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.
Charles E. Metz, Ph.D.
Maryellen L. Giger, Ph.D.

CONTRACTING ORGANIZATION: University of Chicago
Chicago, IL 60637

REPORT DATE: May 2005

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. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2005	Annual (1 May 04 - 30 Apr 05)

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

5. FUNDING NUMBERS
W81XWH-04-1-0495

6. AUTHOR(S)
Darrin C. Edwards, Ph.D.
Charles E. Metz, Ph.D.
Maryellen L. Giger, Ph.D.

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

E-Mail: d-edwards@uchicago.edu

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. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

Our goal is to develop a fully automated classification scheme for computer-aided diagnosis in mammography. Our proposed scheme would classify computer detections into three groups: malignant lesions, benign lesions, and false-positive computer detections. This method presents considerable difficulties in terms of both signal detection theory and performance evaluation methods. During the past year we have focused our efforts on theoretical understanding of the behavior and properties of the three-group classifier. We have proven that an obvious generalization of the well-known two-group performance metric is not in fact a useful performance metric for classification tasks with three or more groups. We have developed an evaluation technique, independent of those we have previously developed, for assessing the accuracy of estimates of ideal observer decision variables. We have analyzed several recently proposed three-group classification methods in terms of the three-group ideal observer. Finally. we have shown that the three decision boundary lines used by the three-group ideal observer are not arbitrary, but are intricately related to one another. A three-group classifier could potentially allow radiologists to detect more malignant breast lesions without increasing their false-positive biopsy rates.

14. SUBJECT TERMS
Computer-aided diagnosis, X-ray imaging

15. NUMBER OF PAGES
79

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

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. 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 been more theoretical than practical so far, but our initial results 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.

Our research accomplishments to date have focused almost entirely 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.

By far the most important result achieved so far was our discovery and proof that an obvious generalization of the well-known performance metric, the area under the ROC curve (AUC), is not in fact useful in tasks with three or more groups [19]. (See Appendix A.) This accomplishment relates directly to Task 4.(b) above, which implicitly requires a well-defined performance metric with respect to which the statistical significance of differences in performance may be computed. Although arguably a "negative" rather than "positive" result — a well-defined performance metric has not yet been found — this result has been very well received in the observer performance and CAD research communities. First, it serves as a striking yet typical example of how intuition can often be an unreliable guide in extending methodology from the two-group classification task to tasks with three or more groups. Second, it clearly indicates that the search for such a well-defined performance metric will yield a deeper understanding of the properties of three-group observer performance, particularly as characterized by ROC analysis.

We stated above that exact determination of the ideal observer's decision variables requires knowledge of the PDFs from which the observational data to be classified were drawn. The tool we have been using for some time now to estimate ideal observer decision variables from samples of observational data is the Bayesian artificial neural network (BANN) [20]. In previous simulation studies in which the PDFs of the observational data are known, the output of the BANN was found to agree with the calculated ideal observer decision variables for two-group [4] and three-group [21] classification tasks. In practice, one does not have the PDFs of real observational data, but we previously developed a means of evaluating three-group BANN decision variables by comparing them with two-group BANN decision variables obtained from simplified two-group tasks using the same observational data [18]. During the past year, we developed an independent technique for evaluating three-group

BANN estimates of ideal observer decision variables, again based on theoretical properties of the three-group ideal observer [22]. (See Appendix B.) This result is important because the three-group classifier we are developing under the current research will be trained and tested using feature data from actual mammograms; thus, we will not have access to the PDFs from which those data are drawn. In addition to three-group ROC analysis methods to be developed by extension from existing two-group methods [5], it will be beneficial to have a direct method of judging the ability of the BANN decision variables to accurately estimate ideal observer decision variables.

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 [23–25]. 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 evaluated the different methods in terms of the three-group ideal observer, both in preliminary work [26] (see Appendix C) and later through more in-depth analysis [27] (see Appendix D). In addition to providing us with valuable insight and experience in comparing different classifiers, which should ultimately prove directly relevant to Task 4. above, 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.

Most recently, we have thoroughly investigated 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 [28]. (See Appendix E.) 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. 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

- Proof that an obvious generalization of the well-known two-group performance metric, the AUC, is not useful in classification tasks with three or more groups (Appendix A)
- Development of a novel technique for evaluating the quality of BANN estimates of ideal observer decision variables in the absence of three-group ROC analysis methodology and observational data PDFs (Appendix B)
- Analysis of several proposed three-group classification methods in the literature in terms of the three-group ideal observer (Appendices C, D)

- Detailed investigation of the relationships among the decision boundary lines used by the three-group ideal observer (Appendix E)

4 Reportable Outcomes

- 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.

- D. C. Edwards and C. E. Metz, "Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions," 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. 174–182. [Conference presentation and proceedings paper.]

- 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. [Conference presentation and proceedings paper.]

- 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, (submitted).

- D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.*, 2005, (submitted).

5 Conclusions

During the past year we have focused our efforts on theoretical understanding of the behavior and properties of the three-group classifier. We have proven that an obvious generalization of the well-known two-group performance metric, the AUC, is not in fact a useful performance metric for classification tasks with three or more groups. We have developed an evaluation technique, independent of those we have previously developed, for assessing the ability of BANN decision variables to accurately estimate ideal observer decision variables. We have analyzed several recently proposed three-group classification methods in terms of the three-group ideal observer. Finally. we have shown that the three decision boundary lines used by the three-group ideal observer are not arbitrary, but are intricately related to one another.

Although these results are 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, 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.

[20] D. J. S. MacKay, "Bayesian methods for adaptive models," Ph.D. Thesis, California Institute of Technology, Pasadena, CA, 1992.

[21] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network," in Proc. SPIE Vol. 4686 *Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, Dev P. Chakraborty and Elizabeth A. Krupinski, Eds., SPIE, Bellingham, WA, 2002, pp. 1–12.

[22] D. C. Edwards and C. E. Metz, "Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions," 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. 174–182.

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

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

[25] 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.

[26] 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.

[27] ——, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.*, 2005, (submitted).

[28] ——, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.*, 2005, (submitted).

The Hypervolume Under the ROC Hypersurface of "Near-Guessing" and "Near-Perfect" Observers in N-Class Classification Tasks

Darrin C. Edwards*, Charles E. Metz, and Robert M. Nishikawa

Abstract—**We express the performance of the N-class "guessing" observer in terms of the $N^2 - N$ conditional probabilities which make up an N-class receiver operating characteristic (ROC) space, in a formulation in which sensitivities are eliminated in constructing the ROC space (equivalent to using false-negative fraction and false-positive fraction in a two-class task). We then show that the "guessing" observer's performance in terms of these conditional probabilities is completely described by a degenerate hypersurface with only $N - 1$ degrees of freedom (as opposed to the $N^2 - N - 1$ required, in general, to achieve a true hypersurface in such a ROC space). It readily follows that the hypervolume under such a degenerate hypersurface must be zero when $N > 2$. We then consider a "near-guessing" task; that is, a task in which the N underlying data probability density functions (pdfs) are nearly identical, controlled by $N - 1$ parameters which may vary continuously to zero (at which point the pdfs become identical). With this approach, we show that the hypervolume under the ROC hypersurface of an observer in an N-class classification task tends continuously to zero as the underlying data pdfs converge continuously to identity (a "guessing" task). The hypervolume under the ROC hypersurface of a "perfect" ideal observer (in a task in which the N data pdfs never overlap) is also found to be zero in the ROC space formulation under consideration. This suggests that hypervolume may not be a useful performance metric in N-class classification tasks for $N > 2$, despite the utility of the area under the ROC curve for two-class tasks.**

Index Terms—**N-class classification, ROC analysis, ROC performance metrics.**

I. 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 [1]–[5], and one for classifying known lesions as malignant or benign [6]–[10]. 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 nonlesions (FP computer detections), and the classifier to be estimated is the ideal observer decision function for this task. Such an approach presents considerable difficulties of its own. On the one hand, decision functions, in particular ideal observer decision functions, increase rapidly in complexity with the number of classes involved. On the other hand, fully general performance evaluation methods, in particular a fully general three-class extension of receiver operating characteristic (ROC) analysis, have yet to be developed for such a task.

Although we have had preliminary success in using Bayesian artificial neural networks (BANNs) [11], [12] to estimate three-class ideal-observer-related decision variables [13], [14], the task of developing an extension of ROC analysis to classification tasks with three or more classes has proved somewhat more daunting. Our initial efforts in this direction have, thus, been more theoretical than practical so far [15]. One issue we began to investigate recently was the calculation of an obvious generalization of the well-known area under the ROC curve (AUC) performance metric, a quantity we are calling the "hypervolume under the ROC hypersurface." Detailed consideration of the integrals involved in calculating this quantity led us to the counterintuitive conclusion that, despite the great success and utility of the AUC performance metric in two-class classification tasks, the hypervolume under the ROC hypersurface does not appear to be a useful performance metric in N-class classification tasks for $N > 2$. The proof of this claim is arrived at by considering observer performance in two extremes: the "guessing" observer and the "perfect" observer. It should be explicitly noted that in our formulation, sensitivities are eliminated in constructing the ROC space; this is equivalent to using false-negative fraction (FNF) and false-positive fraction (FPF) in a two-class task. In such a formulation, the "guessing" observer in a two-class task achieves an AUC of 0.5 as expected, but the "perfect" observer in a two-class task achieves an AUC of zero.

In Section II, we consider the properties of the "guessing" observer in an N-class classification task, and of its ROC

Manuscript received June 23, 2004; revised November 8, 2004. This work was supported in part by the National Cancer Institute under Grant R01-CA60187 and in part by the US Army Breast Cancer Research Program under Grant W81XWH-04-1-0495. 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 So. Maryland Avenue, Chicago, IL 60637 USA (e-mail: d-edwards@uchicago.edu)

C. E. Metz and R. M. Nishikawa are with the Department of Radiology, The University of Chicago, Chicago, IL 60037 USA.

Digital Object Identifier 10.1109/TMI.2004.841227

hypersurface. In Section III, we consider the properties of the ROC hypersurface of a so-called "near-guessing" observer, i.e., an observer in a task for which the observational data probability density functions (pdfs) are not identical, but differ only by arbitrarily small amounts. In Section IV, we then show that the hypervolume under the ROC hypersurface of such a "near-guessing" observer will continuously approach the hypervolume under the ROC hypersurface of the "guessing" observer as the observational data pdfs continuously approach identity; furthermore, the hypervolume under the ROC hypersurface of the "guessing" observer is shown to be zero.

We then show in Section V that the hypervolume under the ROC hypersurface of the "perfect" observer is zero (as expected by analogy with the two-class task), and that the hypervolume under the ROC hypersurface of a "near-perfect" observer will approach zero continuously as the observational data pdfs are separated. Finally, in Section VI, we argue that these results taken together imply that the hypervolume under the ROC hypersurface is not a useful performance metric in N-class classification tasks for $N > 2$, despite the utility of the AUC performance metric in two-class tasks.

II. The ROC Hypersurface of the N-Class "Guessing" Observer

The performance of an observer in an N-class classification task is completely determined by a hypersurface with $N^2 - N - 1$ degrees of freedom in an $(N^2 - N)$-dimensional ROC space [16]. Without loss of generality, we can specify any point in the ROC space by a vector of the misclassification probabilities $[P(\mathbf{d} = \pi_1 \,|\, \mathbf{t} = \pi_2), \ldots, P(\mathbf{d} = \pi_1 \,|\, \mathbf{t} = \pi_N), P(\mathbf{d} = \pi_2 \,|\, \mathbf{t} = \pi_1), P(\mathbf{d} = \pi_2 \,|\, \mathbf{t} = \pi_3), \ldots, P(\mathbf{d} = \pi_2 \,|\, \mathbf{t} = \pi_N), P(\mathbf{d} = \pi_N \,|\, \mathbf{t} = \pi_{N-1}), P(\mathbf{d} = \pi_N \,|\, \mathbf{t} = \pi_1)]^\dagger$ [15]. Here the N classes are denoted by the labels $\pi_1, \ldots, \pi_N$; $\mathbf{d}$ denotes the class to which an observation is assigned (the "decision"); and $\mathbf{t}$ is the class to which it actually belongs (the "truth"). We use boldface type to denote statistically variable quantities. For simplicity, we write $P(\mathbf{d} = \pi_i \,|\, \mathbf{t} = \pi_j)$ as P_{ij}.

We can also, again without loss of generality, consider the ROC hypersurface to be given by P_{N1} considered as a function of the other $N^2 - N - 1$ misclassification probabilities [15]. Note that this formulation is equivalent, in a two-class classification task, to using FPF and FNF to characterize the ROC curve, rather than FPF and true-positive fraction (TPF), as is more common. In a two-class classification task, this produces ROC curves which are "upside-down" with respect to the standard formulation; we have adopted the nonstandard formulation described above because it has proven easier to generalize to classification tasks with more than two classes.

Some researchers have suggested [17], [18] that in, e.g., a three-class classification task, the set of three "sensitivities" ($P(\mathbf{d} = \pi_i \,|\, \mathbf{t} = \pi_i)$ in our notation) provides a complete description of observer performance. This is incorrect in general, because it ignores the $N^2 - N$ misclassification probabilities, not all of which are determined uniquely by the "sensitivities" when $N > 2$ unless particular restrictions are imposed on the observer's behavior. Complete quantification of the trade-offs available among the probabilities of various kinds of misclassification error is important in medical diagnosis, where different misclassification errors often have substantially different clinical consequences. Moreover, restrictions concerning the observer's behavior are inappropriate when considering the general behavior of ideal observers, human observers, or automated observers (such as automated schemes for computer-aided diagnosis) designed to approximate ideal or human observer behavior. Other researchers have reduced the three-class ROC hypersurface to more tractable two-dimensional surfaces in three-dimensional ROC spaces by explicitly imposing restrictions on the form of the observer's decision rule [19], [20], or on the utilities used by an ideal observer [21]. While such restrictions may ultimately prove to be of great pragmatic importance given the inherent complexity of multi-class classification tasks, our approach so far has been to attempt as general an understanding as possible of the unrestricted classification task.

Consider the performance of an observer which makes decisions by "guessing," that is, in a random fashion unrelated to the actual class $\mathbf{t}$ from which a given observation is drawn. (Note that this corresponds to the performance of the ideal observer when the pdfs of the observational data are identical, i.e., $p(\vec{\mathbf{x}} \,|\, \pi_1) = p(\vec{\mathbf{x}} \,|\, \pi_2) = \cdots = p(\vec{\mathbf{x}} \,|\, \pi_N)$.) In this case, we clearly must have

$$P_{12} = P_{13} = \cdots = P_{1N} \tag{1}$$

$$P_{21} = P_{23} = \cdots = P_{2N} \tag{2}$$

$$\cdots$$

$$P_{N1} = P_{N2} = \cdots = P_{N\,(N-1)}. \tag{3}$$

Defining $\alpha_i \equiv P_{iN}$ for $1 \leq i \leq N-1$, and $\alpha_N \equiv P_{N\,(N-1)}$, we see that the performance of the "guessing" observer is given by a locus of vectors of the form

$$\begin{bmatrix} \left.\begin{matrix}\alpha_1\\ \alpha_1\\ \vdots\end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix}\alpha_i\\ \alpha_i\\ \vdots\end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix}\alpha_N\\ \alpha_N\\ \vdots\end{matrix}\right\} N-1 \text{ elements} \end{bmatrix} \tag{4}$$

where all of the α_i are restricted to the range $[0, 1]$. Furthermore, note that

$$\begin{aligned} P(\mathbf{d} = \pi_i) &= \sum_{j=1}^{N} P_{ij} P(\mathbf{t} = \pi_j) \\ &= \sum_{j=1}^{N} \alpha_i \, P(\mathbf{t} = \pi_j) \\ &= \alpha_i \end{aligned} \tag{5}$$

which immediately gives $\alpha_N = 1 - \sum_{i=1}^{N-1} \alpha_i$. Thus, the performance of the "guessing" observer is given by

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{N(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \left.\begin{matrix} \alpha_1 \\ \alpha_1 \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} \alpha_i \\ \alpha_i \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} 1-\sum_{j=1}^{N-1}\alpha_j \\ 1-\sum_{j=1}^{N-1}\alpha_j \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \end{bmatrix}$$

$$= \vec{v}_0 + \sum_{i=1}^{N-1} \alpha_i \vec{v}_i. \tag{6}$$

This is the parametric equation for an $(N-1)$-dimensional plane in an $(N^2 - N)$-dimensional space; the actual performance of the "guessing" observer will of course be further restricted to a region within this plane such that $0 \leq \alpha_i \leq 1, 0 \leq 1 - \sum \alpha_i \leq 1$.

III. The ROC Hypersurface of an N-Class "Near-Guessing" Observer

Consider observational data $\vec{\mathbf{x}}$ drawn from N pdfs

$$p(\vec{x} \,|\, \mathbf{t} = \pi_1) = p(\vec{x} \,|\, \mathbf{t} = \pi_N) + \delta_1 h_1(\vec{x}) \tag{7}$$

$$\cdots$$

$$p(\vec{x} \,|\, \mathbf{t} = \pi_j) = p(\vec{x} \,|\, \mathbf{t} = \pi_N) + \delta_j h_j(\vec{x}) \tag{8}$$

$$\cdots$$

$$p(\vec{x} \,|\, \mathbf{t} = \pi_N) \tag{9}$$

where $0 \leq \delta_j \leq 1$, $\int h_j(\vec{x}) d^n\vec{x} = 0$, and $|h_j(\vec{x})| \leq p(\vec{x} \,|\, \mathbf{t} = \pi_N)$ for $1 \leq j \leq N-1$. In the limit as the δ_j all approach zero, we expect the performance of any observer for this task to converge smoothly to that of the "guessing" observer.

Decisions are made by partitioning the decision variable space into N regions, determined by a total of $N^2 - N - 1$ parameters; we denote these parameters by the components of a vector $\vec{\gamma}$. An observer which uses more than $N^2 - N - 1$ parameters for an N-class classification task can always be replaced by a simplified observer, such that the "excess" parameters are eliminated by the requirement that P_{N1} be minimized, thereby collapsing the dimensionality of the parameter space to $N^2 - N - 1$. On the other hand, an observer which uses fewer than $N^2 - N - 1$ decision parameters will fail to generate a true ROC hypersurface—i.e., one with $N^2 - N - 1$ degrees of freedom in the $(N^2 - N)$-dimensional ROC space. (An example in a three-class classification task would be an observer which sequentially performs a pair of binary classification tasks by first classifying observations as being "π_1" or "not π_1" based on the value of a single decision parameter, and then further classifying the "not π_1" observations as "π_2" or "π_3" based on the value of a second decision parameter [17], thus depending on fewer than the five degrees of freedom needed in a three-class classification task.) Such "degenerate" observers will not be considered here (apart from the "guessing" observer itself).

We can, thus, define N regions which partition the original data space, given particular values of the parameters $\vec{\gamma}$, by

$$\mathcal{D}_1(\vec{\gamma}) \equiv \{\vec{x} : \mathbf{d} = \pi_1 \text{ given } \vec{\gamma}\} \tag{10}$$

$$\cdots$$

$$\mathcal{D}_i(\vec{\gamma}) \equiv \{\vec{x} : \mathbf{d} = \pi_i \text{ given } \vec{\gamma}\} \tag{11}$$

$$\cdots$$

$$\mathcal{D}_N(\vec{\gamma}) \equiv \{\vec{x} : \mathbf{d} = \pi_N \text{ given } \vec{\gamma}\}. \tag{12}$$

For a nonrandom observer, the $\mathcal{D}_i$ can be expected to depend implicitly on the pdfs (7)–(9) and, therefore, on the δ_j. The misclassification probabilities which define the ROC hypersurface are then given by

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{N(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \int_{\mathcal{D}_1} p(\vec{x} \,|\, \mathbf{t} = \pi_2)\, d^n\vec{x} \\ \int_{\mathcal{D}_1} p(\vec{x} \,|\, \mathbf{t} = \pi_3)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_1} p(\vec{x} \,|\, \mathbf{t} = \pi_N)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_i} p(\vec{x} \,|\, \mathbf{t} = \pi_1)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_i} p(\vec{x} \,|\, \mathbf{t} = \pi_j)\, d^n\vec{x} \quad \{i \neq j\} \\ \vdots \\ \int_{\mathcal{D}_i} p(\vec{x} \,|\, \mathbf{t} = \pi_N)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_N} p(\vec{x} \,|\, \mathbf{t} = \pi_{N-1})\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_N} p(\vec{x} \,|\, \mathbf{t} = \pi_1)\, d^n\vec{x} \end{bmatrix}. \tag{13}$$

Using (7) and (8), we can rewrite this as

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{N(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} P_{1N} + \delta_2 \int_{\mathcal{D}_1} h_2(\vec{x})\, d^n\vec{x} \\ P_{1N} + \delta_3 \int_{\mathcal{D}_1} h_3(\vec{x})\, d^n\vec{x} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{iN} + \delta_1 \int_{\mathcal{D}_i} h_1(\vec{x})\, d^n\vec{x} \\ \vdots \\ P_{iN} + \delta_j \int_{\mathcal{D}_i} h_j(\vec{x})\, d^n\vec{x} \; \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{NN} + \delta_{N-1} \int_{\mathcal{D}_N} h_{N-1}(\vec{x})\, d^n\vec{x} \\ \vdots \\ P_{NN} + \delta_1 \int_{\mathcal{D}_N} h_1(\vec{x})\, d^n\vec{x} \end{bmatrix}. \tag{14}$$

Defining the functions $H_{ij} \equiv \int_{\mathcal{D}_i} h_j(\vec{x})\, d^n\vec{x}$ allows us to simplify the notation slightly

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} P_{1N} + \delta_2 H_{12} \\ P_{1N} + \delta_3 H_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{iN} + \delta_1 H_{i1} \\ \vdots \\ P_{iN} + \delta_j H_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{NN} + \delta_{N-1} H_{N(N-1)} \\ \vdots \\ P_{NN} + \delta_1 H_{N1} \end{bmatrix}. \tag{15}$$

Now of course $P_{NN} = 1 - \sum_{i=1}^{N-1} P_{iN}$; for simplicity, we will write $\alpha_i \equiv P_{iN}$. Equation (15) can now be written as

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \alpha_1 + \delta_2 H_{12} \\ \alpha_1 + \delta_3 H_{13} \\ \vdots \\ \alpha_1 \\ \vdots \\ \alpha_i + \delta_1 H_{i1} \\ \vdots \\ \alpha_i + \delta_j H_{ij} \quad \{i \neq j\} \\ \vdots \\ \alpha_i \\ \vdots \\ 1 - \sum_{j=1}^{N-1} \alpha_j + \delta_{N-1} H_{N(N-1)} \\ \vdots \\ 1 - \sum_{j=1}^{N-1} \alpha_j + \delta_1 H_{N1} \end{bmatrix} \tag{16}$$

which further simplifies to

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \left.\begin{matrix} \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} \alpha_i \\ \alpha_i \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} 1 - \sum_{j=1}^{N-1} \alpha_j \\ 1 - \sum_{j=1}^{N-1} \alpha_j \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \end{bmatrix} + \sum_{j=1}^{N-1} \delta_j \vec{w}_j \tag{17}$$

where the vectors $\vec{w}_j$ have components which depend only on H_{ij}. The first term on the right-hand side of this equation is just the expression for the "guessing" observer [cf. the left-hand side of (6)]. The other term on the righthand side of this equation tends to zero as the δ_j tend to zero. Note that the H_{ij} may in general depend on the δ_k via (10)–(12), but

$$\begin{aligned} |H_{ij}| &= \left| \int_{\mathcal{D}_i} h_j(\vec{x})\, d^n\vec{x} \right| \\ &\leq \int_{\mathcal{D}_i} |h_j(\vec{x})|\, d^n\vec{x} \\ &\leq \int_{\mathcal{D}_i} p(\vec{x}\,|\,\mathbf{t} = \pi_N)\, d^n\vec{x} \\ &= P_{iN} \\ &\leq 1. \end{aligned} \tag{18}$$

Thus, the H_{ij} are bounded, and will possess Taylor expansions in δ_k (i.e., will not depend on terms of the form δ_k^{-m} for positive integers m). Therefore, operating points on the ROC hypersurface of a "near-guessing" observer tend continuously toward points on the ROC hypersurface of the "guessing" observer. Note that the $N(N-1)$ terms $\alpha_i, \delta_j H_{ij}$ are not all independent, since they all depend implicitly for fixed δ_j on the $N^2 - N - 1$ decision parameters $\vec{\gamma}$. That is, the ROC hypersurface given by (17) possesses only $N^2 - N - 1$ degrees of freedom.

IV. The Hypervolume Under the ROC Hypersurface of an N-Class "Near-Guessing" Observer

In the preceding section, it was shown that the ROC hypersurface of a "near-guessing" observer tends continuously to the ROC hypersurface of a "guessing" observer as the pdfs of the observational data tend arbitrarily toward identical distributions. Intuitively, one would expect that the hypervolumes under these hypersurfaces should also tend toward each other. Since intuition can occasionally be an unreliable guide in analyzing N-class classification tasks, it would be reassuring if the results of the preceding section could be applied directly to the calculation of the relevant hypervolumes.

For this section, we will write P_{ij} as $P_{ij}(\vec{\gamma})$, emphasizing that it is a function of the decision parameters chosen. We, thus, rewrite (15) to obtain

$$\begin{bmatrix} P_{12}(\vec{\gamma}) \\ P_{13}(\vec{\gamma}) \\ \vdots \\ P_{1N}(\vec{\gamma}) \\ \vdots \\ P_{i1}(\vec{\gamma}) \\ \vdots \\ P_{ij}(\vec{\gamma}) \quad \{i \neq j\} \\ \vdots \\ P_{iN}(\vec{\gamma}) \\ \vdots \\ P_{N,(N-1)}(\vec{\gamma}) \\ \vdots \\ P_{N,1}(\vec{\gamma}) \end{bmatrix} = \begin{bmatrix} P_{1N}(\vec{\gamma}) + \delta_2 H_{12}(\vec{\gamma}) \\ P_{1N}(\vec{\gamma}) + \delta_3 H_{13}(\vec{\gamma}) \\ \vdots \\ P_{1N}(\vec{\gamma}) \\ \vdots \\ P_{iN}(\vec{\gamma}) + \delta_1 H_{i1}(\vec{\gamma}) \\ \vdots \\ P_{iN}(\vec{\gamma}) + \delta_j H_{ij}(\vec{\gamma}) \quad \{i \neq j\} \\ \vdots \\ P_{iN}(\vec{\gamma}) \\ \vdots \\ P_{N,(N-1)}(\vec{\gamma}) \\ \vdots \\ P_{N,(N-1)}(\vec{\gamma}) - \delta_{N-1} H_{N(N-1)}(\vec{\gamma}) \\ +\delta_1 H_{N1}(\vec{\gamma}) \end{bmatrix}. \tag{19}$$

To find the hypervolume under the ROC surface given by P_{N1} considered as a function of $(P_{12}, P_{13}, \ldots, P_{ij}, \ldots, P_{N(N-1)}, \ldots, P_{N2})$, one must evaluate the integral

$$\int \cdots \int P_{N1} d^{N^2-N-1}\vec{P}. \tag{20}$$

(The domain of the integral is simply the set of all P_{ij} such that P_{N1} is defined.) Note that, for the "guessing" observer, we expect this integral to be zero when $N > 2$ due to dimensionality considerations—the ROC hypersurface has only $N-1$ degrees of freedom (cf. (6)), not the N^2-N-1 required in this (N^2-N)-dimensional ROC space. To see this explicitly, one can rearrange the order of integration and consider the innermost integral $\int P_{N1} dP_{N(N-1)}$ for fixed values of the other misclassification probabilities. Then the limits of integration of this innermost definite integral become, again by (6)

$$\int_{1-\sum P_{jN}}^{1-\sum P_{jN}} P_{N1} dP_{N(N-1)} \quad \{j < N\} \tag{21}$$

which is zero by inspection.

We now return to the general case of a "near-guessing" observer. One way to evaluate the integral in (20) is to reexpress it explicitly in terms of the decision parameters $\vec{\gamma}$, via the Jacobian

$$J \equiv \begin{vmatrix} \frac{\partial P_{12}}{\partial \gamma_1} & \frac{\partial P_{12}}{\partial \gamma_2} & \frac{\partial P_{12}}{\partial \gamma_3} & \cdots & \frac{\partial P_{12}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{1N}}{\partial \gamma_1} & \frac{\partial P_{1N}}{\partial \gamma_2} & \frac{\partial P_{1N}}{\partial \gamma_3} & \cdots & \frac{\partial P_{1N}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{ij}}{\partial \gamma_1} & \frac{\partial P_{ij}}{\partial \gamma_2} & \frac{\partial P_{ij}}{\partial \gamma_3} & \cdots & \frac{\partial P_{ij}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N(N-1)}}{\partial \gamma_1} & \frac{\partial P_{N(N-1)}}{\partial \gamma_2} & \frac{\partial P_{N(N-1)}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N2}}{\partial \gamma_1} & \frac{\partial P_{N2}}{\partial \gamma_2} & \frac{\partial P_{N2}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N2}}{\partial \gamma_{N^2-N-1}} \end{vmatrix} \tag{22}$$

where the vertical bars indicate that the determinant of the enclosed matrix is to be taken, and where γ_i denotes the ith component of $\vec{\gamma}$. (We assume that indices of the parameters $\vec{\gamma}$ have been chosen appropriately so that no negative sign is introduced, i.e., volumes remain positive.) For the "guessing" observer, this reduces to

$$J_{\text{guessing}} \equiv \begin{vmatrix} \frac{\partial P_{1N}}{\partial \gamma_1} & \frac{\partial P_{1N}}{\partial \gamma_2} & \frac{\partial P_{1N}}{\partial \gamma_3} & \cdots & \frac{\partial P_{1N}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{1N}}{\partial \gamma_1} & \frac{\partial P_{1N}}{\partial \gamma_2} & \frac{\partial P_{1N}}{\partial \gamma_3} & \cdots & \frac{\partial P_{1N}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{iN}}{\partial \gamma_1} & \frac{\partial P_{iN}}{\partial \gamma_2} & \frac{\partial P_{iN}}{\partial \gamma_3} & \cdots & \frac{\partial P_{iN}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N(N-1)}}{\partial \gamma_1} & \frac{\partial P_{N(N-1)}}{\partial \gamma_2} & \frac{\partial P_{N(N-1)}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N(N-1)}}{\partial \gamma_1} & \frac{\partial P_{N(N-1)}}{\partial \gamma_2} & \frac{\partial P_{N(N-1)}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_{N^2-N-1}} \end{vmatrix} \tag{23}$$

where $P_{N(N-1)} = P_{NN} = 1 - \sum_{j=1}^{N-1} P_{jN}$. For a "near-guessing" observer, we combine (19) and (22) to obtain

$$J_{\text{near}} \equiv \begin{vmatrix} \cdots & \frac{\partial (P_{1N}+\delta_2 H_{12})}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial P_{1N}}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial (P_{iN}+\delta_j H_{ij})}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial (P_{N(N-1)}-\delta_{N-1} H_{N(N-1)}+\delta_2 H_{N2})}{\partial \gamma_k} & \cdots \end{vmatrix}. \tag{24}$$

From the properties of determinants [22], it can be shown that, to first order in the δ_j,

$$J_{\text{near}} = J_{\text{guessing}} + \sum_{j=1}^{N-1} \delta_j J_j + \cdots \tag{25}$$

where the J_j are bounded and continuous with respect to the δ_j.

If we denote the hypervolume under the ROC hypersurface of the "guessing" observer by

$$\begin{aligned} I_{\text{guessing}} &= \int \cdots \int P_{N1}\, d^{N^2-N-1}\vec{P} \\ &= \int \cdots \int P_{N(N-1)}(\vec{\gamma}) J_{\text{guessing}}\, d^{N^2-N-1}\vec{\gamma} \end{aligned} \tag{26}$$

then the hypervolume under the ROC hypersurface of a "near-guessing" observer becomes, again to first order in the δ_j

$$\begin{aligned} I_{\text{near}} &= \int \cdots \int [P_{N,(N-1)}(\vec{\gamma}) - \delta_{N-1} H_{N(N-1)}(\vec{\gamma}) \\ &\qquad + \delta_1 H_{N1}(\vec{\gamma})] \\ &\qquad \times \left[J_{\text{guessing}} + \sum_{j=1}^{N-1} \delta_j J_j + \cdots \right] d^{N^2-N-1}\vec{\gamma} \end{aligned} \tag{27}$$

$$= I_{\text{guessing}} + \sum_{j=1}^{N-1} \delta_j I_j + \cdots \tag{28}$$

where the integrals I_j are bounded (i.e., they may depend on higher integral powers of δ_j, but not on δ_j^{-m} for positive integers m). That is, in the limit as the δ_j tend toward zero, I_{near} tends toward I_{guessing} in a continuous fashion.

V. The Hypervolume Under the ROC Hypersurface of an N-Class "Near-Perfect" Observer

In the preceding sections, we established that the hypervolume under the ROC hypersurface of a "guessing" observer is zero, and furthermore that this result is not singular: an observer in a "near-guessing" task will achieve a ROC hypersurface with hypervolume approaching zero continuously as the data pdfs approach identity. An ideal observer in a "perfect" task—i.e., in which the data pdfs never overlap—will also achieve a ROC hypersurface with zero hypervolume, because it can achieve the operating point $\vec{0}$ and, thus, will not, for any rational decision rule, achieve points interior to the unit hypercube defining ROC space. It is reasonable to ask whether "near-perfect" observers, performing tasks for which

the overlap in the underlying data pdfs is nearly negligible, behave similarly to "near-guessing" observers, in the sense that the hypervolume under the ROC hypersurface of such an observer will approach zero in a continuous fashion.

Consider observational data $\vec{\mathbf{x}}$ drawn from N pdfs $p(\vec{x} \mid \mathbf{t} = \pi_j)$ where $1 \leq j \leq N$. We denote the mean of $p(\vec{x} \mid \mathbf{t} = \pi_j)$ by $\vec{\mu}_j$ and note that, without loss of generality, the mean of $p(\vec{x} \mid \mathbf{t} = \pi_N)$ can be taken to be $\vec{0}$. Furthermore, note that we can apply a linear transformation to the data $\vec{\mathbf{x}}$ and, thus, effectively to the $\vec{\mu}_j$, such that each of the resulting $\vec{\mu}_j$ is either 1) mutually orthogonal to, or 2) a scalar multiple of, any of the other $\vec{\mu}_i$. Because the transformation applied is linear, the ideal observer for this task will remain the same, and hence the task itself can be considered essentially unchanged.

Let us consider now an observer for this task which is generally not ideal; in fact, we will consider only a single operating point achieved by this observer. The observer decides $d = \pi_i$ for a given observation $\vec{x}$ if

$$(\vec{x} - \vec{\mu}_i) \cdot \frac{(\vec{\mu}_j - \vec{\mu}_i)}{|\vec{\mu}_j - \vec{\mu}_i|} < \frac{1}{2}|\vec{\mu}_j - \vec{\mu}_i| \quad \{j : 1 \leq j \leq N, j \neq i\} \quad (29)$$

with equality for any such relation between two classes being decided in an arbitrary but consistent manner. That is, the observer places hyperplanes between the means of any two classes when attempting to decide between those classes (rather than placing those hyperplanes in the likelihood ratio decision variable space, as would the ideal observer).

Now suppose the task is made slightly "easier," while the observer itself remains unchanged. That is, consider the mean of one pdf, say $\vec{\mu}_i$ for $i \neq N$, being increased by a factor $1 + \delta$ for $0 \leq \delta \leq 1$, while the location of the decision hyperplanes does not change, except in the special case where $\vec{\mu}_j = \alpha\vec{\mu}_i$ for some other pdf (again with $j \neq N$). In this latter case we increase both means ($\vec{\mu}'_j = (1+\delta)\vec{\mu}_j, \vec{\mu}'_i = (1+\delta)\vec{\mu}_i$), and the location of the corresponding decision hyperplane shifts accordingly.

Note that $\vec{\mu}'_i$ is now further away from each decision hyperplane relevant to $\mathbf{d} = \pi_i$ in (29). In the case $\vec{\mu}_j = \alpha\vec{\mu}_i$, the decision hyperplane is now a distance of $|(\vec{\mu}'_j) - (\vec{\mu}'_i)/(2)| = (1+\delta)|(\vec{\mu}_j) - (\vec{\mu}_i)/(2)|$ from $\vec{\mu}'_i$. For noncollinear $\vec{\mu}_j$, the direction from $\vec{\mu}'_i$ to the decision hyperplane is given by $\vec{\mu}_j - \vec{\mu}_i$, and since $\vec{\mu}_j$ and $\vec{\mu}'_i$ are orthogonal, $(\vec{\mu}'_i - \vec{\mu}_i) \cdot (\vec{\mu}_j - \vec{\mu}_i) = -\delta|\vec{\mu}_i|^2$; since this quantity is negative, it follows that $\vec{\mu}'_i$ is further from that decision plane than $\vec{\mu}_i$.

It immediately follows from this that none of the misclassification probabilities making up the coordinates of the observer's operating point can increase when moving from the old task to the new one. To see this, consider a change of coordinates in the data space such that $\vec{\mu}'_i$ is now the origin. All of the decision hyperplanes separating this class from the others are effectively moving away from the center of its pdf; since the hyperplanes are translating without rotating, we see immediately that the probability P_{ii} cannot decrease (and will increase in general), while the other probabilities P_{ji} $(j \neq i)$ cannot increase (and will decrease in general).

Note that any pdf $p(\vec{x})$ must decrease more rapidly than $|\vec{x}|^{-n}$ for sufficiently large $|\vec{x}|$, where n is the dimensionality of $\vec{x}$. This allows us to state qualitatively the sense in which the observer under consideration is "near-perfect": we hypothesize that the $|\vec{\mu}_i|$ are all sufficiently large that this limiting condition is met. Given this condition, the only situation in which an error probability P_{ji} $(j \neq i)$ will fail to decrease is if this probability is already zero. By allowing all of the $|\vec{\mu}_i|$ to increase in the manner described above, we can clearly obtain in general a situation in which each of the misclassification probabilities is either decreasing, or equal to zero.

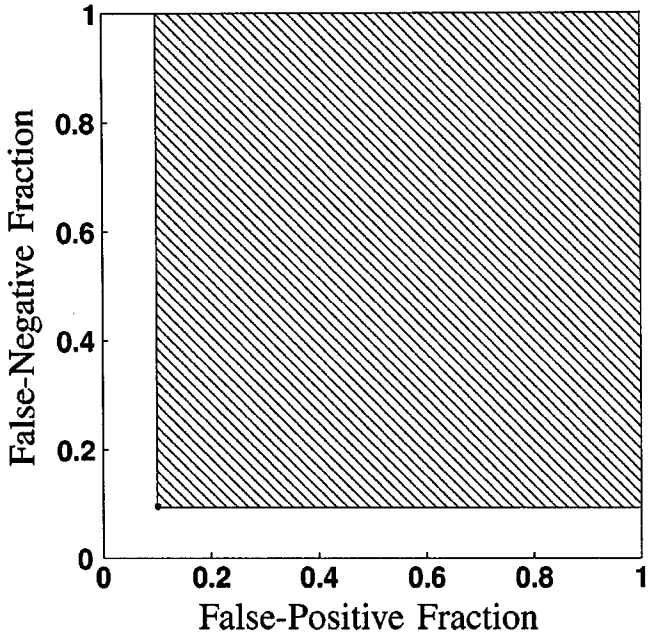


Fig. 1. Operating point of an observer in a two-class classification task with coordinates $(\mathrm{FPF_0}, \mathrm{FNF_0})$, denoted by the point at the lower left corner of the crosshatched region. Since no rational observer will achieve points in the crosshatched region, the area under this observer's ROC curve cannot be greater than $1 - (1 - \mathrm{FPF_0})(1 - \mathrm{FNF_0})$.

This implies that the hypervolume under the ROC hypersurfaces of the observers under consideration (however we chose to define their decision rules for operating points other than those described above) must also decrease as the task is made "easier" as described above. To see this, note that if a given observer achieves an operating point $\vec{P}$ on its ROC hypersurface, it cannot achieve another point $\vec{P}'$ such that the components of these points satisfy $P'_i > P_i (1 \leq i \leq N^2 - N)$ (because such an observer could be replaced by an observer which achieved $\vec{P}$ for all such points by using the original decision rule for the point $\vec{P}$, thereby achieving unambiguously better performance at those points). Thus, knowing that a given observer achieves an operating point of $\vec{P}$ implies that that observer's ROC hypersurface must have a hypervolume under it of no greater than $1 - \prod_{i=1}^{N^2-N}(1 - P_i)$; as the (nonzero) P_i decrease, this upper limit on the hypervolume must also decrease to zero. This point is illustrated in Fig. 1 for the two-class case; here the observer's false-negative fraction, $\mathrm{FNF_0}$, corresponds to P_{21}, and the false-positive fraction, $\mathrm{FPF_0}$, corresponds to P_{12}.

To summarize, we have shown that the known operating point of our simple observer will move closer to the origin for arbitrary data pdfs as those pdfs are moved further apart (i.e., as the underlying task is made "easier"), implying that the hypervolume under its ROC hypersurface will also converge to zero. In fact, reasoning as above, one can see that the ideal observer

will also be unable to achieve operating points within the region $P_i' > P_i$ $(1 \leq i \leq N^2 - N)$, since the ideal observer's ROC hypersurface is never above that of any other observer at any given point in the domain of the ROC space [15]. The hypervolume under the ideal observer's ROC hypersurface will, thus, also converge to zero as the underlying data pdfs are moved apart.

VI. Conclusion

In N-class classification tasks where $N > 2$, it can be shown that the hypervolume under the ROC hypersurface of both the "guessing" observer and the "perfect" observer are zero. More importantly, we have shown in each of these performance extremes that the convergence to zero is smooth rather than discontinuous. This convergence can be considered completely general for "near-guessing" observers and generally true for "near-perfect" observers which follow rational decision rules (analogous to false-negative fraction and false-positive fraction being monotonically related in a two-class task); that is, the conclusions appear to hold true for arbitrary underlying data pdfs.

In the two-class classification task, the area under the ROC curve (AUC) is considered a useful performance metric for a variety of reasons. One of the most pleasing and straightforward of these is the simple relationship between AUC and the "separability" of the two underlying data pdfs (i.e., the difficulty of the task). Namely, the AUC (with the two-class ROC defined as a plot of false-negative fraction versus false-positive fraction) of a "perfect" observer is zero, and increases in some sense uniformly as the task is made more difficult, until one arrives at the "guessing" observer with an AUC of 0.5. In an N-class classification task, this straightforward relationship appears to break down, and both "perfect" and "guessing" observers yield ROC hypersurfaces with zero hypervolume. It would appear that, due to this ambiguity, hypervolume under the ROC hypersurface of an N-class observer is not a useful performance metric: Does a hypervolume of 0.005 indicate an observer faced with an exceptionally difficult or exceptionally easy task? One hopes that some other performance metric from two-class classification can be generalized usefully for N-class classification; perhaps a quantity which is equal to AUC in the two-class case has a generalization which is not equal to the hypervolume, but can be shown to be of use for other reasons.

Acknowledgment

The authors thank the associate editor and anonymous reviewers for their helpful suggestions for improving the quality of this manuscript.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] M. A. Kupinski, "Computerized Pattern Classification in Medical Imaging," Ph.D. Thesis, The University of Chicago, Chicago, IL, 2000.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] D. J. S. MacKay, "Bayesian Methods for Adaptive Models," Ph.D. Thesis, California Institute of Technology, Pasadena, CA, 1992.

[12] 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.

[13] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network," in *Proc. SPIE, Vol. 4686, Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, D. P. Chakraborty and E. A. Krupinski, Eds. Bellingham, WA, 2002, pp. 1–12.

[14] 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.

[15] 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, Jul. 2004.

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

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

[18] S. Dreiseitl, L. Ohno-Machado, and M. Binder, "Comparing three-class diagnostic tests by three-way ROC analysis," *Med. Decis. Making*, vol. 20, pp. 323–331, 2000.

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

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

[21] 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*, vol. 5032, M. Sonka and J. M. Fitzpatrick, Eds. Bellingham, WA, 2003, pp. 567–578.

[22] S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations*, 2nd ed. San Diego, CA: Harcourt Brace Jovanovich, 1986.

Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions

Darrin C. Edwards* and Charles E. Metz

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

ABSTRACT

Bayesian artificial neural networks (BANNs) have proven useful in two-class classification tasks, and are claimed to provide good estimates of ideal-observer-related decision variables (the *a posteriori* class membership probabilities). We wish to apply the BANN methodology to three-class classification tasks for computer-aided diagnosis, but we currently lack a fully general extension of two-class receiver operating characteristic (ROC) analysis to objectively evaluate three-class BANN performance. It is well known that "the likelihood ratio of the likelihood ratio is the likelihood ratio." Based on this, we found that the decision variable which is the *a posteriori* class membership probability of an observational data vector is in fact equal to the *a posteriori* class membership probability of that decision variable. Under the assumption that a BANN can provide good estimates of these *a posteriori* probabilities, a second BANN trained on the output of such a BANN should perform very similarly to an identity function. We performed a two-class and a three-class simulation study to test this hypothesis. The mean squared error (deviation from an identity function) of a two-class BANN was found to be 2.5×10^{-4}. The mean squared error of the first component of the output of a three-class BANN was found to be 2.8×10^{-4}, and that of its second component was found to be 3.8×10^{-4}. Although we currently lack a fully general method to objectively evaluate performance in a three-class classification task, circumstantial evidence suggests that two- and three-class BANNs can provide good estimates of ideal-observer-related decision variables.

Keywords: Bayesian artificial neural networks, ideal observers, three-class classification

1. INTRODUCTION

In the past, computerized methods for the detection[1–5] and classification[6–11] of mammographic mass lesions have been investigated at the University of Chicago. The classification scheme currently analyzes lesions which have been manually identified by a radiologist. We are attempting to develop a fully automated classification scheme by combining the existing detection and classification schemes; we have argued previously[12] that this will require a three-class classifier to account for the presence of false-positive (FP) computer detections, in addition to the malignant and benign lesions, in the output of the detection scheme.

For some time now we have explored the use of Bayesian artificial neural networks (BANNs) for a variety of detection[5,13,14] and classification[11] tasks in computer-aided diagnosis (CAD). Our motivation for investigating BANNs is based, first, on our theoretical observation that, in the limit of infinite training data, a BANN will yield an ideal observer decision function for that data population;[15] and second, on empirical observations that even given a finite sample of training data, a BANN can estimate an ideal observer decision function reasonably well.[16] (We note that the BANN implementation we are using is that of MacKay,[17] which employs a multivariate normal function for the prior distribution on the network weight values.) We have also performed simulation studies showing that BANNs can accurately estimate ideal observer decision variables in a three-class classification task.[15] Moreover, we showed recently that a three-class BANN could produce decision variables for actual mammographic mass lesion feature data, and that these decision variables are related to two-class BANN decision variable data in a particular way consistent with a theoretical relationship between three-class and two-class ideal observer decision variables.[12] We consider this to be strong circumstantial evidence for the ability of a BANN to estimate three-class ideal observer decision variables, though we currently lack a fully general method for evaluating three-class classifiers (*i.e.*, a three-class extension to receiver operating characteristic (ROC) analysis).

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

In this work, we present further circumstantial evidence toward the claim that a BANN can provide good estimates of three-class ideal observer decision variables. We develop a theoretical relationship between the *a posteriori* class membership probabilities of a given observational data variable and the *a posteriori* class membership probabilities of those *a posteriori* probabilities treated as a set of observational data in their own right. (It is known that *a posteriori* class membership probabilities are equivalent to ideal observer decision variables in a two-class task,[16] and related in a straightforward way to the ideal observer decision variables in a task with three or more classes.[15]) We then describe simulation studies to train and test a set of BANNs, and present results of such a simulation study verifying that the BANNs we examined did indeed obey the theoretical relationship predicted for ideal observer decision variables, to within experimental error. In the final section, we present our conclusions drawn from this work.

2. THEORY

It is well known that the ideal observer decision variable, *i.e.*, the likelihood ratio or any monotonic transformation of this value, yields optimal performance in a two-class classification task.[18] It can also be shown, in a classification task with N classes ($N > 2$), that the ideal observer decision rule becomes more complicated than a simple threshold on a single decision variable, but that the optimal decision variables remain a set of $N-1$ likelihood ratios.[18,19]

We can define the ith likelihood ratio as

$$\mathbf{\Lambda}_i \equiv \mathrm{LR}_i(\vec{\mathbf{x}}) \equiv \frac{p(\vec{\mathbf{x}}|\pi_i)}{p(\vec{\mathbf{x}}|\pi_N)}, \tag{1}$$

where $\vec{\mathbf{x}}$ represents statistically variable observational data (which we assume to have dimensionality n), and π_j represents one of the N classes from which the data are drawn (here $1 \leq i \leq N-1$). Clearly the vector (of dimensionality $N-1$) of decision variables $\mathbf{\Lambda}_i$ is itself statistically variable, and one might ask what the likelihood ratios of these variables are. In fact,[20]

$$\begin{aligned} p(\vec{\Lambda}|\pi_i) &= \int\cdots\int \sum_j \frac{p(\vec{x}_j|\pi_i)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \int\cdots\int \sum_j \mathrm{LR}_i(\vec{x}_j) \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n, \end{aligned} \tag{2}$$

where we have assumed that $N-1 < n$; if $N-1 = n$, then no integration is performed. (If $N-1 > n$, then at least one of the likelihood ratio decision variables will be expressible as a function of the others; we will not consider this degenerate case here.) The sum is over all solutions to Eq. 1 for a given $\vec{\Lambda}$; this yields

$$\begin{aligned} p(\vec{\Lambda}|\pi_i) &= \int\cdots\int \sum_j \Lambda_i \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \Lambda_i \int\cdots\int \sum_j \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \Lambda_i\, p(\vec{\Lambda}|\pi_N) \\ \frac{p(\vec{\Lambda}|\pi_i)}{p(\vec{\Lambda}|\pi_N)} &\equiv \mathrm{LR}_i(\vec{\Lambda}) = \Lambda_i, \end{aligned} \tag{3}$$

the source of the well-known adage that "the likelihood ratio of the likelihood ratio is the likelihood ratio."

Consider now a different set of decision variables, the *a posteriori* class membership probabilities considered as functions of the statistically variable observational data

$$\mathbf{y}_i \equiv P(\pi_i|\vec{\mathbf{x}}). \tag{4}$$

(Since $P(\pi_N|\vec{x}) = 1 - \sum_{i=1}^{N-1} P(\pi_i|\vec{x})$, we still have $N-1$ decision variables.) Note that in a two-class classification task, this decision variable is known to be a monotonic function of the likelihood ratio, and is therefore an ideal observer decision variable;[16] while in a classification task with more than two classes, the *a posteriori* class membership probabilities can be shown to be related to the likelihood ratios in a straightforward way.[15]

Reasoning as above, we may ask what the *a posteriori* class membership probability of these decision variables, or $P(\pi_i|\vec{y})$, is. In fact,

$$\begin{aligned} P(\pi_i|\vec{x}) &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{p(\vec{x})} \\ &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{\sum_{k=1}^{N} p(\vec{x}|\pi_k)P(\pi_k)} \\ &= \frac{\mathrm{LR}_i(\vec{x})P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \mathrm{LR}_k(\vec{x})P(\pi_k)/P(\pi_N)}, \end{aligned} \tag{5}$$

and this relation can also be inverted to yield

$$\begin{aligned} \mathrm{LR}_i(\vec{x}) &= \frac{P(\pi_i|\vec{x})}{1 - \sum_{k=1}^{N-1} P(\pi_k|\vec{x})P(\pi_k)/P(\pi_N)} \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)}. \end{aligned} \tag{6}$$

We again start with Eq. 2, this time obtaining

$$\begin{aligned} p(\vec{y}|\pi_i) &= \int\cdots\int \sum_j \mathrm{LR}_i(\vec{x}_j) \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \int\cdots\int \sum_j \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} \int\cdots\int \sum_j \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} p(\vec{y}|\pi_N), \end{aligned} \tag{7}$$

where the sums in j are over all solutions to Eq. 4 for a given $\vec{y}$. (The fraction can be taken out of the integral because the relations in Eqs. 5 and 6 are one-to-one, and thus the set of all solutions to Eq. 4 correspond to a single value of $\mathrm{LR}_i(\vec{x}_j)$.) This again yields

$$\mathrm{LR}_i(\vec{y}) = \mathrm{LR}_i(\vec{x}_j) \tag{8}$$

where $\vec{y}$ is the vector of *a posteriori* class membership probabilities of $\vec{x}$ from Eq. 4, and $\vec{x}_j$ is any solution to that equation for a given $\vec{y}$.

It follows that

$$\begin{aligned} P(\pi_i|\vec{y}) &= \frac{\mathrm{LR}_i(\vec{y})P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \mathrm{LR}_k(\vec{y})P(\pi_k)/P(\pi_N)} \\ &= \frac{\mathrm{LR}_i(\vec{x}_j)P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \mathrm{LR}_k(\vec{x}_j)P(\pi_k)/P(\pi_N)} \\ &= P(\pi_i|\vec{x}_j) = y_i, \end{aligned} \tag{9}$$

where $\vec{x}_j$ is again any solution to Eq. 4 for a given $\vec{y}$. This shows that a similar adage to that for likelihood ratios holds true, namely that "the *a posteriori* class probabilities of the (data) *a posteriori* class probabilities are the (data) *a posteriori* class probabilities."

3. MATERIALS AND METHOD

We have shown in the past[16] that a BANN can provide good estimates of the *a posteriori* class membership probabilities in a two-class classification task, and we have presented the results of simulation studies[15] and experiments with real mammographic feature data[12] strongly suggesting that the same holds true for three-class BANNs as well. The theoretical relationship given by Eq. 9, derived in the preceding section, provides a basis for another simulation study which should provide further circumstantial evidence for the claim that two-class and three-class BANNs can provide good estimates of the two- and three-class *a posteriori* class membership probabilities (directly related to the ideal observer decision variables *via* Eq. 5), respectively.

Specifically, for the two-class simulation study, we drew 500 samples pseudorandomly from each of two distributions:

$$p(x|\pi_1) \equiv N(x; \mu_1 = 1, \sigma_1^2 = 2) \tag{10}$$

$$p(x|\pi_2) \equiv N(x; \mu_2 = 0, \sigma_2^2 = 1). \tag{11}$$

We then trained a two-class BANN with one input, five hidden units, and one output on this data, obtaining a classifier we denote by

$$y = B_1^2(x). \tag{12}$$

(The superscript denotes the number of classes being classified.) We then used this output, given the known truth states for the original observations x from which it was obtained, as training data for a second BANN with one input, five hidden units, and one output:

$$z = B_2^2(y). \tag{13}$$

Finally, we pseudorandomly sampled an independent testing set of 500 observations x from each of the two classes given in Eqs. 10 and 11. This testing set was used as input to the first BANN to obtain a testing set y^{test}; this in turn was given as input to the second BANN, for which the output was z^{test}.

Given Eq. 9, together with the assumption that an adequately trained two-class BANN yields good estimates of the *a posteriori* class membership probabilities of the observations being classified, it should be the case that z^{test} estimates y^{test} at least to within experimental error. To verify this, we plotted z^{test} as a function of y^{test} for each of the two classes, and we computed the mean squared error

$$\text{MSE}_2 = \frac{1}{1000} \sum (z^{\text{test}} - y^{\text{test}})^2, \tag{14}$$

where the sum is over all the observations in the two classes.

Similarly, for the three-class simulation study, we drew 500 two-dimensional samples pseudorandomly from each of three distributions:

$$p(\vec{x}|\pi_1) \equiv N\left(\vec{x}; \vec{\mu}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 4 & .75 \times 2 \\ .75 \times 2 & 1 \end{bmatrix}\right) \tag{15}$$

$$p(\vec{x}|\pi_2) \equiv N\left(\vec{x}; \vec{\mu}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & -.4 \times 1.5 \\ -.4 \times 1.5 & 2.25 \end{bmatrix}\right) \tag{16}$$

$$p(\vec{x}|\pi_3) \equiv N\left(\vec{x}; \vec{\mu}_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{17}$$

We then trained a three-class BANN with two inputs, five hidden units, and two outputs on this data, obtaining a classifier we denote by

$$\vec{y} = B_1^3(\vec{x}). \tag{18}$$

We then used this output, given the known truth states for the original observations $\vec{x}$ from which it was obtained, as training data for a second BANN with two inputs, five hidden units, and two outputs:

$$\vec{z} = B_2^3(\vec{y}). \tag{19}$$

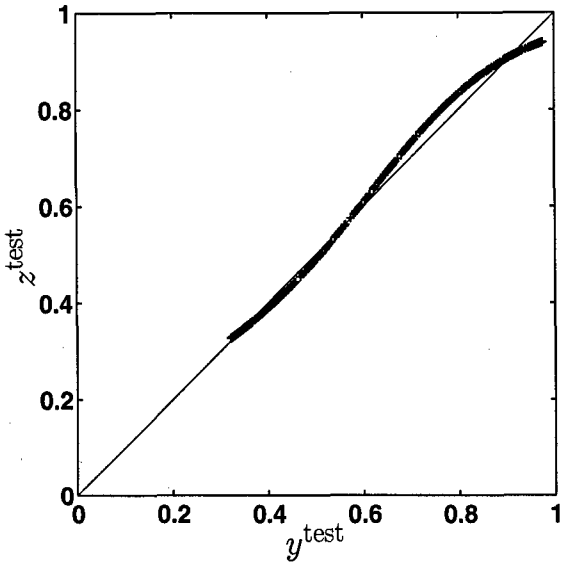


Figure 1. Output of the second two-class BANN as a function of its input for the observations actually drawn from class π_1 in the two-class simulation study.

Finally, we pseudorandomly sampled an independent testing set of 500 observations $\vec{x}$ from each of the three classes given in Eqs. 15-17. This testing set was used as input to the first BANN to obtain a testing set $\vec{y}^{\,\text{test}}$; this in turn was given as input to the second BANN, for which the output was $\vec{z}^{\,\text{test}}$.

Again, given Eq. 9, together with the assumption that an adequately trained two-class BANN yields good estimates of the *a posteriori* class membership probabilities of the observations being classified, it should be the case that z_1^{test} estimates y_1^{test}, and z_2^{test} estimates y_2^{test}, at least to within experimental error. To verify this, we plotted z_1^{test} as a function of y_1^{test}, and z_2^{test} as a function of y_2^{test}, for each of the three classes, and we computed the mean squared errors

$$\text{MSE}_{3i} = \frac{1}{1500}\sum(z_i^{\text{test}} - y_i^{\text{test}})^2, \tag{20}$$

$\{i : 1, 2\}$, where the sum is over all the observations in the three classes.

4. RESULTS

Figure 1 shows z^{test} as a function of y^{test} for the observations in class π_1, and Fig. 2 shows z^{test} as a function of y^{test} for the observations in class π_2 from the two-class simulation study. The mean squared error for the complete set of 1000 observations was 2.5×10^{-4}.

Figure 3 shows the components of $\vec{z}^{\,\text{test}}$ as a function of the corresponding components of $\vec{y}^{\,\text{test}}$ for the observations in class π_1. Similarly Fig. 4 shows the components of $\vec{z}^{\,\text{test}}$ as a function of the corresponding components of $\vec{y}^{\,\text{test}}$ for the observations in class π_2, and Fig. 5 shows the components of $\vec{z}^{\,\text{test}}$ as a function of the corresponding components of $\vec{y}^{\,\text{test}}$ for the observations in class π_3. The mean squared error for the complete set of 1500 observations was 2.8×10^{-4} for the first component and 3.8×10^{-4} for the second component.

5. DISCUSSION AND CONCLUSIONS

We developed a theoretical relationship between the *a posteriori* class membership probabilities, directly related to ideal observer decision variables, and the *a posteriori* class membership probabilities of those *a posteriori* class membership probabilities treated as statistically variable observer data in their own right. The identity relationship found is, perhaps unsurprisingly, quite similar in spirit to the identity relationship between the likelihood ratio decision variables and the likelihood ratio of those likelihood ratio decision variables for a given task.

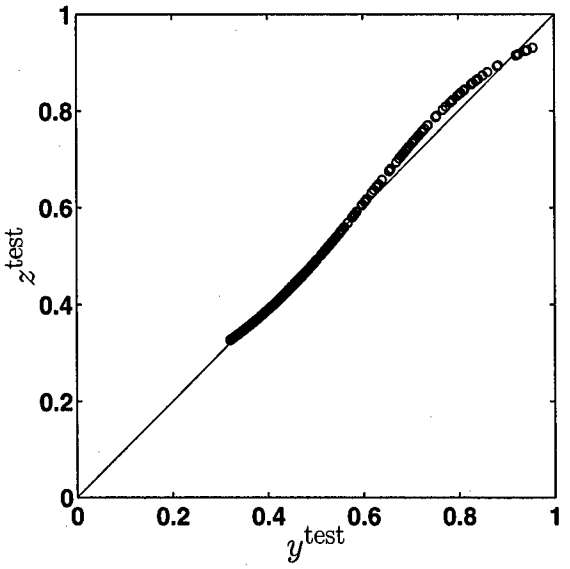
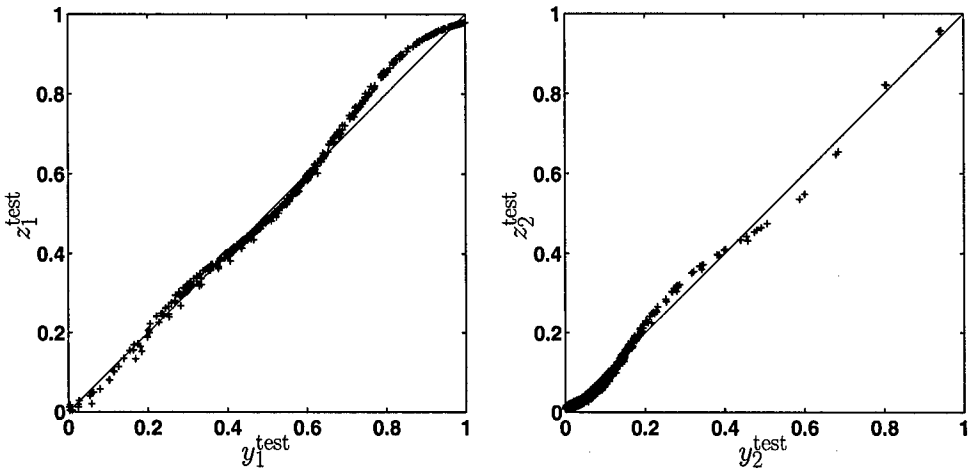
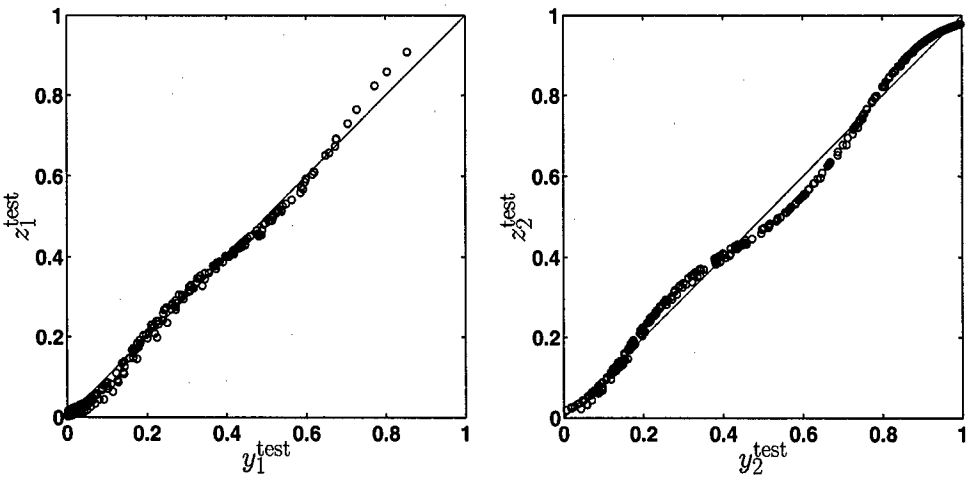


Figure 2. Output of the second two-class BANN as a function of its input for the observations actually drawn from class π_2 in the two-class simulation study.



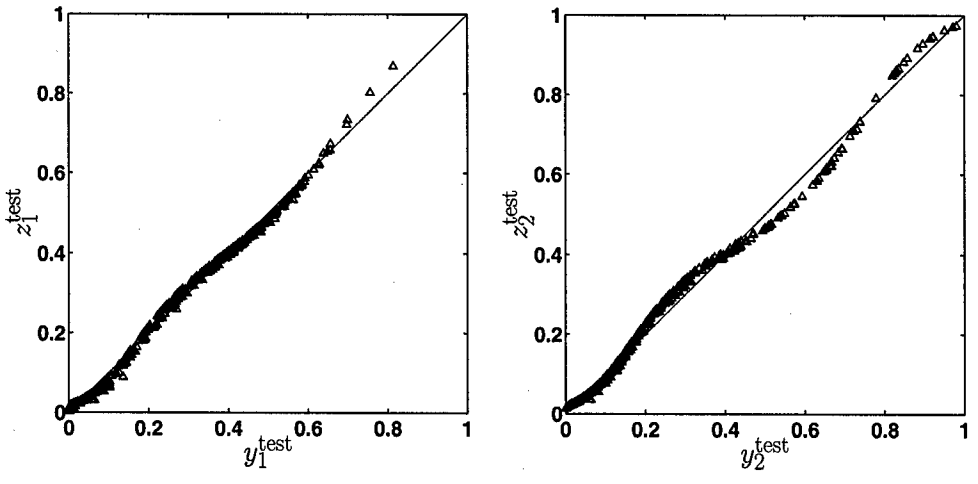
(a) (b)

Figure 3. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_1 in the three-class simulation study.



(a) (b)

Figure 4. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_2 in the three-class simulation study.



(a) (b)

Figure 5. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_3 in the three-class simulation study.

We currently lack a fully general method for three-class classification or for practically evaluating the performance of a three-class classifier. As a first step toward such a classification method, we are investigating the use of BANNs to estimate three-class ideal observer decision variables for such a task. Since, in a practical situation, we will not have access to the underlying probability distributions from which the observational data are drawn, we must rely on circumstantial evidence in support of our claim that a three-class BANN can adequately estimate decision variables directly related to ideal observer decision variables.

Previously, we presented work relating the output of a three-class BANN to the outputs of two-class BANNs trained for various "simplified" cases in which the three-class classification task was reduced to a two-class classification task, and showed that the relationships found were consistent with the relationship between three- and two-class ideal observers for the same tasks.[12] In the present work, we showed that the output of two- and three-class BANNs was consistent, to within experimental error, with the theoretical relationship developed for actual *a posteriori* class membership probabilities. This is of limited practical use in the complete development of a three-class classifier, mainly because the three-class ideal observer decision rule is considerably more complicated than its two-class counterpart (a simple threshold on a single decision variable). It does, however, bolster our confidence in the choice of the BANN as an appropriate tool for estimating the decision variables which would eventually be incorporated in such a classifier.

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. U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.* **2**, pp. 1–9, 1995.
2. 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.* **18**, pp. 955–963, 1991.
3. 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.* **28**, pp. 473–481, 1993.
4. 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.* **21**, pp. 445–452, 1994.
5. M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging.* Ph.D. thesis, The University of Chicago, Chicago, IL, 2000.
6. 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.* **5**, pp. 155–168, 1998.
7. 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.* **44**, pp. 2579–2595, 1999.
8. 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.* **7**, pp. 1077–1084, 2000.
9. 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.* **20**, pp. 1285–1292, 2001.
10. 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* **224**, pp. 560–568, 2002.

11. Z. Huo and M. L. Giger, "Effect of case mix on feature selection in the computerized classification of mammographic lesions," in Proc. SPIE Vol. 4684 *Medical Imaging 2002: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 762–767, (SPIE, Bellingham, WA), 2002.
12. 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.
13. D. C. Edwards, M. A. Kupinski, R. H. Nagel, R. M. Nishikawa, and J. Papaioannou, "Using a Bayesian neural network to optimally eliminate false-positive microcalcification detections in a CAD scheme," in *IWDM 2000: 5th International Workshop on Digital Mammography*, M. J. Yaffe, ed., *Proceedings of the Workshop*, pp. 168–173, Medical Physics Publishing, (Madison, WI), 2001.
14. D. C. Edwards, J. Papaioannou, Y. Jiang, M. A. Kupinski, and R. M. Nishikawa, "Eliminating false-positive microcalcification clusters in a mammography CAD scheme using a Bayesian neural network," in Proc. SPIE Vol. 4322 *Medical Imaging 2001: Image Processing*, Milan Sonka and Kenneth Hanson, eds., pp. 1954–1960, (SPIE, Bellingham, WA), 2001.
15. D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network," in Proc. SPIE Vol. 4686 *Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, Dev P. Chakraborty and Elizabeth A. Krupinski, eds., pp. 1–12, (SPIE, Bellingham, WA), 2002.
16. 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.* **20**, pp. 886–899, 2001.
17. D. J. S. MacKay, *Bayesian Methods for Adaptive Models.* Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.
18. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
19. 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.
20. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., New York, 1991.

Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule

Darrin C. Edwards* and Charles E. Metz

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

ABSTRACT

We analyzed a variety of recently proposed decision rules for three-class classification from the point of view of ideal observer decision theory. We considered three-class decision rules which have been proposed recently: one by Scurfield, one by Chan *et al.*, and one by Mossman. Scurfield's decision rule can be shown to be a special case of the three-class ideal observer decision rule in two different situations: when the pair of decision variables is the pair of likelihood ratios used by the ideal observer, and when the pair of decision variables is the pair of logarithms of the likelihood ratios. Chan *et al.* start with an ideal observer model, where two of the decision lines used by the ideal observer overlap, and the third line becomes undefined. Finally, we showed that the Mossman decision rule (in which a single decision line separates one class from the other two, while a second line separates those two classes) cannot be a special case of the ideal observer decision rule. Despite the considerable difficulties presented by the three-class classification task compared with two-class classification, we found that the three-class ideal observer provides a useful framework for analyzing a wide variety of three-class decision strategies.

Keywords: 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,[1–5] and one for classifying known lesions as malignant or benign.[6–10] 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.

The explicit form of the ideal observer in a three-class classification task has been known for some time.[11] For the reasons just stated, however, a practical method for estimating and evaluating observer performance based on an ideal observer model has proven elusive, despite the success of the two-class binormal ideal observer model.[12] Nevertheless, pragmatic 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.

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

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,[13] one by Chan *et al.*,[14] and one by Mossman.[15] 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 in the works discussed; we are specifically interested only in the form of the decision rule which serves as the starting point for each work. The results of our analyses are briefly summarized in Sec. 6.

2. THE THREE-CLASS IDEAL OBSERVER

It can be shown[11,16] 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:

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

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

and

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

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

$$\mathrm{LR}_i \equiv \frac{p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_i)}{p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_N)} \tag{3}$$

for $i < N$. 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.)

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

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

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 Eq. 5).

Finally, note that the hyperplanes represented by Eqs. 1 and 2 are unchanged if we multiply all of these equations 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.

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

$$\gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 = \gamma_{313} - \gamma_{323} \tag{7}$$

$$\gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 = \gamma_{313} \tag{8}$$

$$(\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 = \gamma_{323}, \tag{9}$$

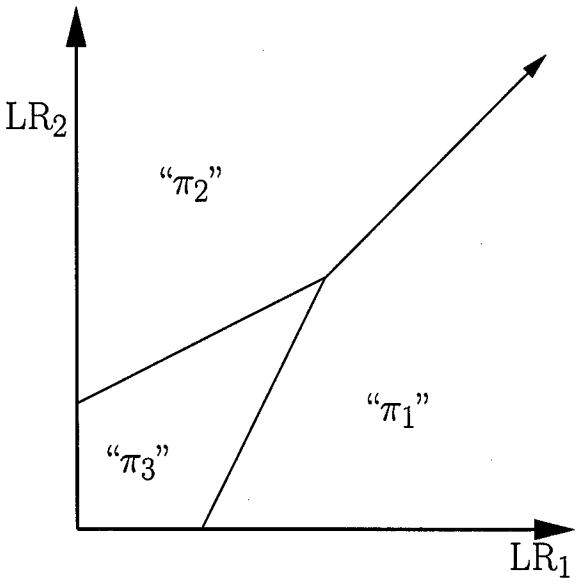


Figure 1. Example three-class ideal observer decision rule, given the values of the decision parameters $\gamma_{121} = \gamma_{212} = 3/14$ and $\gamma_{131} = \gamma_{313} = \gamma_{232} = \gamma_{323} = 1/7$. Note $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\mathbf{t} = \pi_k)$.

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 Eq. 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).

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

$$\frac{3}{14}\mathrm{LR}_1 - \frac{3}{14}\mathrm{LR}_2 = 0 \quad \{\text{"1-}vs.\text{-2"}\} \tag{10}$$

$$\frac{1}{7}\mathrm{LR}_1 - \frac{1}{14}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{"1-}vs.\text{-3"}\} \tag{11}$$

$$-\frac{1}{14}\mathrm{LR}_1 + \frac{1}{7}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{"2-}vs.\text{-3"}\}. \tag{12}$$

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.[13] 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.[17] We will not review that prior work here, because at present we are interested in relating given observer models to the three-class ideal observer model for multivariate observational data, which yield two-dimensional decision variable data by Eq. 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.

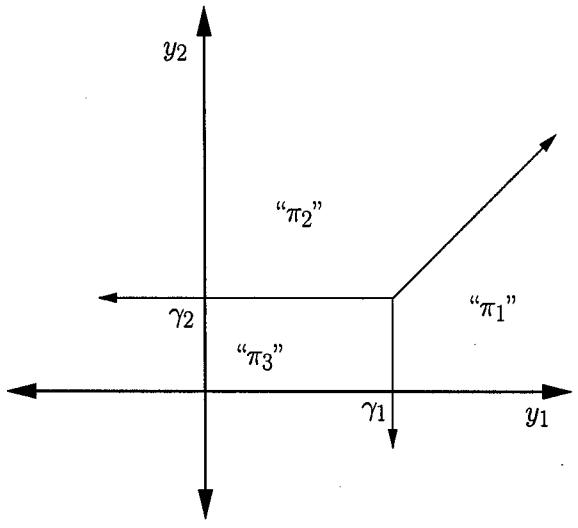


Figure 2. Decision rule investigated by Scurfield, for the decision parameters γ_1 and γ_2.

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

$$\text{decide} \quad d = \pi_1 \quad \text{iff} \quad y_1 - y_2 \geq \gamma_1 - \gamma_2 \quad \text{and} \quad y_1 \geq \gamma_1; \tag{13}$$

$$\text{decide} \quad d = \pi_2 \quad \text{iff} \quad y_1 - y_2 < \gamma_1 - \gamma_2 \quad \text{and} \quad y_2 \geq \gamma_2; \tag{14}$$

$$\text{decide} \quad d = \pi_3 \quad \text{iff} \quad y_1 < \gamma_1 \quad \text{and} \quad y_2 < \gamma_2. \tag{15}$$

This decision rule is illustrated in Fig. 2.

From these relations, one can define the decision boundary lines

$$y_1 - y_2 = \gamma_1 - \gamma_2 \quad \{\text{"1-}vs\text{.-2"}\} \tag{16}$$

$$y_1 = \gamma_1 \quad \{\text{"1-}vs\text{.-3"}\} \tag{17}$$

$$y_2 = \gamma_2 \quad \{\text{"2-}vs\text{.-3"}\}. \tag{18}$$

Note the similarity in form between these equations and Eqs. 7-9. 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 a special case of Eqs. 7-9, which is illustrated in Fig. 3.

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 $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$; note that a line of the form $\log(\mathrm{LR}_2) = \log(\mathrm{LR}_1) + \alpha$ corresponds to a line of the form $\mathrm{LR}_2 = \beta \mathrm{LR}_1$ for appropriate constants α and β. By inspection, this is again a special case of Eqs. 7-9, which is illustrated in Fig. 4.

Scurfield points out[13] 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.

4. THE CHAN DECISION RULE

Chan *et al.* are investigating three-class classifiers for computer-aided diagnosis.[14] 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,

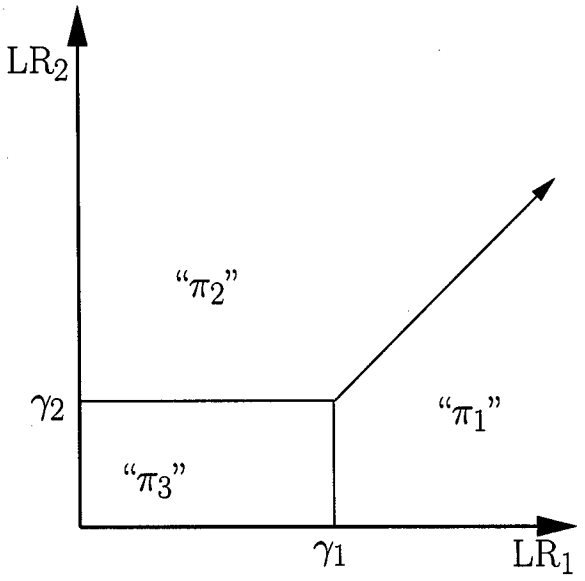


Figure 3. A special case of the ideal observer decision rule, which is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$.

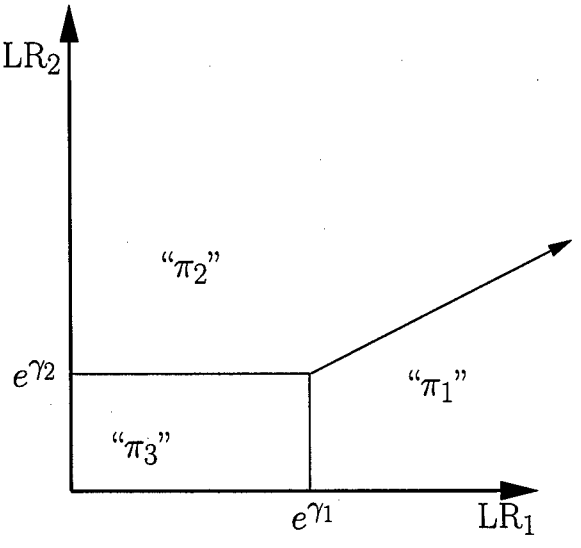


Figure 4. A special case of the ideal observer decision rule which is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$.

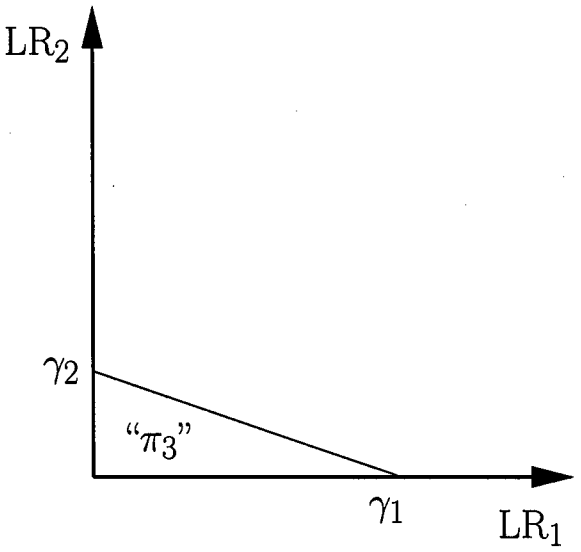


Figure 5. The decision rule investigated by Chan *et al.*, which as they state is a special case of the ideal observer decision rule. Observations in the unlabelled region are decided "not π_3", *i.e.*, either "π_1" or "π_2".

or normal. Because the goal of their work is to optimize the performance of a system to aid a radiologist or clinician, rather than to measure the psychophysical performance of an existing observer, they choose 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 labelling 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 Eq. 4, we obtain decision boundary lines of the form

$$0\,\mathrm{LR}_1 + 0\,\mathrm{LR}_2 = 0 \qquad \{\text{"1-}vs\text{.-2"}\} \tag{19}$$

$$\frac{(1-U_{3|1})P(\mathbf{t}=\pi_1)}{\alpha}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\mathbf{t}=\pi_2)}{\alpha}\mathrm{LR}_2 = \frac{P(\mathbf{t}=\pi_3)}{\alpha} \qquad \{\text{"1-}vs\text{.-3"}\} \tag{20}$$

$$\frac{(1-U_{3|1})P(\mathbf{t}=\pi_1)}{\alpha}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\mathbf{t}=\pi_2)}{\alpha}\mathrm{LR}_2 = \frac{P(\mathbf{t}=\pi_3)}{\alpha} \qquad \{\text{"2-}vs\text{.-3"}\} \tag{21}$$

where $\alpha \equiv 1 + P(\mathbf{t}=\pi_3) - U_{3|1}P(\mathbf{t}=\pi_1) - U_{3|2}P(\mathbf{t}=\pi_2)$. 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 Eqs. 7-9; if any two of these equations yield identical lines, the third line must be either identical to them or undefined. The decision rule considered by Chan *et al.* is illustrated in Fig. 5.

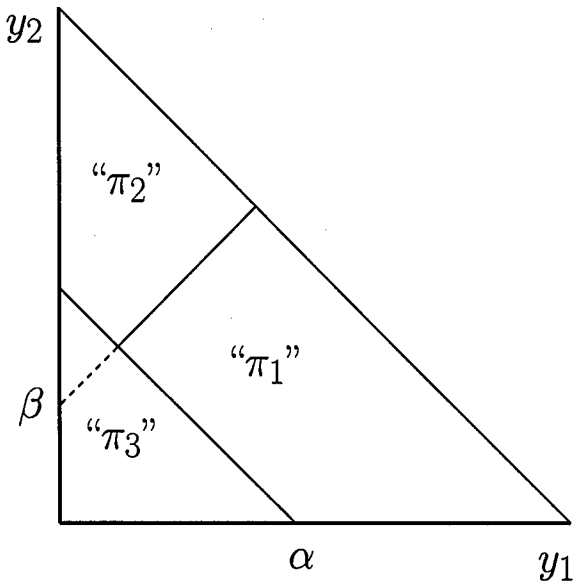


Figure 6. Decision rule investigated by Mossman, for the decision parameters α and β, shown in the *a posteriori* class probability space.

5. THE MOSSMAN DECISION RULE

Mossman investigates 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

$$\mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 = 1, \tag{22}$$

as well as $0 \leq \mathbf{y}_i \leq 1 \ \{1 \leq i \leq 3\}$. 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.[18,19] (In this section we will write $P(\pi_i|\vec{x})$ instead of $P(\mathbf{t} = \pi_i|\vec{x})$ for simplicity.) Mossman does not explicitly require, however, that the decision variables in Eq. 22 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 α and β, is

$$\text{decide} \quad d = \pi_1 \quad \text{iff} \quad y_2 - y_1 \leq \beta \quad \text{and} \quad y_3 \leq \alpha; \tag{23}$$

$$\text{decide} \quad d = \pi_2 \quad \text{iff} \quad y_2 - y_1 > \beta \quad \text{and} \quad y_3 \leq \alpha; \tag{24}$$

$$\text{decide} \quad d = \pi_3 \quad \text{iff} \quad y_3 > \alpha. \tag{25}$$

where $0 \leq \alpha \leq 1$ and $-1 \leq \beta \leq 1$. From these relations, and given the relation $y_3 = 1 - y_1 - y_2$ from Eq. 22, one can define the decision boundary lines

$$y_1 - y_2 = -\beta \quad \{\text{"1-}vs\text{.-2"}\} \tag{26}$$

$$y_1 + y_2 = 1 - \alpha \quad \{\text{"1-}vs\text{.-3"}\} \tag{27}$$

$$y_1 + y_2 = 1 - \alpha \quad \{\text{"2-}vs\text{.-3"}\}. \tag{28}$$

This decision rule is illustrated in Fig. 6. 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 $\vec{\mathbf{x}}$. This version of the decision rule is illustrated in Fig. 7.

Although the Mossman decision rule 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

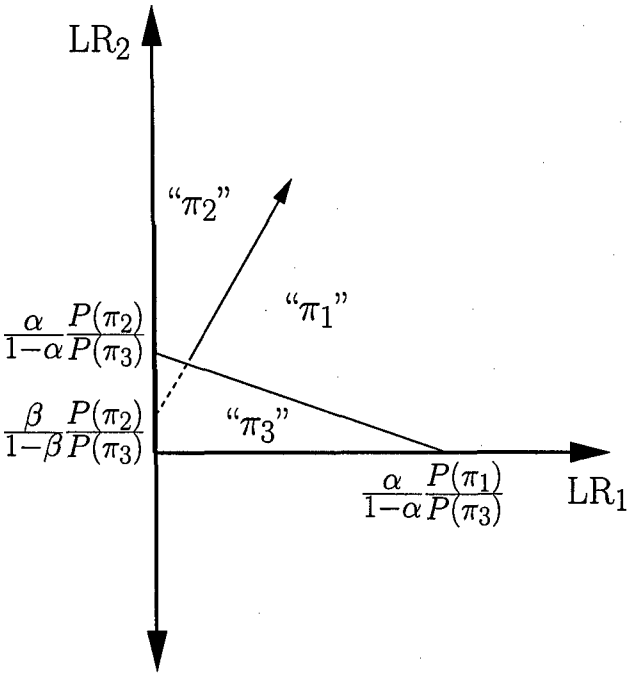


Figure 7. Decision rule investigated by Mossman, for the decision parameters α and β, shown in likelihood ratio space.

the first two or be undefined. Another way to see this is to note that the coefficients of Eq. 9 are differences of the corresponding coefficients of Eqs. 7 and 8. If the coefficients of Eqs. 8 and 9 are identical, it must be the case that the coefficients of Eq. 7 are all zero. For the Mossman decision rule, this would require $1+\beta = 0$, $1-\beta = 0$, and $\beta = 0$ simultaneously, which is clearly impossible. It follows that the decision rule considered by Mossman cannot represent possible ideal observer performance for any choice of the utilities $U_{i|j}$ in Eqs. 1 and 2.

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.

Although this is not the most general approach to three-class classification, 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.

In particular, Scurfield points out[13] 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.[14] Although this rendered our analysis of that decision rule in terms of ideal observer decision theory largely trivial, it provided an intuitive basis for understanding the results of similar analysis of the Mossman decision rule, namely the conclusion that the latter does not correspond to ideal observer behavior for any possible values of the utilities used by the ideal observer. 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.

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.

This work was supported in parts 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. U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.* **2**, pp. 1–9, 1995.
2. 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.* **18**, pp. 955–963, 1991.
3. 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.* **28**, pp. 473–481, 1993.
4. 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.* **21**, pp. 445–452, 1994.
5. M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging.* Ph.D. thesis, The University of Chicago, Chicago, IL, 2000.
6. 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.* **5**, pp. 155–168, 1998.
7. 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.* **44**, pp. 2579–2595, 1999.
8. 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.* **7**, pp. 1077–1084, 2000.
9. 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.* **20**, pp. 1285–1292, 2001.
10. 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* **224**, pp. 560–568, 2002.
11. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
12. C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, pp. 1–33, 1999.
13. 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.
14. 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.
15. D. Mossman, "Three-way ROCs," *Med. Decis. Making* **19**, pp. 78–89, 1999.
16. 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.

17. B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math Psychol.* **40**, pp. 253–269, 1996.
18. 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.* **20**, pp. 886–899, 2001.
19. 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.

APPENDIX D

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-consquence 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 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, Nishikawa, Wolverton, and Doi, 1995; Yin, Giger, Doi, Metz, Vyborny, and

* 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).

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.

The explicit form of 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 method for estimating and evaluating observer performance based on an ideal observer model has proven elusive, despite the success of the two-class binormal ideal observer model (Metz and Pan, 1999). Nevertheless, pragmatic 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 proposed performance evaluation metrics in the works discussed; we are specifically interested only in the form of the decision rule which serves as the starting point for each work. 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 $i < N$. 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

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

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

$$\gamma_{ijk} = \gamma_{kjk} - \gamma_{kik}. \quad (5)$$

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.

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

$$\gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 = \gamma_{313} - \gamma_{323} \tag{7}$$

$$\gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 = \gamma_{313} \tag{8}$$

$$(\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 = \gamma_{323}, \tag{9}$$

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).

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

$$\frac{3}{14}\mathrm{LR}_1 - \frac{3}{14}\mathrm{LR}_2 = 0 \quad \{\text{"1-}vs.\text{-2"}\} \tag{10}$$

$$\frac{1}{7}\mathrm{LR}_1 - \frac{1}{14}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{"1-}vs.\text{-3"}\} \tag{11}$$

$$-\frac{1}{14}\mathrm{LR}_1 + \frac{1}{7}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{"2-}vs.\text{-3"}\}. \tag{12}$$

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.

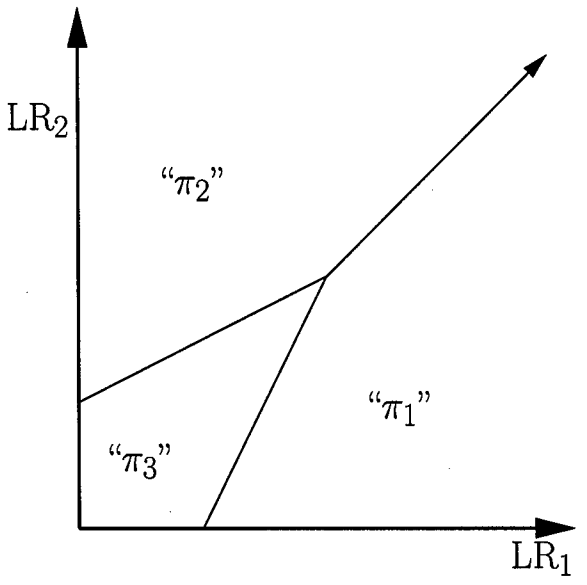


Fig. 1. Example three-class ideal observer decision rule, given the values of the decision parameters $\gamma_{121} = \gamma_{212} = 3/14$ and $\gamma_{131} = \gamma_{313} = \gamma_{232} = \gamma_{323} = 1/7$. Note that $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\mathbf{t} = \pi_i)$.

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 three-class ideal observer model for multivariate observational data, which 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

$$\text{decide} \quad d = \pi_1 \quad \text{iff} \; y_1 - y_2 \geq \gamma_1 - \gamma_2 \quad \text{and} \quad y_1 \geq \gamma_1; \tag{13}$$

$$\text{decide} \quad d = \pi_2 \quad \text{iff} \; y_1 - y_2 < \gamma_1 - \gamma_2 \quad \text{and} \quad y_2 \geq \gamma_2; \tag{14}$$

$$\text{decide} \quad d = \pi_3 \quad \text{iff} \quad y_1 < \gamma_1 \quad \text{and} \quad y_2 < \gamma_2. \tag{15}$$

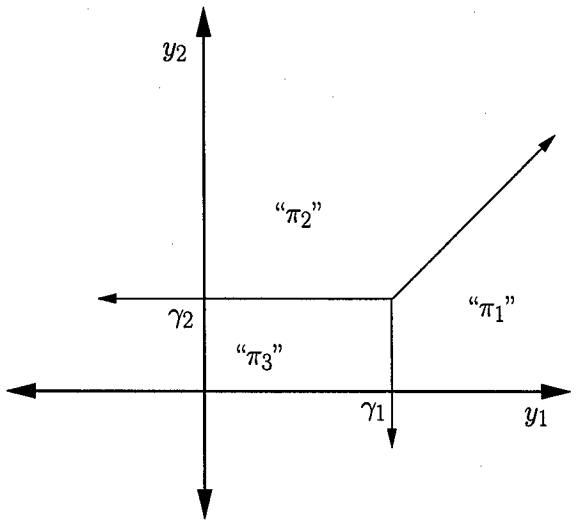


Fig. 2. Decision rule investigated by Scurfield, for the decision parameters γ_1 and γ_2.

This decision rule is illustrated in Fig. 2.

From these relations, one can define the decision boundary lines

$$y_1 - y_2 = \gamma_1 - \gamma_2 \quad \{\text{"1-}vs\text{.-2"}\} \tag{16}$$

$$y_1 = \gamma_1 \quad \{\text{"1-}vs\text{.-3"}\} \tag{17}$$

$$y_2 = \gamma_2 \quad \{\text{"2-}vs\text{.-3"}\}. \tag{18}$$

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

$$\frac{1}{\gamma_0}\mathrm{LR}_1 - \frac{1}{\gamma_0}\mathrm{LR}_2 = \frac{\gamma_1 - \gamma_2}{\gamma_0} \quad \{\text{"1-}vs\text{.-2"}\} \tag{19}$$

$$\frac{1}{\gamma_0}\mathrm{LR}_1 = \frac{\gamma_1}{\gamma_0} \quad \{\text{"1-}vs\text{.-3"}\} \tag{20}$$

$$\frac{1}{\gamma_0}\mathrm{LR}_2 = \frac{\gamma_2}{\gamma_0} \quad \{\text{"2-}vs\text{.-3"}\}, \tag{21}$$

where $\gamma_0 \equiv \gamma_1 + \gamma_2 + 4$. Note the similarity in form between these equations and (7)-(9). If we require γ_1 and γ_2 to be positive, the correspondence is exact, and this special case of (7)-(9) 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.)

A second correspondence between Scurfield's decision rule and the ideal ob-

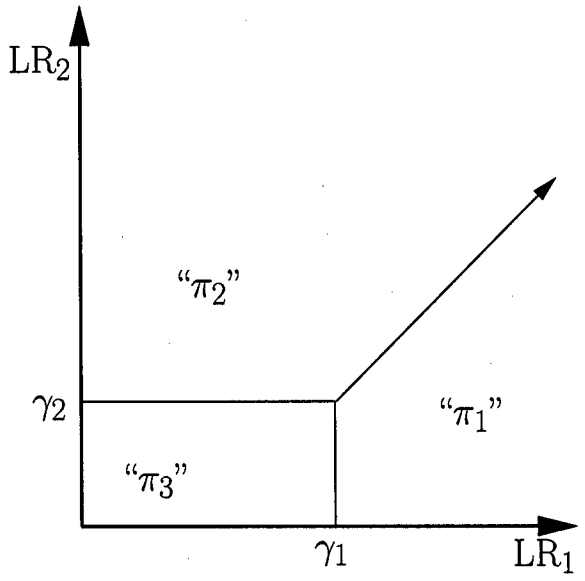


Fig. 3. A special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{212} = \gamma_{131} = \gamma_{232} = 1/(\gamma_1 + \gamma_2 + 4)$, $\gamma_{313} = \gamma_1/(\gamma_1 + \gamma_2 + 4)$, and $\gamma_{323} = \gamma_2/(\gamma_1 + \gamma_2 + 4)$. The parameters γ_1 and γ_2 are positive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

server 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 (16)-(18), we obtain

$$\log(\mathrm{LR}_1) - \log(\mathrm{LR}_2) = \gamma_1 - \gamma_2 \quad \{\text{"1-}vs\text{.-2"}\} \tag{22}$$
$$\log(\mathrm{LR}_1) = \gamma_1 \quad \{\text{"1-}vs\text{.-3"}\} \tag{23}$$
$$\log(\mathrm{LR}_2) = \gamma_2 \quad \{\text{"2-}vs\text{.-3"}\}. \tag{24}$$

Taking exponentials on each side of these equations then gives

$$\frac{\mathrm{LR}_1}{\mathrm{LR}_2} = e^{\gamma_1 - \gamma_2} \quad \{\text{"1-}vs\text{.-2"}\} \tag{25}$$
$$\mathrm{LR}_1 = e^{\gamma_1} \quad \{\text{"1-}vs\text{.-3"}\} \tag{26}$$
$$\mathrm{LR}_2 = e^{\gamma_2} \quad \{\text{"2-}vs\text{.-3"}\}; \tag{27}$$

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

$$\frac{e^{-\gamma_1}}{\gamma_0}\mathrm{LR}_1 - \frac{e^{-\gamma_2}}{\gamma_0}\mathrm{LR}_2 = 0 \quad \{\text{"1-}vs\text{.-2"}\} \tag{28}$$
$$\frac{e^{-\gamma_1}}{\gamma_0}\mathrm{LR}_1 = \frac{1}{\gamma_0} \quad \{\text{"1-}vs\text{.-3"}\} \tag{29}$$

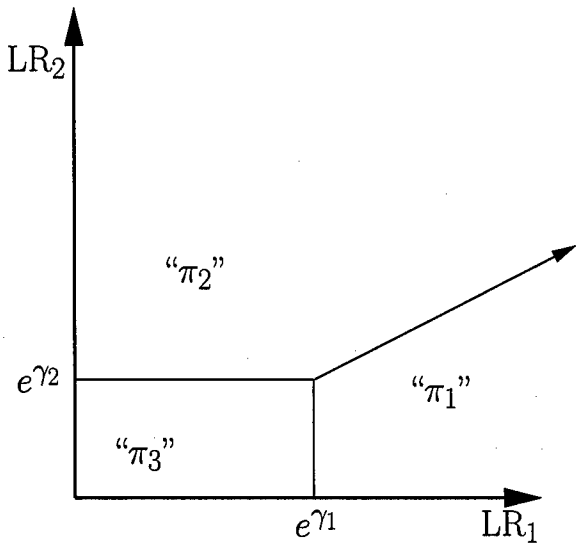


Fig. 4. A special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{131} = e^{-\gamma_1}/\gamma_0$, $\gamma_{212} = \gamma_{232} = e^{-\gamma_1}/\gamma_0$, $\gamma_{313} = \gamma_{323} = 1/\gamma_0$, and $\gamma_0 \equiv 2(e^{-\gamma_1} + e^{-\gamma_2} + 1)$. The parameters γ_1 and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

$$\frac{e^{-\gamma_2}}{\gamma_0}\mathrm{LR}_2 = \frac{1}{\gamma_0} \quad \{\text{"2-}vs.\text{-3"}\}, \tag{30}$$

where $\gamma_0 \equiv 2(e^{-\gamma_1} + e^{-\gamma_2} + 1)$. By inspection, this is again a special case of (7)-(9), which is illustrated in Fig. 4.

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

$$P(\pi_1|\vec{x}) - P(\pi_2|\vec{x}) = \gamma_1 - \gamma_2 \quad \{\text{"1-}vs.\text{-2"}\} \tag{31}$$
$$P(\pi_1|\vec{x}) = \gamma_1 \quad \{\text{"1-}vs.\text{-3"}\} \tag{32}$$
$$P(\pi_2|\vec{x}) = \gamma_2 \quad \{\text{"2-}vs.\text{-3"}\}, \tag{33}$$

as illustrated in Fig. 5.

Note that (3) can be written as

$$\mathrm{LR}_i = \frac{P(\pi_i|\vec{x})p(\vec{x})/P(\pi_i)}{p(\vec{x}|\pi_3)} \qquad \{i : 1 \le i \le 2\}$$
$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i[P(\pi_i)/P(\pi_3)]}{p(\vec{x})/p(\vec{x}|\pi_3)}$$
$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i[P(\pi_i)/P(\pi_3)]}{1 + \mathrm{LR}_1[P(\pi_1)/P(\pi_3)] + \mathrm{LR}_2[P(\pi_2)/P(\pi_3)]}. \tag{34}$$

This allows us to rewrite (31)-(33) as

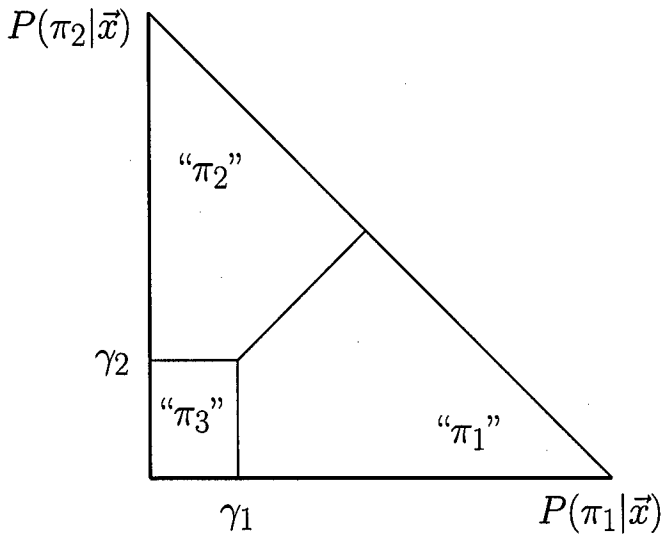


Fig. 5. A special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv P(\pi_1|\vec{x})$ and $\mathbf{y}_2 \equiv P(\pi_2|\vec{x})$.

$$\frac{1-(\gamma_1-\gamma_2)}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 - \frac{1+(\gamma_1-\gamma_2)}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_1-\gamma_2}{\gamma_0} \tag{35}$$

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

$$-\frac{\gamma_2}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 + \frac{1-\gamma_2}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_2}{\gamma_0}, \tag{37}$$

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 (7)-(9), 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 points out (Scurfield, 1998) 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 (16)-(18), 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, which together form what is known as the symmetric group on three letters (Clark, 1984). 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. One can, however, consider an alternative formulation motivated strictly by economy rather than elegance of interpretation. From this point of view, we claim that only four such surfaces are necessary to completely characterize the observer's performance. 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

$$P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_1) = 1 - P(\mathbf{d} = \pi_2|\mathbf{t} = \pi_1) - P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_1) \quad (38)$$
$$P(\mathbf{d} = \pi_2|\mathbf{t} = \pi_2) = 1 - P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_2) - P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_2) \quad (39)$$
$$P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3) = 1 - P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_3) - P(\mathbf{d} = \pi_2|\mathbf{t} = \pi_3). \quad (40)$$

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 (16)-(18)), one can show that a set of $(N-1)^2$ hypersurfaces with $N-1$ degrees of freedom in N-dimensional ROC spaces is sufficient to fully characterize the observer's performance, rather than the set of $N!$ 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. Because the goal of their work is to optimize the performance of a system to aid a radiologist or clinician, rather than to measure the psychophysical performance of an existing observer, they choose 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 labelling 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

$$0\,\mathrm{LR}_1 + 0\,\mathrm{LR}_2 = 0 \qquad \{\text{"1-}vs.\text{-2"}\} \quad (41)$$

$$\frac{(1-U_{3|1})P(\pi_1)}{\gamma_0}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\pi_2)}{\gamma_0}\mathrm{LR}_2 = \frac{P(\pi_3)}{\gamma_0} \qquad \{\text{"1-}vs.\text{-3"}\} \quad (42)$$

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

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 (7)-(9); if any two of these equations yield identical lines, the third line must be either identical to them or 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.)

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 (41) 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

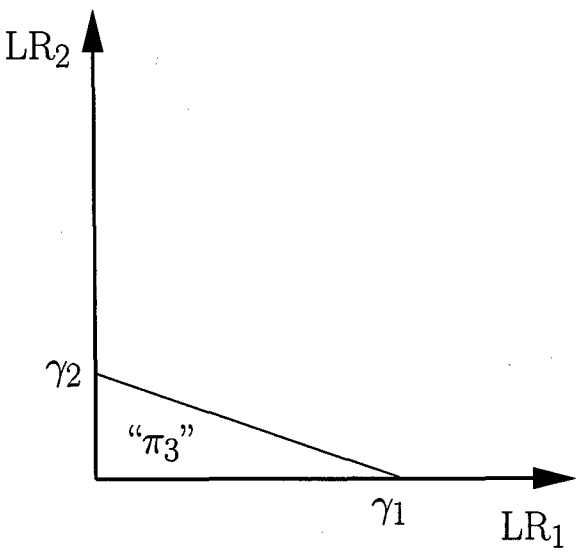


Fig. 6. The decision rule investigated by Chan et al., which is a special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{212} = 0$, $\gamma_{131} = (1 - U_{3|1})P(\pi_1)/\gamma_0$, $\gamma_{232} = (1 - U_{3|2})P(\pi_2)/\gamma_0$, and $\gamma_{313} = \gamma_{323} = P(\pi_3)/\gamma_0$; here $\gamma_0 \equiv 1 + P(\pi_3) - U_{3|1}P(\pi_1) - U_{3|2}P(\pi_2)$. Observations in the unlabelled 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.

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, unlike the case for the Scurfield decision rule (and for the Mossman decision rule, as shown in the next section), this single two-dimensional surface is sufficient to completely characterize the performance 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:

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_1) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_1) = 1 \tag{44}$$

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_2) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_2) = 1 \tag{45}$$

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

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 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

$$\mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 = 1, \tag{47}$$

as well as $0 \leq \mathbf{y}_i \leq 1 \;\; \{1 \leq i \leq 3\}$. 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 (47) 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

$$\begin{aligned}
&\text{decide} \quad d = \pi_1 \quad \text{iff} \; y_2 - y_1 \leq \gamma_2 \; \text{and} \quad y_3 \leq \gamma_1; && (48)\\
&\text{decide} \quad d = \pi_2 \quad \text{iff} \; y_2 - y_1 > \gamma_2 \; \text{and} \quad y_3 \leq \gamma_1; && (49)\\
&\text{decide} \quad d = \pi_3 \quad \text{iff} \quad\quad\quad\quad\quad\quad\quad\quad y_3 > \gamma_1. && (50)
\end{aligned}$$

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 (47), one can define the decision boundary lines

$$\begin{aligned}
&y_1 - y_2 = -\gamma_2 \quad \{\text{“1-}vs\text{.-2”}\} && (51)\\
&y_1 + y_2 = 1 - \gamma_1 \quad \{\text{“1-}vs\text{.-3”}\} && (52)\\
&y_1 + y_2 = 1 - \gamma_1 \quad \{\text{“2-}vs\text{.-3”}\}. && (53)
\end{aligned}$$

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 $\vec{\mathbf{x}}$. Note that (3) can be written as

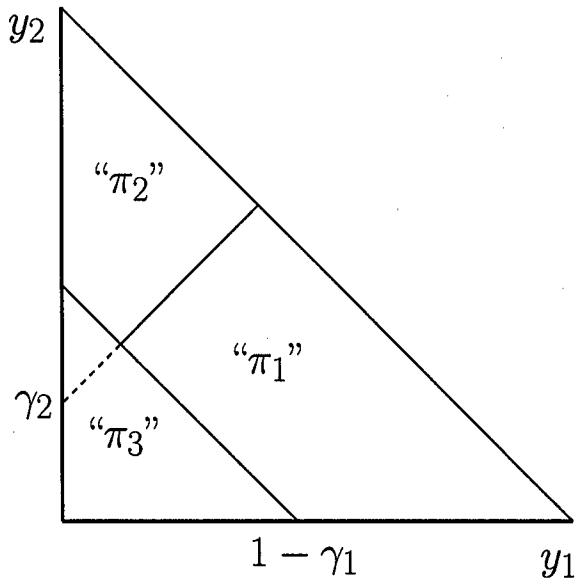


Fig. 7. Decision rule investigated by Mossman, for the decision parameters γ_1 and γ_2, shown in the *a posteriori* class probability space.

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

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

$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i P(\pi_i)/P(\pi_3)}{1 + \mathrm{LR}_1[P(\pi_1)/P(\pi_3)] + \mathrm{LR}_2[P(\pi_2)/P(\pi_3)]}. \tag{54}$$

This allows us to rewrite (51)-(53) as

$$(1+\gamma_2)\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 - (1-\gamma_2)\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = -\gamma_2 \qquad \{\text{"1-}vs\text{.-2"}\} \tag{55}$$

$$\gamma_1\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 + \qquad \gamma_1\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = 1-\gamma_1 \quad \{\text{"1-}vs\text{.-3"}\} \tag{56}$$

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

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

Although the Mossman decision rule 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 (9) are differences of the corresponding coefficients of (7) and (8). If the coefficients of (8) and (9) are identical, it must be the case that the coefficients of (7) 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

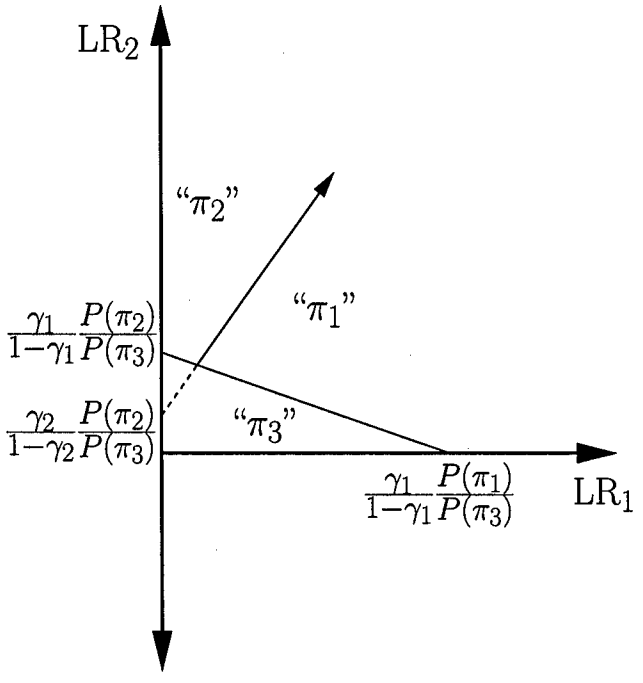


Fig. 8. Decision rule investigated by Mossman, for the decision parameters γ_1 and γ_2, shown in likelihood ratio space.

that 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).

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 observer's performance.

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 a decision rule proposed by Mossman (Mossman, 1999) does not correspond to ideal observer behavior for any possible values of the utilities used by the ideal observer. 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 *always* correspond to ideal observer behavior, while a rule similar in structure *never* can, 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 properties

to become "lost in the noise" so to speak. 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.

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.

Clark, A., 1984. Elements of Abstract Algebra. Dover Publications, Inc., New York.

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.

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.

Metz, C. E., Pan, X., 1999. 'Proper' binormal ROC curves: Theory and maximum-likelihood estimation. J. Math. Psychol. 43, 1–33.

Mossman, D., 1999. Three-way ROCs. Med. Decis. Making 19, 78–89.

Scurfield, B. K., 1996. Multiple-event forced-choice tasks in the theory of signal detectability. J. Math Psychol. 40, 253–269.

Scurfield, B. K., 1998. Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks. J. Math Psychol. 42, 5–31.

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

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

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

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

APPENDIX E

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 the present work 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

ROC analysis, ideal observers, three-class classification

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.

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

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 a particular condition or not 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 them as malignant or benign [13]–[17]. 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.

Our initial efforts toward this goal so far have been more theoretical than practical. 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 non-trivial 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 the present work we attempt to characterize the constraining relationships just mentioned among the three-class ideal observer's decision boundary lines. Although this work is admittedly still removed from image analysis *per se*, 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 Sec. III, we briefly illustrate that the intersection of the three-class ideal observer's decision boundary lines may lie in any of the four quadrants of the decision variable plane. In Sec. IV, 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.

Given the arbitrariness of the labels applied to the three classes (*i. e.*, which classes are considered first, second, or third), one would expect the selection of the fixed line in Sec. IV to be similarly arbitrary, and indeed in Secs. V and VI we show that corresponding 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. These results are summarized, and the principal conclusions of the work are given, in Sec. VII.

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:

$$\text{decide} \quad d = \pi_i \text{ iff } \sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k) \mathrm{LR}_k$$

$$\geq (U_{j|N} - U_{i|N})P(\mathbf{t} = \pi_N) \quad \{j < i\} \tag{1}$$

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

$$> (U_{j|N} - U_{i|N})P(\mathbf{t} = \pi_N) \quad \{j > i\}. \tag{2}$$

(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.) The partitioning is determined by the parameters

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

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

$$\gamma_{ijk} = \gamma_{kjk} - \gamma_{kik}. \tag{4}$$

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.*, $U_{i|i} > U_{j|i} \quad \{i \neq j\}$. This gives, for $j \neq i$,

$$\gamma_{iji} > 0, \tag{5}$$

leaving $N(N-1)$ 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 $1/(\sum_{i \neq j} \gamma_{iji})$. 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

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

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

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

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 *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).

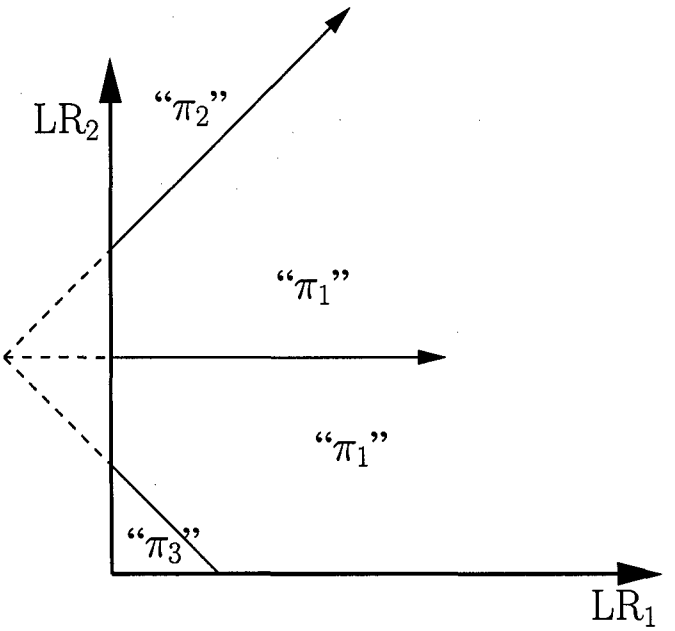


Fig. 1. Example ideal observer decision rule in which the intersection point of the decision boundaries lies in the second quadrant. Here the "1-*vs.*-2" line has unit slope, the "1-*vs.*-3" line has slope -1, and the "2-*vs.*-3" line is horizontal.

III. The Intersection Point

The intersection point of the three decision boundary lines mentioned at the end of the preceding section is often shown as occurring in the first quadrant ($\mathrm{LR}_1 \geq 0$, $\mathrm{LR}_2 \geq 0$). This is not, however, a requirement, and we can demonstrate with special cases that this intersection point can in fact occur in any quadrant. That is, we seek values of the γ_{iji} coefficients in (6)-(8), consistent with the conditions in (5), for which the intersection of the three lines occurs in the appropriate quadrant.

Consider a case in which $\gamma_{121} = 1/10$, $\gamma_{212} = 1/10$, $\gamma_{131} = 1/10$, $\gamma_{313} = 1/10$, $\gamma_{232} = 2/10$, and $\gamma_{323} = 4/10$. This yields the decision boundaries

$$\frac{1}{10}\mathrm{LR}_1 - \frac{1}{10}\mathrm{LR}_2 = -\frac{3}{10} \quad \{\text{"1-}vs.\text{-2"}\} \tag{9}$$

$$\frac{1}{10}\mathrm{LR}_1 + \frac{1}{10}\mathrm{LR}_2 = \frac{1}{10} \quad \{\text{"1-}vs.\text{-3"}\} \tag{10}$$

$$\frac{2}{10}\mathrm{LR}_2 = \frac{4}{10} \quad \{\text{"2-}vs.\text{-3"}\} \tag{11}$$

which intersect at the point $(\mathrm{LR}_1 = -1, \mathrm{LR}_2 = 2)$, as illustrated in Fig. 1.

An example of the intersection point occurring in the third quadrant may be obtained by taking $\gamma_{121} = 3/10$, $\gamma_{212} = 3/10$, $\gamma_{131} = 1/10$, $\gamma_{313} = 1/10$, $\gamma_{232} = 1/10$, and $\gamma_{323} = 1/10$. This yields

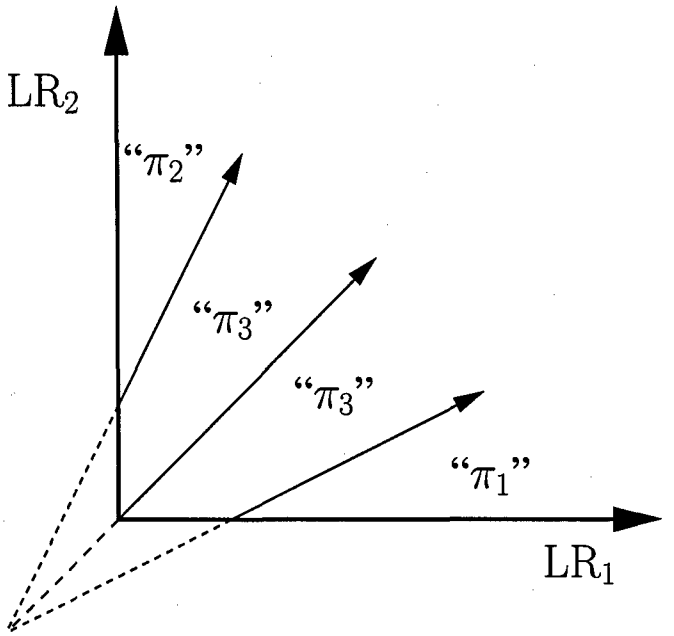


Fig. 2. Example ideal observer decision rule in which the intersection point of the decision boundaries lies in the third quadrant. Here the "1-*vs.*-2" line has unit slope, the "1-*vs.*-3" line has a slope of 1/2, and the "2-*vs.*-3" line has a slope of 2.

the decision boundaries

$$\frac{3}{10}\mathrm{LR}_1 - \frac{3}{10}\mathrm{LR}_2 = 0 \quad \{\text{"1-}vs.\text{-2"}\} \tag{12}$$

$$\frac{1}{10}\mathrm{LR}_1 - \frac{2}{10}\mathrm{LR}_2 = \frac{1}{10} \quad \{\text{"1-}vs.\text{-3"}\} \tag{13}$$

$$-\frac{2}{10}\mathrm{LR}_1 + \frac{1}{10}\mathrm{LR}_2 = \frac{1}{10} \quad \{\text{"2-}vs.\text{-3"}\} \tag{14}$$

which intersect at the point $(\mathrm{LR}_1 = -1, \mathrm{LR}_2 = -1)$, as illustrated in Fig. 2.

Finally, an example of the intersection point occurring in the fourth quadrant may be obtained by taking $\gamma_{121} = 1/10$, $\gamma_{212} = 1/10$, $\gamma_{131} = 2/10$, $\gamma_{313} = 4/10$, $\gamma_{232} = 1/10$, and $\gamma_{323} = 1/10$. This yields the decision boundaries

$$\frac{1}{10}\mathrm{LR}_1 - \frac{1}{10}\mathrm{LR}_2 = \frac{3}{10} \quad \{\text{"1-}vs.\text{-2"}\} \tag{15}$$

$$\frac{2}{10}\mathrm{LR}_1 = \frac{4}{10} \quad \{\text{"1-}vs.\text{-3"}\} \tag{16}$$

$$\frac{1}{10}\mathrm{LR}_1 + \frac{1}{10}\mathrm{LR}_2 = \frac{1}{10} \quad \{\text{"2-}vs.\text{-3"}\} \tag{17}$$

which intersect at the point $(\mathrm{LR}_1 = 2, \mathrm{LR}_2 = -1)$, as illustrated in Fig. 3.

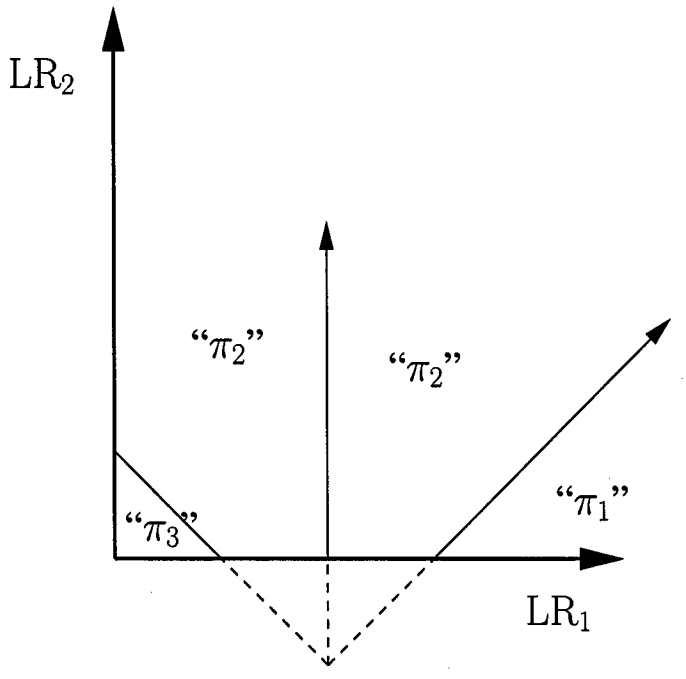


Fig. 3. Example ideal observer decision rule in which the intersection point of the decision boundaries lies in the fourth quadrant. Here the "1-*vs.*-2" line has unit slope, the "1-*vs.*-3" line is vertical, and the "2-*vs.*-3" line has a slope of -1.

IV. Restrictions Determined by the "1-vs.-3" Line

From the conditions on the γ_{iji} parameters, 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

$$m_{12} = \frac{\gamma_{121}}{\gamma_{212}} > 0 \tag{18}$$

$$\chi_{13} = \frac{\gamma_{313}}{\gamma_{131}} > 0 \tag{19}$$

$$b_{23} = \frac{\gamma_{323}}{\gamma_{232}} > 0. \tag{20}$$

These are the three conditions stated in [22].

Further constraints on the decision boundaries can be obtained by considering the two cases $\gamma_{232} - \gamma_{212} > 0$ and $\gamma_{232} - \gamma_{212} < 0$. In the first case (*i. e.*, $\gamma_{232} > \gamma_{212}$, or $U_{1|2} > U_{3|2}$), we have

$$m_{13} = \frac{-\gamma_{131}}{\gamma_{232} - \gamma_{212}} < 0 \tag{21}$$

$$b_{13} = \frac{\gamma_{313}}{\gamma_{232} - \gamma_{212}} > 0. \tag{22}$$

We also have

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

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 (18) and (21), it must therefore be the case that

$$m_{13} \leq m_{23} \leq m_{12}. \tag{24}$$

Furthermore,

$$\begin{aligned} b_{23} &= \frac{\gamma_{323}}{\gamma_{232}} \\ &= \frac{\gamma_{313}-(\gamma_{313}-\gamma_{323})}{\gamma_{232}} \\ &= \frac{(\gamma_{232}-\gamma_{212})b_{13}+\gamma_{212}b_{12}}{\gamma_{232}} \\ &= \left(1-\frac{\gamma_{212}}{\gamma_{232}}\right)b_{13}+\frac{\gamma_{212}}{\gamma_{232}}b_{12}. \end{aligned} \tag{25}$$

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 (24), we have the condition

$$\min(b_{12}, b_{13}) \leq b_{23} \leq \max(b_{12}, b_{13}). \tag{26}$$

If $b_{12} < 0$, then (26) immediately reduces to $b_{12} \leq b_{23} \leq b_{13}$ (by (22), we are considering a special case in which $b_{13} > 0$). This is illustrated in Fig. 4 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 (24) and (26) together imply two possible situations, depending on whether $b_{12} < b_{13}$ or $b_{12} \geq b_{13}$. These possibilities are illustrated in Fig. 5.

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

$$m_{13} = \frac{-\gamma_{131}}{\gamma_{232}-\gamma_{212}} > 0 \tag{27}$$

$$b_{13} = \frac{\gamma_{313}}{\gamma_{232}-\gamma_{212}} < 0. \tag{28}$$

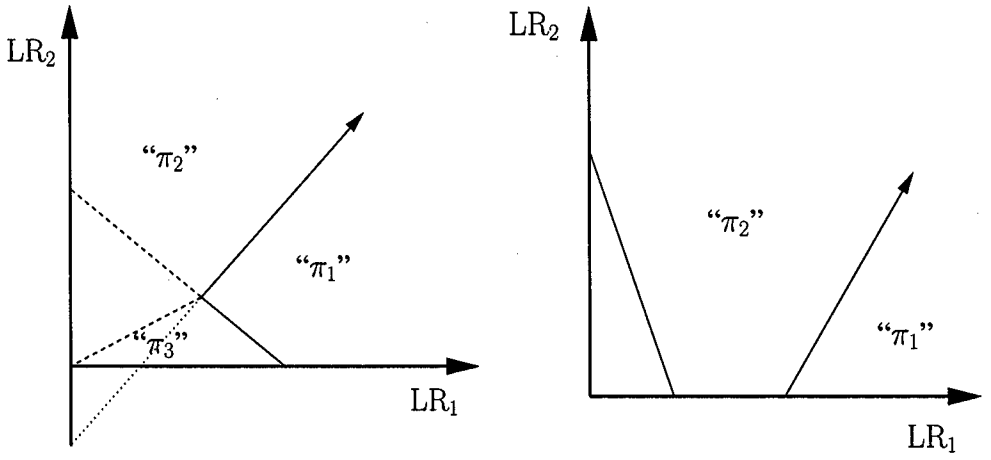


Fig. 4. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} > 0$ and $b_{12} < 0$. In (a), the "2-*vs.*-3" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $b_{23} > 0$); observations in the unlabeled region above this line will be decided "π_2", and those below this line will be decided "π_3". In (b), 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".

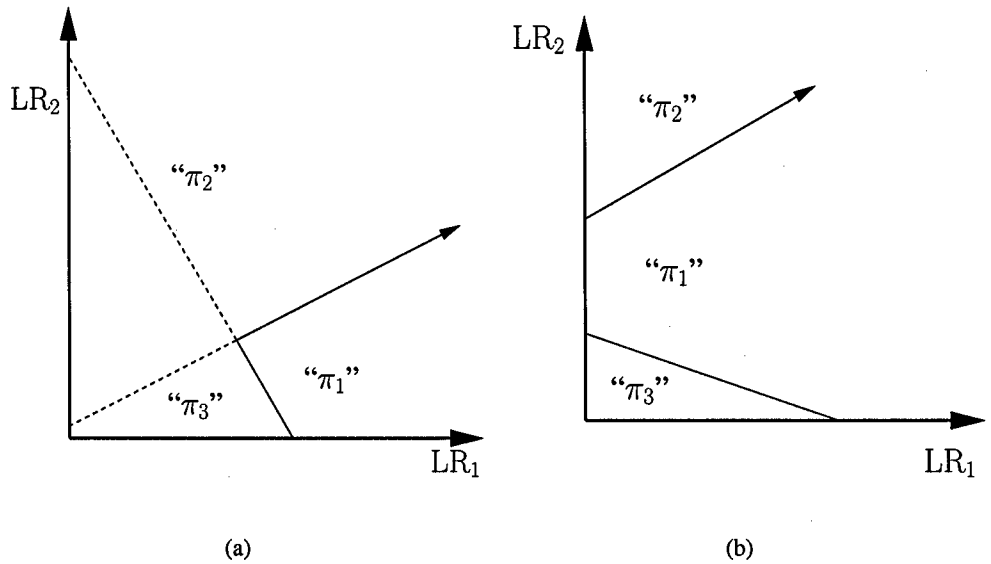


Fig. 5. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} > 0$ and $b_{12} \geq 0$. In (a), the "2-*vs.*-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2", and those below this line will be decided "π_3". In (b), the "2-*vs.*-3" 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.

We now have

$$\begin{aligned} m_{12} &= \frac{\gamma_{121}}{\gamma_{212}} \\ &= \frac{\gamma_{131} - (\gamma_{131} - \gamma_{121})}{\gamma_{212}} \\ &= \frac{-(\gamma_{232} - \gamma_{212})m_{13} + \gamma_{232}m_{23}}{\gamma_{212}} \\ &= \left(1 - \frac{\gamma_{232}}{\gamma_{212}}\right) m_{13} + \frac{\gamma_{232}}{\gamma_{212}} m_{23}. \end{aligned} \tag{29}$$

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

$$\min(m_{13}, m_{23}) \leq m_{12} \leq \max(m_{13}, m_{23}). \tag{30}$$

Furthermore,

$$\begin{aligned} b_{12} &= \frac{\gamma_{313} - \gamma_{323}}{-\gamma_{212}} \\ &= \frac{-\gamma_{313} + \gamma_{323}}{\gamma_{212}} \\ &= \frac{-(\gamma_{232} - \gamma_{212})b_{13} + \gamma_{232}b_{23}}{\gamma_{212}} \\ &= \left(1 - \frac{\gamma_{232}}{\gamma_{212}}\right) b_{13} + \frac{\gamma_{232}}{\gamma_{212}} b_{23}. \end{aligned} \tag{31}$$

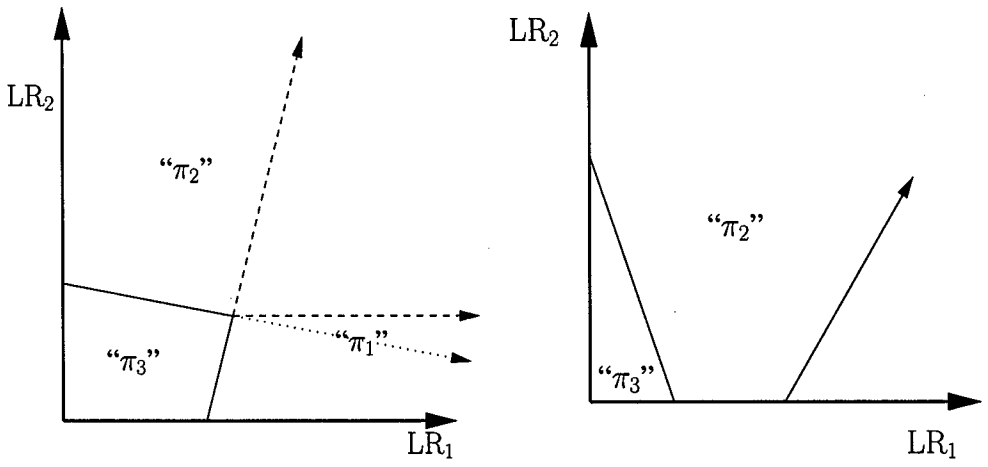
This is a weighted sum of the y-intercepts b_{13} and b_{23}, where the weights are positive and sum to one; thus, in addition to (30), we have the condition

$$b_{13} \leq b_{12} \leq b_{23}, \tag{32}$$

since $b_{13} < b_{23}$ by (20) and (28).

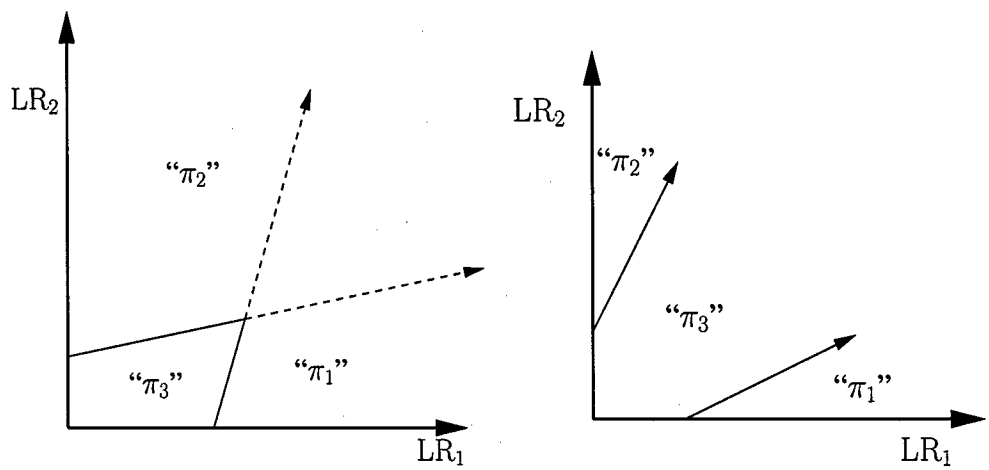
If $m_{23} < 0$, then (30) immediately reduces to $m_{23} \leq m_{12} \leq m_{13}$ (by (27), we are considering a special case in which $m_{13} > 0$). This is illustrated in Fig. 6 for the slightly different situations $\chi_{13} < \chi_{23}$ and $\chi_{13} \geq \chi_{23}$. If, on the other hand, $m_{23} \geq 0$, then (30) and (32) together imply two possible situations, depending on whether $m_{23} < m_{13}$ or $m_{23} \geq m_{13}$. These possibilities are illustrated in Fig. 7.

One may of course ask what happens when $\gamma_{232} - \gamma_{212} = 0$ (*i. e.*, $\gamma_{232} = \gamma_{212}$, or $U_{1|2} = U_{3|2}$).



(a) (b)

Fig. 6. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} < 0$ and $m_{23} < 0$. In (a), 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), 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".



(a) (b)

Fig. 7. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} < 0$ and $m_{23} \geq 0$. In (a), 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), 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.

In this case, both m_{13} and b_{13} are infinite. Furthermore,

$$\begin{aligned} m_{23} &= \frac{-(\gamma_{131}-\gamma_{121})}{\gamma_{232}} \\ &= \frac{-\gamma_{131}}{\gamma_{232}} + \frac{\gamma_{121}}{\gamma_{212}} \\ &= \frac{-\gamma_{131}}{\gamma_{232}} + m_{12} \\ &\leq m_{12}, \end{aligned} \tag{33}$$

and

$$\begin{aligned} b_{12} &= \frac{\gamma_{323}-\gamma_{313}}{\gamma_{212}} \\ &= \frac{\gamma_{323}}{\gamma_{232}} + \frac{-\gamma_{313}}{\gamma_{212}} \\ &= b_{23} + \frac{-\gamma_{313}}{\gamma_{212}} \\ &\leq b_{23}. \end{aligned} \tag{34}$$

Together, (33) and (34) can be considered *either* a special case of the inequalities (24) and (26), if we take $m_{13} = -\infty$ and $b_{13} = +\infty$; *or* of the inequalities (30) and (32), 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. 8.

V. Restrictions Determined by the "2-vs.-3" Line

In the preceding 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 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).

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

$$\frac{1}{m_{23}} = \frac{-\gamma_{232}}{\gamma_{131}-\gamma_{121}} < 0 \tag{35}$$

$$\chi_{23} = \frac{\gamma_{323}}{\gamma_{131}-\gamma_{121}} > 0. \tag{36}$$

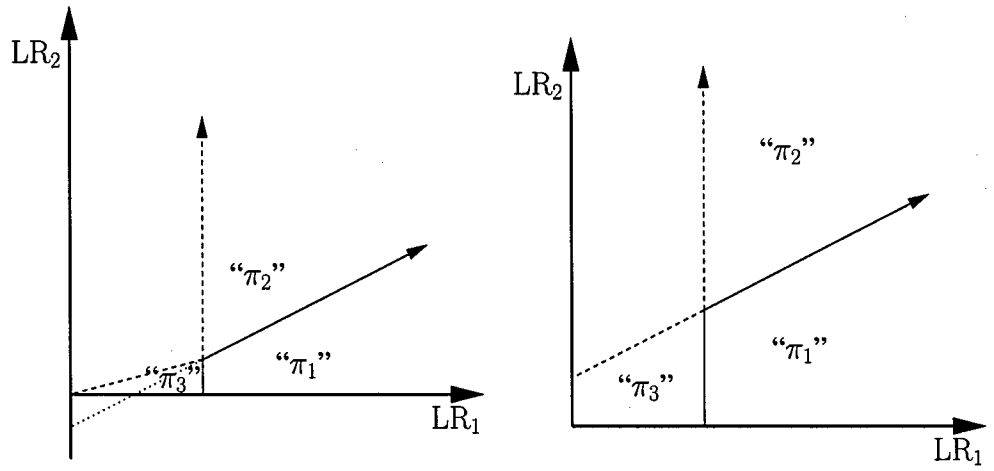


Fig. 8. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} = 0$. In (a), the "2-*vs.*-3" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $b_{23} > 0$); observations in the unlabeled region above this line will be decided "π_2", and those below this line will be decided "π_3". In (b), 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".

Through reasoning similar to that of the preceding section, we also have

$$\frac{1}{m_{23}} \leq \frac{1}{m_{13}} \leq \frac{1}{m_{12}} \tag{37}$$

and

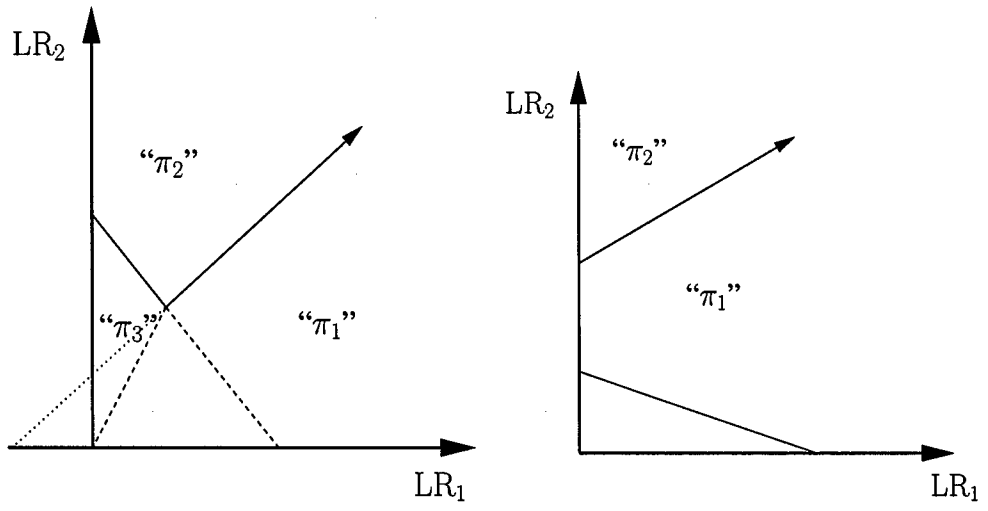
$$\min(\chi_{12}, \chi_{23}) \leq \chi_{13} \leq \max(\chi_{12}, \chi_{23}). \tag{38}$$

If $\chi_{12} < 0$, then (38) immediately reduces to $\chi_{12} \leq \chi_{13} \leq \chi_{23}$ (by (36), we are considering a special case in which $\chi_{23} > 0$). This is illustrated in Fig. 9 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 (37) and (38) together imply two possible situations, depending on whether $\chi_{12} < \chi_{23}$ or $\chi_{12} \geq \chi_{23}$. These possibilities are illustrated in Fig. 10.

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

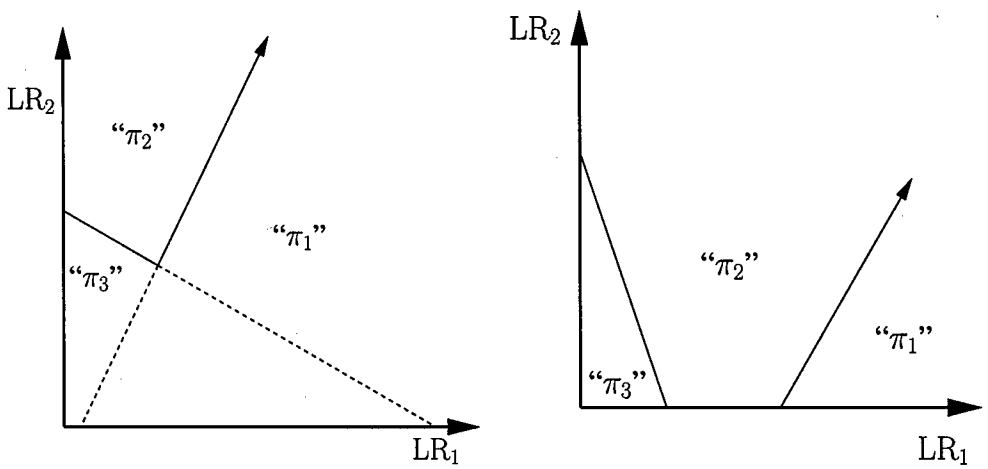
$$\frac{1}{m_{23}} = \frac{-\gamma_{232}}{\gamma_{131} - \gamma_{121}} > 0 \tag{39}$$

$$\chi_{23} = \frac{\gamma_{323}}{\gamma_{131} - \gamma_{121}} < 0. \tag{40}$$



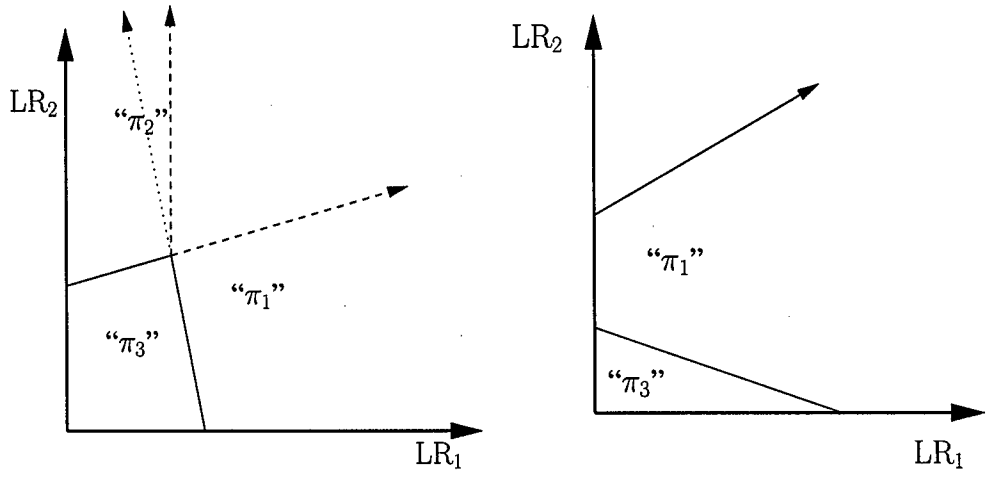
(a) (b)

Fig. 9. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} > 0$ and $\chi_{12} < 0$. In (a), 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), 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".



(a) (b)

Fig. 10. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} > 0$ and $\chi_{12} \geq 0$. In (a), 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), 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.



(a) (b)

Fig. 11. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} < 0$ and $1/m_{13} < 0$. In (a), 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), 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".

One can also show

$$\min\left(\frac{1}{m_{13}}, \frac{1}{m_{23}}\right) \leq \frac{1}{m_{12}} \leq \max\left(\frac{1}{m_{13}}, \frac{1}{m_{23}}\right) \tag{41}$$

and

$$\chi_{23} \leq \chi_{12} \leq \chi_{13}. \tag{42}$$

If $1/m_{13} < 0$, then (41) immediately reduces to $1/m_{13} \leq 1/m_{12} \leq 1/m_{23}$ (by (39), we are considering a special case in which $1/m_{23} > 0$). This is illustrated in Fig. 11 for the slightly different situations $b_{23} < b_{13}$ and $b_{23} \geq b_{13}$. If, on the other hand, $1/m_{13} \geq 0$, then (41) and (42) 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. 12.

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

$$\frac{1}{m_{13}} \leq \frac{1}{m_{12}} \tag{43}$$

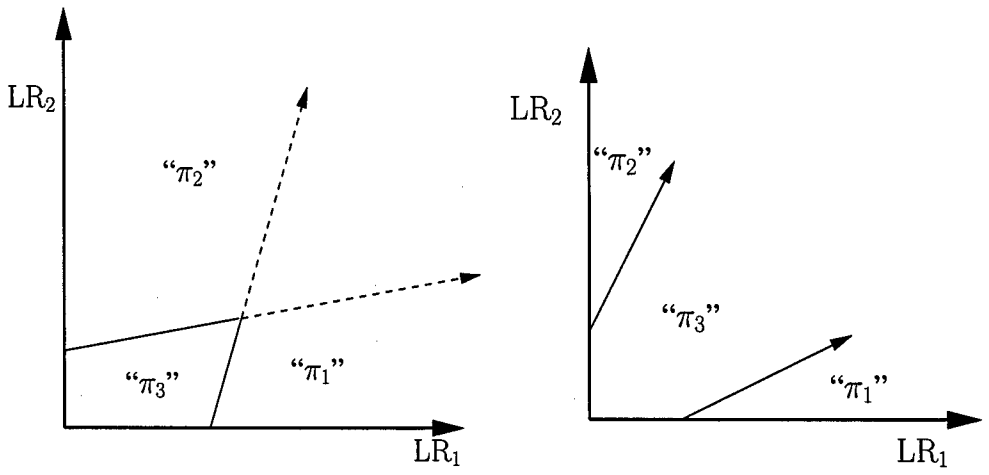


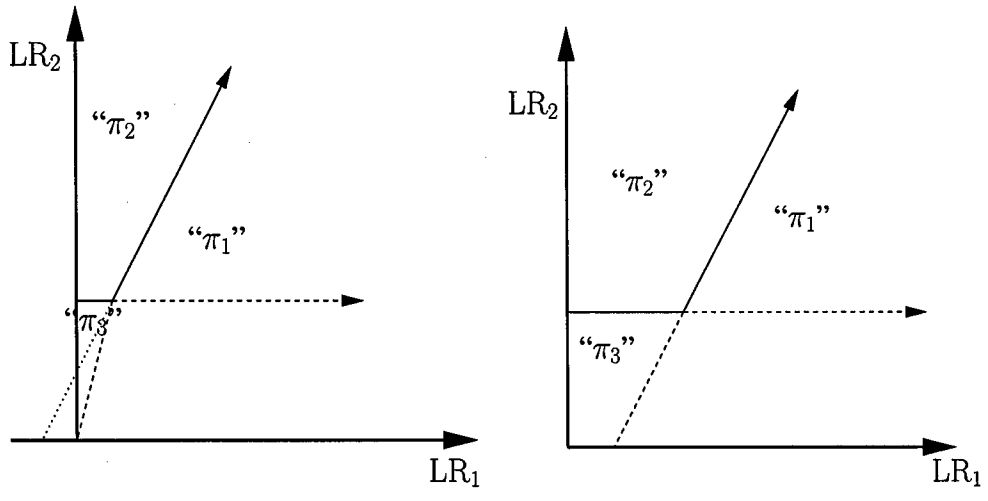
Fig. 12. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} < 0$ and $1/m_{13} \geq 0$. In (a), the "1-*vs.*-2" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2", and those below this line will be decided "π_1". In (b), the "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

$$\chi_{12} \leq \chi_{13}. \tag{44}$$

Together, (43) and (44) can be considered *either* a special case of the inequalities (37) and (38), if we take $1/m_{23} = -\infty$ and $\chi_{23} = +\infty$; *or* of the inequalities (41) and (42), 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. 13.

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



(a) (b)

Fig. 13. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} = 0$. In (a), 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), 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".

Fig. 9(a) $\Rightarrow$ Figs. 5(a), 6(a), 8(b)
Fig. 9(b) $\Rightarrow$ Fig. 5(b)
Fig. 10(a) $\Rightarrow$ Figs. 4(a), 6(a), 8(a)
Fig. 10(b) $\Rightarrow$ Figs. 4(b), 6(b), 8(a)
Fig. 11(a) $\Rightarrow$ Figs. 4(a), 5(a)
Fig. 11(b) $\Rightarrow$ Fig. 5(b)
Fig. 12(a) $\Rightarrow$ Figs. 7(a), 8(a), 8(b)
Fig. 12(b) $\Rightarrow$ Fig. 7(b)
Fig. 13(a) $\Rightarrow$ Figs. 5(a), 7(a), 8(b), 5(b)
Fig. 13(b) $\Rightarrow$ Figs. 4(a), 7(a), 8(a)

That is, none of the conditions derived in this section are inconsistent with those derived in the preceding section. Also note the symmetry between the corresponding equations and figures in Secs. IV and V, if one "swaps" the labels of classes π_1 and π_2, and additionally replaces m_{ij} with $1/m_{ij}$ and χ_{ij} with b_{ij}. (Intuitively, if one "flips" the figures in one section about the $y = x$

line, one obtains the figures in the other section.)

VI. Restrictions Determined by the "1-vs.-2" Line

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

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

$$\frac{1}{b_{12}} = \frac{-\gamma_{212}}{\gamma_{313} - \gamma_{323}} < 0 \tag{45}$$

$$\frac{1}{\chi_{12}} = \frac{\gamma_{121}}{\gamma_{313} - \gamma_{323}} > 0. \tag{46}$$

$$\tag{47}$$

Through reasoning similar to that of the preceding sections, we also have

$$\frac{1}{b_{12}} \leq \frac{1}{b_{13}} \leq \frac{1}{b_{23}} \tag{48}$$

and

$$\min\left(\frac{1}{\chi_{23}}, \frac{1}{\chi_{12}}\right) \leq \frac{1}{\chi_{13}} \leq \max\left(\frac{1}{\chi_{23}}, \frac{1}{\chi_{12}}\right). \tag{49}$$

If $1/\chi_{23} \leq 0$, then (49) immediately reduces to $1/\chi_{23} \leq 1/\chi_{13} \leq 1/\chi_{12}$ (by (46), we are considering a special case in which $1/\chi_{12} > 0$). This is illustrated in Fig. 14 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 (48) and (49) 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. 15.

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

$$\frac{1}{b_{12}} = \frac{-\gamma_{212}}{\gamma_{313} - \gamma_{323}} > 0 \tag{50}$$

$$\frac{1}{\chi_{12}} = \frac{\gamma_{121}}{\gamma_{313} - \gamma_{323}} < 0. \tag{51}$$

$$\tag{52}$$

One can also show

$$\min\left(\frac{1}{b_{13}}, \frac{1}{b_{12}}\right) \leq \frac{1}{b_{23}} \leq \max\left(\frac{1}{b_{13}}, \frac{1}{b_{12}}\right) \tag{53}$$

and

$$\frac{1}{\chi_{12}} \leq \frac{1}{\chi_{23}} \leq \frac{1}{\chi_{13}}. \tag{54}$$

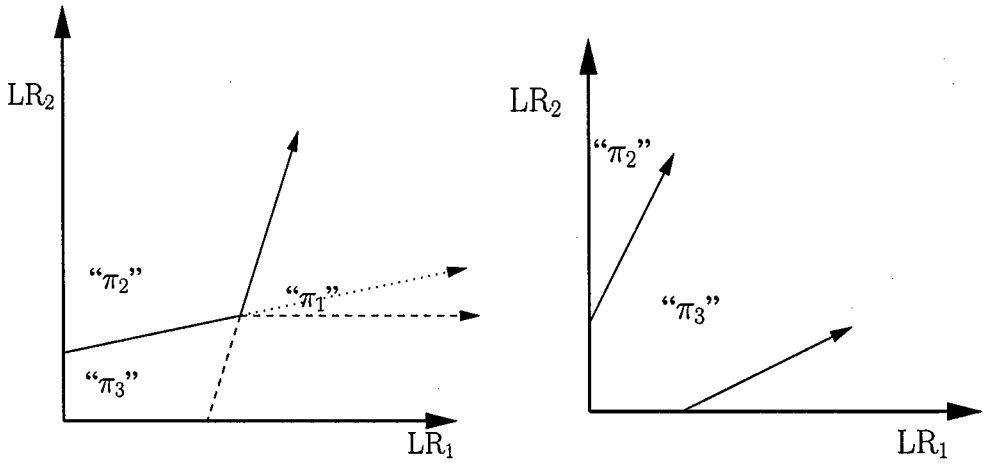


Fig. 14. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} > 0$ and $1/\chi_{23} \leq 0$. In (a), the "1-*vs.*-3" line can lie anywhere between the two dashed lines shown (the region between the horizontal dashed and dotted lines is excluded because $\chi_{13} > 0$, and 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), 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".

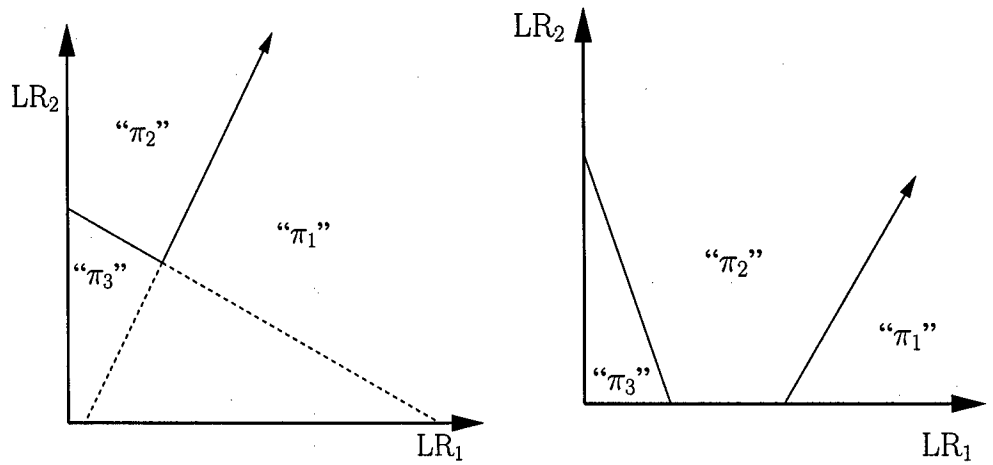


Fig. 15. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} > 0$ and $1/\chi_{23} > 0$. In (a), the "1-*vs.*-3" line can lie anywhere in the unlabeled region; observations to the left of this line will be decided "π_3", and those to the right of this line will be decided "π_1". In (b), 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.

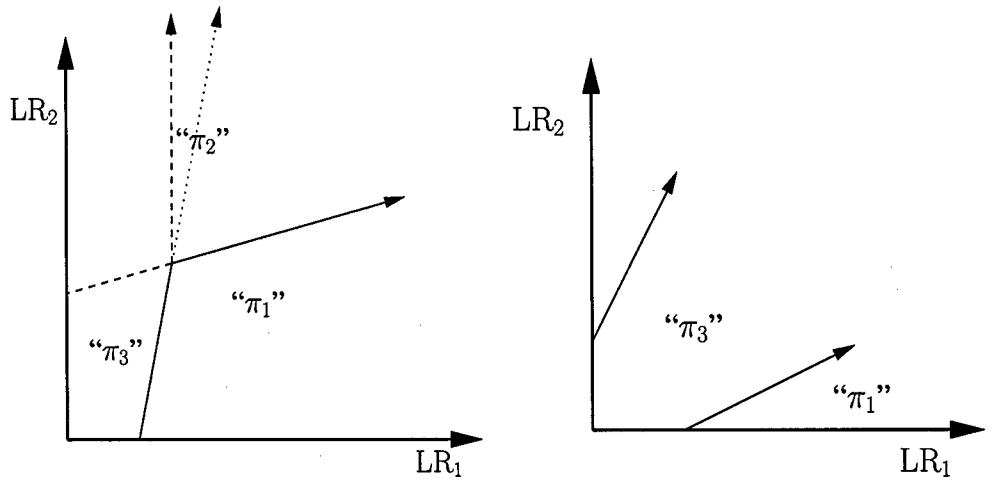


Fig. 16. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} < 0$ and $1/b_{13} \leq 0$. In (a), the "2-*vs.*-3" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $b_{23} > 0$, and therefore $1/b_{23} \geq 0$); 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), 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".

If $1/b_{13} \leq 0$, then (53) immediately reduces to $1/b_{13} \leq 1/b_{23} \leq 1/b_{12}$ (by (50), we are considering a special case in which $1/b_{12} > 0$). This is illustrated in Fig. 16 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 (53) and (54) 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. 17.

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

$$\frac{1}{b_{13}} \leq \frac{1}{b_{23}} \tag{55}$$

and

$$\frac{1}{\chi_{23}} \leq \frac{1}{\chi_{13}}. \tag{56}$$

Together, (55) and (56) can be considered *either* a special case of the inequalities (53) and (54), if we take $1/b_{12} = +\infty$ and $1/\chi_{12} = -\infty$; *or* of the inequalities (48) and (49), if we

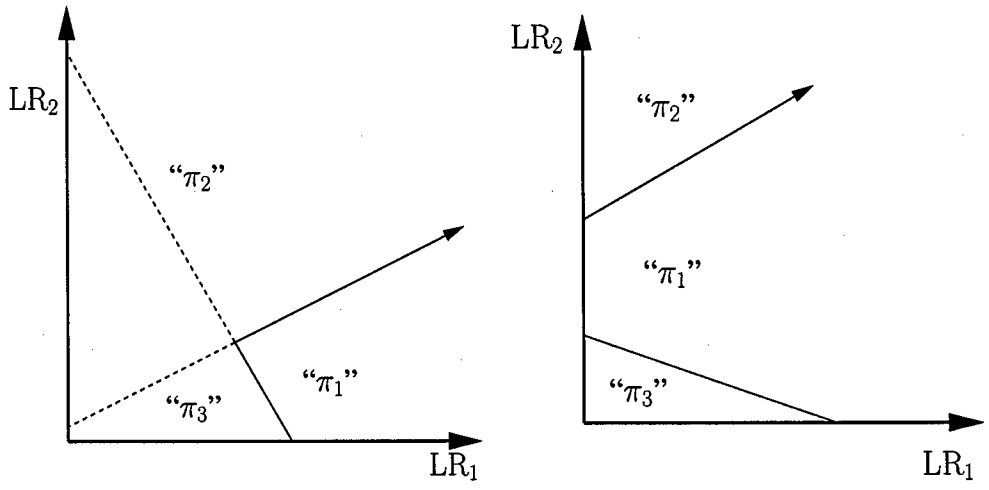


Fig. 17. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} < 0$ and $1/b_{13} > 0$. In (a), the "2-*vs.*-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2", and those below this line will be decided "π_3". In (b), the "2-*vs.*-3" 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.

take $1/b_{12} = -\infty$ and $\chi_{12} = +\infty$. This situation, for the slightly different cases $1/\chi_{23} \leq 0$ and $1/\chi_{23} > 0$, is illustrated in Fig. 18.

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

Fig. 14(a) $\Rightarrow$ Figs. 4(a), 8(a), 7(a)
Fig. 14(b) $\Rightarrow$ Fig. 7(b)
Fig. 15(a) $\Rightarrow$ Figs. 4(a), 8(a), 6(a)
Fig. 15(b) $\Rightarrow$ Figs. 4(b), 6(b), 8(a)
Fig. 16(a) $\Rightarrow$ Figs. 6(a), 7(a), 8(b)
Fig. 16(b) $\Rightarrow$ Fig. 7(b)
Fig. 17(a) $\Rightarrow$ Fig. 5(a)
Fig. 17(b) $\Rightarrow$ Fig. 5(b)
Fig. 18(a) $\Rightarrow$ Figs. 5(a), 8(b), 7(a), 7(b)
Fig. 18(b) $\Rightarrow$ Figs. 5(a), 8(b), 6(a), 6(b)

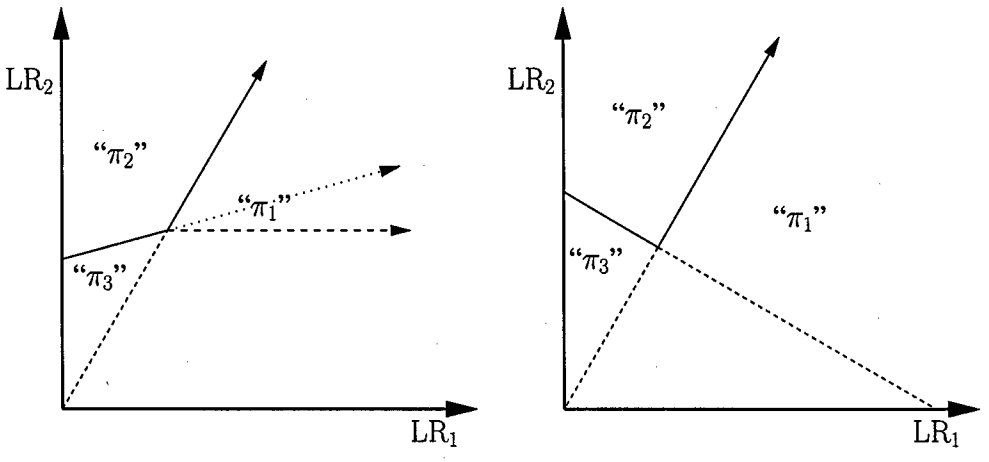


Fig. 18. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} = 0$. In (a), the "1-*vs.*-3" line can lie anywhere between the two dashed lines shown (the region between the horizontal dashed and dotted lines is excluded because $1/\chi_{13} \geq 0$); observations in the unlabeled 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), 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".

That is, none of the conditions derived in this section are inconsistent with those derived in the preceding sections.

VII. Discussion and Conclusions

The repetitive nature of the algebraic manipulations given in the preceding sections 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 LR_1, and are thus monotonic

in the decision criterion γ. For the three-class ideal observer, this straightforward relationship is lost; indeed, Figs. 5(b), 7(b), 10(b), 12(b), 15(b), and 17(b) show that for certain values of four of the five decision criteria γ_{iji}, the misclassification probabilities (*i. e.*, 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 observer's ROC hypersurface given the decision variable probability density functions, will require careful consideration of this issue (and no doubt others yet to be investigated).

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

[5] ——, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psychol.*, 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," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, Eds., SPIE, Bellingham, WA, 2003, pp. 567–578.

[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, 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.

[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, pp. 293–299, 2005.

[20] A. Srinivasan, "Note on the location of optimal classifiers in n-dimensional ROC space," Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, 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 (Spain), Tech. Rep., 2003.

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