{x}|\pi_3)\,d^m\vec{x}\right]. \end{aligned} \tag{77}$$

Collecting terms with given domains of integration yields

$$\begin{aligned} F_{\text{Scurfield:AP}} =\ & -\lambda_1\alpha_1-\lambda_2\alpha_2 \\ & +\int_{Z_1} P(\pi_1)p(\vec{x}|\pi_1)-\lambda_1P(\pi_3)p(\vec{x}|\pi_3)\,d^m\vec{x} \\ & +\int_{Z_2} P(\pi_2)p(\vec{x}|\pi_2)+2(\lambda_1-\lambda_2)P(\pi_2)p(\vec{x}|\pi_2)-\lambda_2P(\pi_3)p(\vec{x}|\pi_3)\,d^m\vec{x} \\ & +\int_{Z_3} \lambda_1P(\pi_2)p(\vec{x}|\pi_2)+\lambda_2P(\pi_1)p(\vec{x}|\pi_1)\,d^m\vec{x}. \end{aligned} \tag{78}$$

$F_{\text{Scurfield:AP}}$ can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the integrand over Z_i in (78) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{aligned}
\text{decide}\quad \pi_1 \quad \text{iff}\quad & P(\pi_1)p(\vec{x}|\pi_1) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\
& > P(\pi_2)p(\vec{x}|\pi_2) + 2(\lambda_1 - \lambda_2)P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\
\text{and}\quad & P(\pi_1)p(\vec{x}|\pi_1) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\
& > \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \qquad (79) \\
\text{decide}\quad \pi_2 \quad \text{iff}\quad & P(\pi_2)p(\vec{x}|\pi_2) + 2(\lambda_1 - \lambda_2)P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\
& \geq P(\pi_1)p(\vec{x}|\pi_1) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\
\text{and}\quad & P(\pi_2)p(\vec{x}|\pi_2) + 2(\lambda_1 - \lambda_2)P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\
& > \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \qquad (80) \\
\text{decide}\quad \pi_3 \quad \text{iff}\quad & \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \geq P(\pi_1)p(\vec{x}|\pi_1) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\
\text{and}\quad & \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \\
& \geq P(\pi_2)p(\vec{x}|\pi_2) + 2(\lambda_1 - \lambda_2)P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3). \qquad (81)
\end{aligned}$$

At this point, we could divide the above equations by $p(\vec{x}|\pi_3$ to obtain decision rules in terms of the likelihood ratios, as in the preceding sections. However, it is in this case more convenient to work with the *a posteriori* class membership probabilities directly; moreover, because we have established that (58)–(60) represent the boundary lines of an ideal observer decision rule, we are justified in doing so. Thus, given that $P(\pi_i)p(\vec{x}|\pi_i) = P(\pi_i|\vec{x})p(\vec{x})$, we divide (79)–(81) by $p(\vec{x})$ to obtain

$$\begin{aligned}
\text{decide}\quad \pi_1 \quad \text{iff}\quad & P(\pi_1|\vec{x}) - \lambda_1 P(\pi_3|\vec{x}) \\
& > P(\pi_2|\vec{x}) + 2(\lambda_1 - \lambda_2)P(\pi_2|\vec{x}) - \lambda_2 P(\pi_3|\vec{x}) \\
\text{and}\quad & P(\pi_1|\vec{x}) - \lambda_1 P(\pi_3|\vec{x}) \\
& > \lambda_1 P(\pi_2|\vec{x}) + \lambda_2 P(\pi_1|\vec{x}) \qquad (82) \\
\text{decide}\quad \pi_2 \quad \text{iff}\quad & P(\pi_2|\vec{x}) + 2(\lambda_1 - \lambda_2)P(\pi_2|\vec{x}) - \lambda_2 P(\pi_3|\vec{x})
\end{aligned}$$

$$\geq P(\pi_1|\vec{x}) - \lambda_1 P(\pi_3|\vec{x})$$

$$\text{and} \quad P(\pi_2|\vec{x}) + 2(\lambda_1 - \lambda_2)P(\pi_2|\vec{x}) - \lambda_2 P(\pi_3|\vec{x})$$

$$> \lambda_1 P(\pi_2|\vec{x}) + \lambda_2 P(\pi_1|\vec{x}) \tag{83}$$

$$\text{decide} \quad \pi_3 \quad \text{iff} \quad \lambda_1 P(\pi_2|\vec{x}) + \lambda_2 P(\pi_1|\vec{x}) \geq P(\pi_1|\vec{x}) - \lambda_1 P(\pi_3|\vec{x})$$

$$\text{and} \quad \lambda_1 P(\pi_2|\vec{x}) + \lambda_2 P(\pi_1|\vec{x})$$

$$\geq P(\pi_2|\vec{x}) + 2(\lambda_1 - \lambda_2)P(\pi_2|\vec{x}) - \lambda_2 P(\pi_3|\vec{x}). \tag{84}$$

As noted at the beginning of this section, $P(\pi_3|\vec{x}) = 1 - P(\pi_1|\vec{x}) - P(\pi_2|\vec{x})$. After rearranging terms, the boundary lines which partition the $(P(\pi_1|\vec{\mathbf{x}}), P(\pi_1|\vec{\mathbf{x}}))$ decision variable plane into the regions Z_1, Z_2, and Z_3 are found to be

$$(1 + \lambda_1 - \lambda_2)P(\pi_1|\vec{x}) - (1 + \lambda_1 - \lambda_2)P(\pi_2|\vec{x}) = \lambda_1 - \lambda_2 \qquad \{\text{“1-}\textit{vs.}\text{-2”}\} \tag{85}$$

$$(1 + \lambda_1 - \lambda_2)P(\pi_1|\vec{x}) = \lambda_1 \qquad \{\text{“1-}\textit{vs.}\text{-3”}\} \tag{86}$$

$$(1 + \lambda_1 - \lambda_2)P(\pi_2|\vec{x}) = \lambda_2 \qquad \{\text{“2-}\textit{vs.}\text{-3”}\}. \tag{87}$$

If we define the quantities $\gamma_1 \equiv \lambda_1/(1 + \lambda_1 - \lambda_2)$ and $\gamma_2 \equiv \lambda_2/(1 + \lambda_1 - \lambda_2)$, and further require $0 < \lambda_1$ and $0 < \lambda_2 < \min\{1, (\lambda_1 + 1)/2\}$ (so that $0 < \gamma_1 < 1$ and $0 < \gamma_2 < 1$), then the resulting decision strategy is found to be identical to that stated in (58)–(60). This special case of the observer proposed by Scurfield, which we have shown to be a special case of the ideal observer [14], has a performance that depends only on the quantities $P(\pi_1)P_{11} + P(\pi_2)P_{22}$, $P(\pi_2)P_{32} + 2P(\pi_2)P_{22} - P(\pi_3)P_{13}$, and $P(\pi_1)P_{31} - 2P(\pi_2)P_{22} - P(\pi_3)P_{23}$ by (75). The observer described above is indeed that which obtains optimal performance with respect to this set of quantities related to the conditional classification rates.

VI. Conclusions

Given the rapidly increase in complexity of the utility constraints and performance evaluation criteria as one proceeds from Secs. II to V, it is quite possible for the main point of the above analyses to become obscured. That main point is that, for each of a variety of constrained special cases of the three-class ideal observer, the performance of that observer is completely describable, in an expected-utility sense, by only two decision criteria and three quantities related

to conditional classification rates. This represents a considerable simplification from the general model, which is known to involve five decision criteria and six conditional classification rates.

It should be immediately acknowledged that such simplified models may ultimately prove to be of limited practical importance. Given an observer known to closely approximate the behavior of the ideal observer, or indeed given a human observer, it is difficult to conceive of a pragmatic way to externally constrain the observer's decision utilities to match a particular model such as the ones described above. On the other hand, an algorithmic observer (such as an implementation of a computerized scheme for computer-aided diagnosis) might readily allow such constraints on its decision *rules* to be implemented; however, the assumption that the probability density functions of the decision *variables* generated by the scheme do indeed follow those required by the ideal observer model would generally be unverifiable, given the limited amount of data typically available for training and testing such a scheme.

Despite these limitations, it remains an acknowledged fact that a fully general extension of ROC analysis to classification tasks with three or more classes has yet to be developed. Although the investigation of constrained and therefore tractable observer models should not be considered an end unto itself, a thorough understanding of such models is almost certain to prove necessary for the development of more general observer models. We believe that demonstrating particular constrained ideal observer models to be complete as well as tractable will be a crucial step toward this understanding.

REFERENCES

[1] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.

[2] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, no. 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.*, vol. 31, pp. 81–90, 2004.

[4] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I.* New York: John Wiley & Sons, 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.*, vol. 23, pp. 891–895, 2004.

[6] C. E. Metz and X. Pan, " 'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.*, vol. 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.*, vol. 24, pp. 293–299, 2005.