Ordered Multiple Class Receiver Operating Characteristic (ROC) Analysis

Christos T. Nakas
Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Samos, Greece

Constantin T. Yiannoutsos
Division of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, U.S.A.

INTRODUCTION

Assessment of diagnostic markers for classification and prediction in two-class classification problems is commonly performed with the use of receiver operating characteristic (ROC) curves. The two classes under study are commonly referred to as the diseased and healthy populations. The ROC curve is a graphical representation resulting from plotting the proportion of diseased subjects with positive test result (TPR, true positive rate) vs. the proportion of healthy subjects with a positive test (FPR, false positive rate) that result from considering all possible values of a diagnostic test used as a cutoff. The area under the ROC curve (AUC) is used as a global measure of the test's discriminatory accuracy between the two groups. Estimation of the AUC and its standard error is performed by U-statistic (nonparametric) methods and parametric means.[1] In three-class diagnostic problems three true class rates (TCR) are defined using an appropriate classification criterion. Extension to multiple-class problems is straightforward. In the three-class case, an ROC surface is generated in three dimensions by considering all possible diagnostic test values. The volume under the ROC surface (VUS) can be used as a global measure of the three-class discriminatory ability of the test under consideration. Estimation and inference involving the VUS are also done in a manner similar to the estimation of the two-dimensional AUC. The VUS can be used whenever the diagnostic marker measurements are continuous or comprised of ordered categories (ordinal data).

ROC SURFACE DEFINITION AND CONSTRUCTION

ROC Curve Definition and Construction

ROC curves are commonly used for the evaluation of diagnostic markers in two-class testing. Let $X_{11}, \ldots, X_{1m}$ be m measurements obtained from the first class and $X_{21}, \ldots, X_{2n}$ be n measurements from the second class. For example, these two classes may involve healthy patients or patients having a certain disease. Without loss of generality, suppose that subjects from the first class tend to have lower measurements than subjects from the second class. The ROC curve is defined as the set of points $\{(\mathrm{FPR}(c),\mathrm{TPR}(c)), c \in R\}$ in the unit square, where $\mathrm{TPR}(c) = P(X_2 > c)$ and $\mathrm{FPR}(c) = P(X_1 > c)$ signify the probability that a diseased and a healthy patient, respectively, will have a positive test defined at the threshold point c. The TPR is the familiar concept of sensitivity of a test. An equivalent construction of the ROC curve is defined by the set of points $\{(\mathrm{TNR}(c),\mathrm{TPR}(c)), c \in R\}$ in the unit square, where $\mathrm{TNR}(c) = P(X_1 < c)$ is the probability that a healthy subject will have a negative test. This is the specificity of a test. The AUC is used as an overall performance index of the diagnostic accuracy of a marker, and it is equal to $P(X_1 < X_2)$. That is, the AUC is the average chance that the test score of a randomly selected subject from the first (healthy patient) group will be lower than the score of a randomly selected subject from the second (diseased patient) group. The AUC derived from a noninformative marker is 0.5 (implying that the test resulting from this marker is no better than the flip of a coin). Conversely, the AUC resulting from a perfectly discriminating marker is 1.0 (that is, a randomly selected pair of observations from the healthy and diseased groups will always have the correct ordering of marker values). The interested reader should review Refs.[1,2], two excellent recent textbooks on ROC curve methodology.

The ROC Surface and Hypersurface

Efforts to generalize the utility of the ROC curve for the assessment of diagnostic markers in multigroup classification problems have led to a number of

Encyclopedia of Biopharmaceutical Statistics DOI: 10.1081/E-EBS-120041740

approaches for the extension of ROC curve theory.[3–9] The difficulty in generalizing the ROC curve to more than two classes results from the fact that a decision rule for a K-group classification will produce K true classification rates (TCR) and $K(K-1)$ false classification rates (FCR), where $\mathrm{TCR}_k = P(C = k|D = k)$, $k = 1, \ldots, K$ and $\mathrm{FCR}_{ij} = P(C = i|D = j), i, j = 1, \ldots, K, i \neq j$, D is the true class membership of an individual while C is the class assigned by the testing procedure. A concise graphical representation of such a decision rule is the K-dimensional ROC hypersurface that can be used for the assessment of the diagnostic capacity of the marker under study.

For the sake of simplicity we illustrate first how an ROC surface is generated from a three-group classification problem based on a diagnostic marker X. Fig. 1 depicts a hypothetical situation for the distributions of marker measurements X_1, X_2, and X_3 obtained from three classes where, in general, measurements from class 1 are lower than those from class 2, and measurements from class 2 are lower than those from class 3. A decision rule that classifies subjects in one of these three classes can be defined as follows: For two ordered threshold points $c_1 < c_2$ (Fig. 1),

IF marker value $X < c_1$ THEN assign subject to class 1

ELSE IF $c_1 < X < c_2$ THEN assign to class 2

ELSE assign to class 3

In practice ties between measurements and threshold values may occur. In that case, the less than signs "$<$" above can arbitrarily be replaced with less than or equal signs "$\leq$"as long as the sets defined by the decision rule remain disjoint (i.e., no subject can be classified as being a member of more than one group). Varying the ordered decision thresholds $c_1 < c_2$ over all possible marker measurements, TCR are defined. These can be plotted in three dimensions to produce an ROC surface in the unit cube $[0,1] \times [0,1] \times [0,1]$. The TCR take values with corner coordinates $\{(1,0,0),(0,1,0),(0,0,1)\}$. This construction is a direct generalization of the ROC curve in three dimensions because ignoring c_2 and the subrules that this implies a conventional ROC curve between classes 1 and 2 is constructed. Similarly, ignoring c_1 results in an ROC curve between classes 2 and 3.

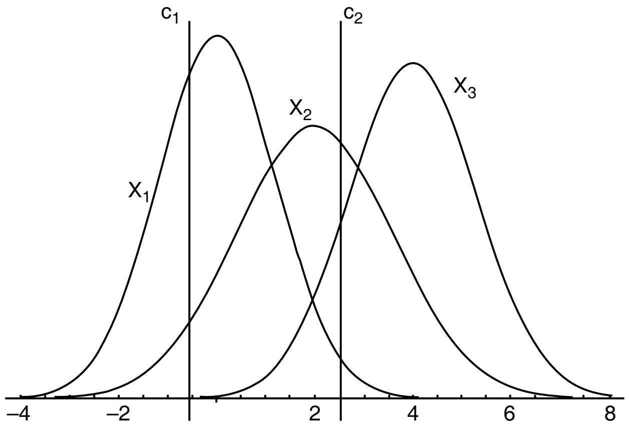

Fig. 1 Hypothetical overlapping distributions of marker measurements in a three-class experiment.

In the case of continuous or ordered categorical (ordinal) data, if the distributions of the marker values for the three groups are $X_1 \sim F_1$, $X_2 \sim F_2$, and $X_3 \sim F_3$, the theoretical TCR are defined by $F_1(c_1)$, $F_2(c_2) - F_2(c_1)$, and $1 - F_3(c_2)$, respectively, with $c_1 < c_2$ and $\{c_1, c_2\} \in R^2$. The ROC surface is constructed by plotting the points $(F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2))$ in three dimensions. In practice, the empirical cumulative distribution functions are used. These are defined from the data as the proportion of subjects in each group that are below the threshold c. For example, based on n_1 observations from the first class, the empirical cumulative distribution estimate of F_1 is defined as

$$\hat{F}_1(c) = \sum_{i=1}^{n_1} \frac{I(X_i < c)}{n_1}$$

where $I(X_i < c)$ is the indicator function and is equal to 1 if observation $X_i < c$, and 0 otherwise. Fig. 2 shows an empirical ROC surface with marker measurements simulated from three different, overlapping normal distributions. The theoretical VUS constructed by the decision rule described above is equal to the probability that three random measurements, one from each class, are classified in the correct order $X_1 < X_2 < X_3$. That is, $\mathrm{VUS} = P(X_1 < X_2 < X_3)$.

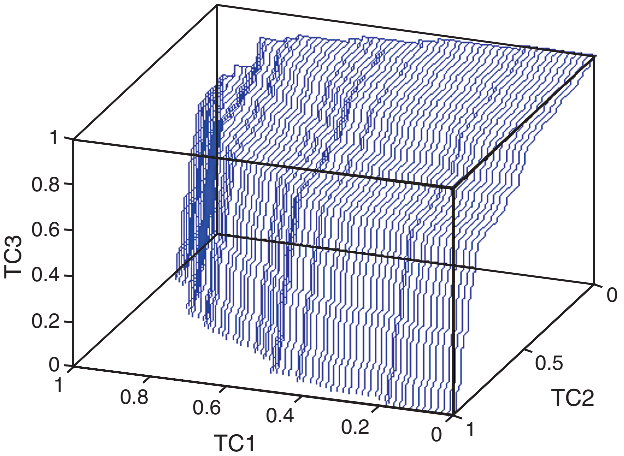

Fig. 2 ROC surface for measurements sampled from the three hypothetical distributions of Fig. 1. *(View this art in color at www.informaworld.com.)*

This is a straightforward generalization of the two-dimensional ROC curve. An unbiased nonparametric estimator of VUS is given by

$$\mathrm{VUS} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I\left(X_{1i}, X_{2j}, X_{3k}\right)$$

where n_1, n_2, and n_3 are the sample sizes from classes 1, 2, and 3, respectively, and $I(X_1,X_2,X_3)$ equals 1 if X_1, X_2, and X_3 are in the correct order, and 0 otherwise. In practice, ties between measurements may occur. In that case we define

$$I(X_1, X_2, X_3) = \begin{cases} 1 & X_1 < X_2 < X_3 \\ 1/2 \quad X_1 < X_2 = X_3 & \text{or} \quad X_1 = X_2 < X_3 \\ 1/6 & X_1 = X_2 = X_3 \\ 0 & \text{otherwise} \end{cases}$$

When the diagnostic marker is completely uninformative, i.e., when there is perfect overlap between the distributions of the three classes, VUS takes the value 1/6 (that is, classification into three groups using that marker is no better than random chance). Conversely, perfect separation between the three classes with $X_1 < X_2 < X_3$ for all measurements yields VUS $= 1$.

Generalization to K-class ($K > 3$) is straightforward. An ROC hypersurface can be constructed (but not visualized) by defining a diagnostic rule with $K - 1$ ordered decision thresholds. The K TCR thus defined should be plotted in a K-dimensional space. The ROC hypersurface is produced by varying the $K - 1$ ordered decision thresholds and calculating the respective TCR. A nonparametric unbiased estimate of the volume under the hypersurface is given by

$$\mathrm{VUS} = \frac{1}{n_1 n_2 \cdots n_k} \times \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_k=1}^{n_k} I(X_{1i_1}, X_{2i_2}, \cdots, X_{ki_k})$$

where n_i, $i = 1,2,\ldots,K$ are the sample sizes from classes $1,2,\ldots,K$, respectively, and IX_{1i_1}, X_{2i_2},..., X_{Ki_K} is defined in analogy to the three-class case. The VUS under this hypersurface takes the value 1 when there is perfect separation between the K-classes in the desired order, and it takes the value $1/(k!)$ when the K-class classification test is uninformative. The variance of the VUS can be calculated based on U-statistics theory or bootstrap techniques.[9] This can be used to produce $(1 - \alpha)$% confidence intervals and perform statistical inference.[9,10]

AN APPLICATION

Glial cells are found in the central nervous system. They provide structural support and protection for neurons. When neurons are injured, glial cells form a fibrous network in the area of injury. The resulting "gliosis" is prominent in a number of conditions that damage the central nervous system such as multiple sclerosis and stroke. In particular, gliosis has been found in neuropathology studies of injury to the central nervous system resulting from infection with the human immunodeficiency virus (HIV).[11] Myoinositol (MI) is a marker of glial cells in the brain.[12] Its levels can be measured by an imaging technique called magnetic resonance spectroscopy (MRS). Increased levels of MI, or its ratio over another marker creatine (Cr) MI/Cr, particularly in the basal ganglia have been associated with neurological impairment resulting from HIV infection.[13,14] MI/Cr measurements were assessed here as an illustration of a diagnostic tool for the discrimination of three subject groups: HIV-negative patients (NEG), HIV-positive patients neurologically asymptomatic (NAS), and HIV-positive patients with neurological impairment (ADC). Details AQ3 of this study have been described in detail elsewhere.[15] There are six possible orderings for MI/Cr measurements of the three classes, i.e., NEG $<$ NAS $<$ ADC, ADC, NAS $<$ NEG $<$ ADC, NEG $<$ ADC $<$ NAS, NAS $<$ ADC $<$ NEG, ADC $<$ NEG $<$ NAS, and ADC $<$ NAS $<$ NEG. The ROC surfaces corresponding to these orderings were constructed, and the volumes under these ROC surfaces were estimated to be 0.4299, 0.2090, 0.1542, 0.0867, 0.0647, and 0.0556, respectively. These results indicate that the strongest evidence supports an increasing trend in the MI/Cr measurements of the HIV-negative

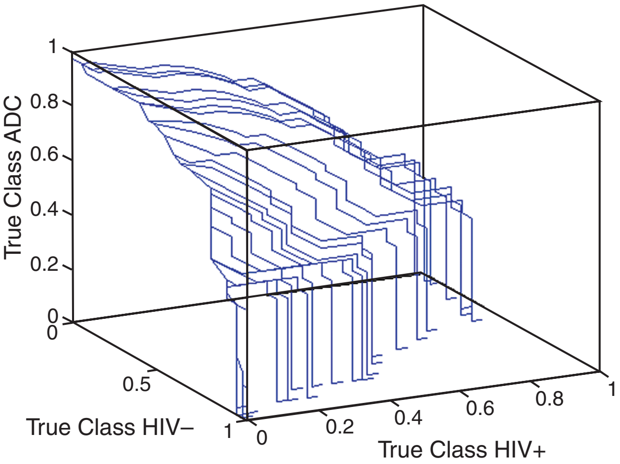

Fig. 3 ROC surface for MI/Cr in the basal ganglia measurements for the classification NEG $<$ NAS $<$ ADC. *(View this art in color at www.informaworld.com.)*

subjects, HIV-positive asymptomatic patients, and HIV-positive patients with neurological impairment. Fig. 3 shows the corresponding ROC surface. A 95% confidence interval for the corresponding VUS is (0.328,0.5318) based on the U-statistics methodology described in Refs.[9,10] Because the uninformative value of 1/6 for a three-class VUS is excluded from the 95% confidence interval, the implication is that MI/Cr measured via MRS in the basal ganglia is informative for the classification of subjects in the NEG, NAS, and ADC classes. In addition, when three measurements are given, one from each class, the probability that the MI/Cr level obtained from an HIV-negative subject is lower than that of the one obtained from an HIV-positive person and that both are lower than the MI/Cr levels measured in an individual suffering from HIV-related neurological impairment is estimated to be about 43% of the time.

CONCLUSIONS

Medical diagnostic tests often result in more than two classes. Traditional techniques such as ROC curve analysis can be generalized to three or more group classification problems by appealing to ROC surfaces in three or higher dimensions. In a similar manner, the area under the two-dimensional ROC curve, a global index of test accuracy, can be extended to three or higher dimensions by use of the VUS or hypersurface, respectively. We describe here nonparametric estimation of the VUS and use well-developed U-statistics methodology to derive estimates of the VUS along with confidence intervals. This formulation is useful especially when an inherent ordering exists in the marker measurements between the classes under study.

Because the relative size, rather than the absolute size, of the measurements is used, transformations of the measurement scale will result in the same estimated ROC surface and VUS in direct extension of this useful property of the two-dimensional ROC curve and AUC. In addition, the VUS defined above is the probability that the marker measurements from k ($k > 2$) subjects, one from each class, will be correctly ordered. This is a helpful interpretation, identical to the two-dimensional AUC. The VUS therefore maintains a number of useful properties of the two-dimensional ROC curve.

Thus, the described methodology addresses directly the question of the marker discriminatory performance over all the classes under study. Pairwise analysis of ROC curves can be performed in a post hoc manner as a detailed approach in multiple-class classification problems. In the example provided, measuring the VUS produced by MRS-measured MI/Cr levels shows that a strong trend exists between HIV-negative subjects, HIV-positive neurologically asymptomatic patients, and HIV-positive patients suffering from neurological complications associated with the infection. This observation has profound implications for the use of this marker in clinical diagnosis of HIV-associated neurological impairment as well as clinical consequences for our understanding of HIV-related neurological injury. Because MI/Cr levels of HIV-positive unimpaired patients were between those of HIV-negative and neurologically impaired subjects, the suggestion is that neurological injury occurs among HIV-infected individuals well before clinical symptoms (leading to an ADC diagnosis) are observed.[15] A three-class ROC surface analysis of these data therefore is a useful means of illuminating these relationships.

ARTICLES OF FURTHER INTEREST

Confidence Interval and Hypothesis Testing, p. 231.
Diagnostic Imaging, p. 288.
ROC Curve, p. 884.
SROC Curve, p. 001.
The Bootstrap, p. 000.

REFERENCES

1. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*; Oxford University Press: Great Britain, 2003.
2. Zhou, X.H.; Obuchowski, N.A.; McClish, D.K. *Statistical Methods in Diagnostic Medicine*; John Wiley & Sons, Inc.: New York, 2002.
3. Scurfield, B.K. Multiple-event forced-choice tasks in the theory of signal detectability. J. Math. Psychol. **1996**, *40*, 253–269.
4. Mossman, D. Three-way ROCs. Med. Decision Making **1999**, *19*, 78–89.
5. Hand, D.J.; Till, R.J. A simple generalization of the area under the ROC curve for multiple class classification problems. Mach. Learning **2001**, *45*, 171–186.
6. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical report HPL-2003-4; HP Laboratories: Palo Alto, CA, USA, 2003.
7. Lachiche, N.; Flach, P.A. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. Proceedings of the 20th International Conference on Machine Learning (ICML'03), Fawcett, T., Mishra, N., Eds.; AAAI Press, 2003; 416–423.
8. Ferri, C.; Hernandez-Orallo, J.; Salido, M.A. Volume under the ROC surface for multi-class problems. In *Lecture Notes in Artificial Intelligence, Machine Learning: ECML 2003*; Springer, 2003; Vol. 2837, 108–120.

9. Nakas, C.T.; Yiannoutsos, C.T. Ordered multiple-class ROC analysis with continuous measurements. Stat. Med. **2004**, *23*, 3437–3449.
10. Dreiseitl, S.; Ohno-Machado, L.; Binder, M. Comparing three-class diagnostic tests by three-way ROC analysis. Med. Decision Making **2000**, *20*, 323–331.
11. Navia, B.A.; Cho, E.S.; Petito, C.K.; Price, R.W. The AIDS dementia complex: II. Neuropathology. Ann. Neurol. **1986**, *19*, 525–535.
12. Brand, A.; Richter-Landsberg, C.; Leibfritz, D. Multinuclear NMR studies on the energy metabolism of glial and neuronal cells. Develop. Neurosci. **1993**, *15*, 289–298.
13. Lopez-Villegas, D.; Lenkinski, R.E.; Frank, I. Biochemical changes in the frontal lobe of HIV-infected individuals detected by magnetic resonance spectroscopy. Proc. Natl. Acad. Sci. **1997**, *94*, 9854–9859.
14. Yiannoutsos, C.T.; Ernst, T.; Chang, L.; Lee, P.L.; Richards, T.; Marra, C.M.; Meyerhoff, D.J.; Jarvik, J.G.; Richards, T.; Kolson, D.; Schifitto, G.; Ellis, R.J.; Swindells, S.; Simpson, D.M.; Miller, E.N.; Gonzalez, R.G.; Navia, B.A. Regional patterns of brain metabolism in AIDS Dementia Complex. Neuroimage **2004**, *23*, 928–935.
15. Chang, L.; Lee, P.L.; Yiannoutsos, C.T.; Ernst, T.; Marra, C.M.; Richards, T.; Kolson, D.; Schifitto, G.; Jarvik, J.G.; Miller, E.N.; Lenkinski, R.; Gonzalez, G.; Navia, B.A. HIV MRS Consortium. A multicenter in vivo proton-MRS study of HIV-associated dementia and its relationship to age. Neuroimage **2004**, *23*, 1336–1347.