Determining the properties of a newly developed test for comparing Receiver Operating Characteristic (ROC) curves

A.N. Meyen and M.R. Sooriyarachchi
Department of Statistics, University of Colombo, Colombo 3, Sri Lanka
roshinis@hotmail.com

ABSTRACT

Receiver operating characteristic (ROC) curves are graphical plots used for visualizing the performance of binary classifiers. A commonly used summary statistic to describe the ROC curve is its Area Under the Curve (AUC). The AUC's can be estimated either parametrically or non- parametrically. The parametric approach assumes that the signal present and signal absent groups can be represented by two overlapping Gaussian distributions. A novel asymptotic test for comparing multiple AUC's of several ROC curves was considered for this study. The objective of this study is to verify the properties of the proposed test. A simulation study was carried out for the case where the AUC's are independent and to study the behavior of the test for various sample sizes and varying degrees of overlap between the Gaussian distributions. Inferences were made regarding the Type I error and power of the test for the varying parameters. The proposed test performed better with respect to sample sizes above 140 when 3 ROC curves were being compared simultaneously. When the overlap between the Gaussian distributions were less the test statistic performed better with respect to the power of the test.

Keywords: Receiving Operating Characteristic (ROC) curve, Area Under the Curve (AUC), Beta Distribution.

1. INTRODUCTION

A Receiver Operating Characteristic (ROC) curve is a graphical plot of the true positive rate versus the false positive rate of a binary classifier. The most commonly used for summarizing the performance of a ROC curve is the value of the Area under the Curve (AUC) which ranges from 0 to 1, where the higher the value of the AUC, the better the discrimination power [1]. There are parametric, nonparametric and semi parametric methods of estimating the area under a ROC curve.

ROC curves are applied in diverse fields such as Medicine to Machine learning and Data Mining. In practice it is often required to compare several alternative binary classifiers. This involves the comparison of several AUC's under the ROC curves.It was of interest to study the Type I error and power of an asymptotic test proposed for comparing several AUC's , the details of which are given in the methods and materials section.

2. MATERIALS AND METHODS

Binormal ROC Curves: The signal detection paradigm on which ROC curves are based is important to understand the underlying principle behind ROC curve analysis. According to [2], the signal-detection paradigm consists simply of a subject successively choosing between a signal present population (with background noise), SN, or signal absent population (just noise), N. The model then assumes that the response of the subject can be represented by a random variable X with cumulative distribution function,

$F_{SN}(x)F_{SN}(x)$ if the signal was present,

$F_N(x)F_N(x)$ if no signal was present.

For the purposes of this study $F_N(x) = \Phi(x)$, $F_{SN}(x) = \Phi(bx - a)$ where b and a are the two principal parameters of the ROC curve which can be seen to depend on the means and standard deviation of $F_N(x)F_N(x)$ and $F_{SN}(x)$ $F_{SN}(x)$ and $\Phi(.)\Phi(.)$ denotes the cumulative density function of the standard normal distribution. The values of a and b along with other parameters of the ROC curve were estimated using the method of scoring proposed in [2].

Simulation: The method of scoring used is an iterative process which uses initial parameter estimates. The start for the initial iteration was used as the parameter estimates of the simple linear regression as given in [2]. Iteration continues until either, two successive iterates differ by less than $10^{-3}10^{-3}$ in all of their components and the final iterate is a possible solution. Degenerate solution for the parameter estimates of the ROC curve can occur from empty cells in the data matrix. Therefore in order to overcome the problem of degeneracy similar to [3]

the method of scoring developed adds a small positive constant in order to avoid degeneracy in the case of empty cells.

Calculation of the AUC and variance of the AUC:

It is possible to obtain the AUC of a ROC curve using the following formula where $\Phi(.)$ denotes the cumulative standard normal distribution.

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \quad (1)$$

In order to calculate the variance of the AUC, the delta method [4] is made use of, giving the formula as follows for the variance.

$$Var(\widehat{AUC}) = \left(\frac{\partial AUC}{\partial a}\right)^2 var(\hat{a}) + \left(\frac{\partial AUC}{\partial b}\right)^2 var(\hat{b}) + 2\left(\frac{\partial AUC}{\partial a}\right)\left(\frac{\partial AUC}{\partial b}\right)cov(\hat{a},\hat{b})$$

Proposed test statistic: The test was developed using various results from multivariate statistics along with the properties of ROC curves. The derivation of the test developed is given below,

Let $\underline{AUC} = \begin{pmatrix} AUC_1 \\ AUC_2 \\ \vdots \\ AUC_p \end{pmatrix}\underline{AUC} = \begin{pmatrix} AUC_1 \\ AUC_2 \\ \vdots \\ AUC_p \end{pmatrix}$

which is a p×1 vector , where $AUC_i AUC_i$ denotes the $AUCAUC$ of the i^{th} i^{th} ROC curve.

Let $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$ denote the Maximum Likelihood Estimate (MLE) of the $\underline{AUC}\ \underline{AUC}$ vector, and $\underline{\mu}$ $\underline{\mu}$ be the expected value of $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$ and $\underline{\Sigma\Sigma}$ be the associated variance-covariance matrix of $\underline{\widehat{AUC}}$ $\underline{\widehat{AUC}}$. As $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$ is the MLE of $\underline{AUC}\ \underline{AUC}$ and as MLE's are asymptotically normal, for large samples $\underline{\widehat{AUC}} \sim N_p(\mu,\Sigma).\underline{\widehat{AUC}} \sim N_p(\mu,\Sigma).$ If the estimate $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$ of $\underline{AUC}\ \underline{AUC}$ of a ROC curve is made up of the sum of nn independent quantities where nn is a function of $n_1 n_1$ (the number of positive responses) and $n_2 n_2$ (the number of negative responses) according to [5]. Therefore $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$ is made up of $n_1 n_2 n_1 n_2$ quantities of which $n = \min(n_1, n_2) n = \min(n_1, n_2)$ are independent. $\underline{\hat{\Sigma}\hat{\Sigma}}$ is the MLE of the covariance matrix $\underline{\Sigma\Sigma}$ of $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$. According to [6] the sampling distribution of the MLE of the $(\underline{\widehat{AUC}} - \mu)(\underline{\widehat{AUC}} - \mu)'$ $(\underline{\widehat{AUC}} - \mu)(\underline{\widehat{AUC}} - \mu)'$ matrix is asymptotically distributed as Wishart, $W_p(\underline{\Sigma}, \mathrm{n})W_p(\underline{\Sigma}, \mathrm{n})$ as $\underline{\widehat{AUC}}$ $\underline{\widehat{AUC}}$ has an asymptotic multivariate normal distribution. Therefore, $\underline{\hat{\Sigma}} \sim W_p(\underline{\Sigma}, \mathrm{n})\underline{\hat{\Sigma}} \sim W_p(\underline{\Sigma}, \mathrm{n})$.

It is needed to test the null hypothesis $H_0 H_0$ that all $\underline{AUC}'\underline{AUC}'$ s are same on average versus the alternative hypothesis $H_1 H_1$ that all the $\underline{AUC}'\underline{AUC}'$ s are not the same on average.

i.e. $H_0: \underline{\mu} = \underline{K} = H_0: \underline{\mu} = \underline{K} =$ constant vector versus $H_1: \underline{\mu} \neq \underline{K} H_1: \underline{\mu} \neq \underline{K}$. It is possible to estimate $\underline{KK}$ as the simple average of $\underline{\widehat{AUC}}\ \underline{\widehat{AUC}}$ (i.e. the simple average of the individual $\widehat{AUC}_i'\widehat{AUC}_i'$ s.) $\underline{KK}$ can be then estimated by $\overline{KK}$ (under $H_0 H_0$) where,

$$\overline{K} = \frac{\sum_{i=1}^{p} \widehat{AUC}_i}{p} \quad (3)$$

As $\underline{KK}$ is not known it has to be estimated. From [7] the general form of the Hotelling's $T^2 T^2$ statistic is as follows,

$$T_G^2 = (\underline{\widehat{AUC}} - \overline{K})'\hat{\Sigma}^{-1}(\underline{\widehat{AUC}} - \overline{K}) \quad (4)$$

The dimensionality pp needs to be reduced by 1 for estimating $\overline{KK}$. Therefore taking $q = p - 1$ $q = p - 1$ instead of pp for large samples gives the following,

$$T_G^2 \frac{n}{(n-1)^2} \sim Beta\left(\frac{q}{2}, \frac{n-q-1}{2}\right) \quad (5)$$

Here pp is the number of AUC's and nn is the number of independent quantities used to calculate the AUC's. For the case of large samples (large $n_1 n_1$ and $n_2 n_2$) nn will be large. The test statistic $T_G^2 T_G^2$ can be used to test $H_0 H_0$.

3. RESULTS AND DISCUSSION

Simulation of the size and power of the test under the null and alternative hypotheses respectively when 3 AUC's of ROC curves are compared:

The size and power of the test was simulated for sample sizes of 20, 50, 100, 120, 140, 250 and 500 for varying values of *a* and *b*. The values of *a* and *b* were selected according to previous research [8]. For this study only independent ROC curves were considered. Table I gives the results of the simulation related to the proportion of rejections out of 1000 under the null hypothesis(H_0) (H_0). This indicates the size of the test.

Table 2 gives the results of the simulation related to the proportion of rejections out of 1000 under the alternative hypothesis. This indicates the power of the test. As the sample size increases the power of the test also increases for the different combinations of $a_1, a_2, a_3 a_1, a_2, a_3$ and $b_1, b_2, b_3 b_1, b_2, b_3$.

In conclusion it can be seen that the significance level of the test and power of the test becomes better when the sample size increases and that the test statistic performs better above sample sizes of 120.The power of the test increases with decreasing overlap. When the sample size is above 140 the Type I error is often within the 95% confidence limits given by [0.036, 0.064].

4. CONCLUSIONS

The size of the test lies within the 95% confidence limits for sample sizes above 140. Also, in general, the power of the test increases when the sample size increases. When the overlap between the signal present and signal absent distributions for the ROC curves decreased as seen by the different $a_1, a_2, a_3 a_1, a_2, a_3$ and b_1, b_2, b_3 b_1, b_2, b_3 values, the power of the test increased. This can be attributed to the fact that the method of scoring [2] has been used for determining the AUC's and this method is based on maximum likelihood estimation. As maximum likelihood estimates are asymptotically normally distributed and the theory of this test is based on normality of the AUC's this test does well in conjunction with the method of scoring [2] for large samples.

Table 1: Under H_0 (Type I error of test): (comparing 3 AUC's simultaneously)

Sample size	a_1	a_2	a_3	b_1	b_2	b_3	Proportion of rejections
20	0.33	0.33	0.33	0.67	0.67	0.67	0.1440
	0.5	0.5	0.5	1	1	1	0.1270
	0.5	0.5	0.5	0.67	0.67	0.67	0.1330
	0.75	0.75	0.75	1	1	1	0.1340
	0.66	0.66	0.66	0.67	0.67	0.67	0.1070
	1.0	1.0	1.0	1	1	1	0.1060
50	0.33	0.33	0.33	0.67	0.67	0.67	0.0730
	0.5	0.5	0.5	1	1	1	0.0700
	0.5	0.5	0.5	0.67	0.67	0.67	0.0720
	0.75	0.75	0.75	1	1	1	0.0640
	0.66	0.66	0.66	0.67	0.67	0.67	0.0750
	1.0	1.0	1.0	1	1	1	0.0790
100	0.33	0.33	0.33	0.67	0.67	0.67	0.0510
	0.5	0.5	0.5	1	1	1	0.0610
	0.5	0.5	0.5	0.67	0.67	0.67	0.0680
	0.75	0.75	v	1	1	1	0.0650
	0.66	0.66	0.66	0.67	0.67	0.67	0.0750
	1.0	1.0	1.0	1	1	1	0.0850
120	0.33	0.33	0.33	0.67	0.67	0.67	0.0720
	0.5	0.5	0.5	1	1	1	0.0710
	0.5	0.5	0.5	0.67	0.67	0.67	0.0750
	0.75	0.75	0.75	1	1	1	0.0840
	0.66	0.66	0.66	0.67	0.67	0.67	0.0740
	1.0	1.0	1.0	1	1	1	0.0770
140	0.33	0.33	0.33	0.67	0.67	0.67	0.0650
	0.5	0.5	0.5	1	1	1	0.0430
	0.5	0.5	0.5	0.67	0.67	0.67	0.0650
	0.75	0.75	0.75	1	1	1	0.0540
	0.66	0.66	0.66	0.67	0.67	0.67	0.0610
	1.0	1.0	1.0	1	1	1	0.0670
250	0.33	0.33	0.33	0.67	0.67	0.67	0.0630
	0.5	0.5	0.5	1	1	1	0.0490
	0.5	0.5	0.5	0.67	0.67	0.67	0.0580
	0.75	0.75	0.75	1	1	1	0.0420
	0.66	0.66	0.66	0.67	0.67	0.67	0.0390
	1.0	1.0	1.0	1	1	1	0.0580
0.33	0.33	0.33	0.33	0.67	0.67	0.67	0.0500
	0.5	0.5	0.5	1	1	1	0.0370
	0.5	0.5	0.5	0.67	0.67	0.67	0.0480
	0.75	0.75	0.75	1	1	1	0.0530
	0.66	0.66	0.66	0.67	0.67	0.67	0.0370
	1.0	1.0	1.0	1	1	1	0.0480

Table 2: Under H_1H_1 (Power of test):(comparing 3 AUC's simultaneously)

Sample size	a_1	a_2	a_3	b_1	b_2	b_3	Proportion of rejections
20	0.67	0.5	0.67	0.67	1	0.67	0.1370
	0.67	0.33	0.67	0.67	0.67	0.67	0.1580
	1.0	0.5	1.0	1	1	1	0.1730
	1.0	0.33	1.0	1	0.67	1	0.2030
50	0.67	0.5	0.67	0.67	1	0.67	0.0920
	0.67	0.33	0.67	0.67	0.67	0.67	0.1330
	1.0	0.5	1.0	1	1	1	0.2050
	1.0	0.33	1.0	1	0.67	1	0.2700
100	0.67	0.5	0.67	0.67	1	0.67	0.1280
	0.67	0.33	0.67	0.67	0.67	0.67	0.1950
	1.0	0.5	1.0	1	1	1	0.2470
	1.0	0.33	1.0	1	0.67	1	0.3680
120	0.67	0.5	0.67	0.67	1	0.67	0.1290
	0.67	0.33	0.67	0.67	0.67	0.67	0.2330
	1.0	0.5	1.0	1	1	1	0.3310
	1.0	0.33	1.0	1	0.67	1	0.4360
140	0.67	0.5	0.67	0.67	1	0.67	0.1560
	0.67	0.33	0.67	0.67	0.67	0.67	0.2350
	1.0	0.5	1.0	1	1	1	0.3400
	1.0	0.33	1.0	1	0.67	1	0.4810
250	0.67	0.5	0.67	0.67	1	0.67	0.1820
	0.67	0.33	0.67	0.67	0.67	0.67	0.3500
	1.0	0.5	1.0	1	1	1	0.5270
	1.0	0.33	1.0	1	0.67	1	0.7240
500	0.67	0.5	0.67	0.67	1	0.67	0.3320
	0.67	0.33	0.67	0.67	0.67	0.67	0.6610
	1.0	0.5	1.0	1	1	1	0.8430
	1.0	0.33	1.0	1	0.67	1	0.9640

5. REFERENCES

[1] Fawcett, T."An introduction to ROC analysis." Pattern Recognition Letters, Volume 27, pp. 861-874, 2006.

[2] Grey, D. M. & Morgan, B. T. " Some aspects of ROC curve fitting: Normal and Logistic models."Journal of Mathematical Psychology,Volume 9, pp. 128-139, 1972.

[3] Dorfman, D. D. & Berbaum, K. S. "Degeneracy and discrete reciever operating characteristic rating data." Academic Radiology, 2(10), pp. 907-915, 1995.

[4] Casella, G & Berger, R. L. "Statistical Inference." Duxbury Press. Second Edition, 2002.

[5] Vergara, I. A. et al. "StAR: a simple tool for the statistical comparison of ROC curves."BMC Bioinformatics, 9(265), 2008.

[6] Mardia, K. V., Kent, J. T. & Bibby, J. M., "Multivariate Analysis."London: Academic Press, 1979.

[7] Hotelling, H. Multivariate Quality Control. In: C. Eisenhart, M. W. Hastay & W. A. Wallis, eds. "Techniques of Statistical Analysis."New York: McGraw-Hill, 1947.

[8] Cleeves, A. M. "Comparative assessment of three common algorithms for estimating the variance of the area under the nonparametric reciever operating characteristic curve." The Stata Journal, 2(3), pp. 280-289, 2002.