Marker selection by a threshold independent measure associated with partial AUC

Park, Eunsik, Yu, Wenbao
Chonnam National University, Department of statistics
77 Yongbong-ro, Buk-gu
Gwangju (500-757), South Korea
E-mail: wbaoyu@ejnu.net, espark02@chonnam.ac.kr

1 Introduction

The receiver operating characteristic (ROC) curve is a statistical tool for evaluating the accuracy of diagnostic tests. The ROC curve demonstrates the balance between true positive rate (TPR or sensitivity) and false positive rate (FPR or 1-specificity). Let T be a feature measurement, a simple rule based on such a single feature can usually be expressed as $T > c$ for a threshold c. A sample is classified as positive if $T > c$. For a specific threshold c, $TPR(c) = Pr(T > c|P)$ and $FPR(c) = Pr(T > c|N)$, where P and N denote positive and negative, respectively. ROC curve is a two-dimensional plot of $(FPR(c), TPR(c)) : -\infty < c < \infty$. Classification rule that have $(FPR(c), TPR(c))$ close to $(0, 1)$ indicate satisfactory discrimination, while those with $(FPR(c), TPR(c))$ near the 45° line cannot discriminate between the two groups.

However, comparing curves directly has never been easy. In biomedical studies, investigators often compare the validity of two tests based on the estimated areas under the respective ROC curves (AUC). AUC is a popular metric because it has a simple probabilistic interpretation (Bamber (1975)). AUC also has an advantage over single measures of performance such as the odd ratio or relative risk (Pepe et al. (2004)). AUC is threshold independent because it considers all possible thresholds. The problem is not all possible threshold is of practical use. Many markers may have the same AUC but perform very differently for a range of thresholds of interest. So AUC alone seems to be not sufficient to measure the performance of markers/classifiers. Many authors used partial area under the ROC curve (PAUC) instead, which limits FPR in a low range or TPR in a high range, (Dodd and Pepe (2003); Komoriet al. (2010); Wang and Chang (2010)). On the other hand, all existing methods based on PAUC needs to give prerequisite range of FPR (or TPR).

In this paper, alternatively, we propose an method to automatically select markers with larger PAUC, where FPR is proved to be in a practicable low range. Consequently, prerequisite range of FPR is not necessary. By comparing the cross point of two ROC curves, we found a fact that the marker with larger ratio of standard deviation of measurement of case group to standard deviation of measurement of control group (RSD) is associated with larger PAUC with a reasonably low FPR, under normal distribution assumption. Then together with AUC, we give a scheme to select markers.

This paper is organized as follows. In section 2, we discuss the cross point of two ROC curves. In section 3, we show the numerical study to demonstrate our method. Section 4 gives conclusion.

2 ROC Curves

2.1 Cross point of ROC curves

Let $\tilde{x}$ and x be two markers. Cases and controls follow $N(\tilde{\mu}_1, \tilde{\sigma}_1^2)$ and $N(\tilde{\mu}_0, \tilde{\sigma}_0^2)$, respectively, for marker $\tilde{x}$, and $N(\mu_1, \sigma_1^2)$ and $N(\mu_0, \sigma_0^2)$ for marker x, respectively. By the probability interpretation of AUC (Bamber (1975)), the exact AUC of x and $\tilde{x}$ can be written as:

$$
\begin{aligned}
&AUC(x) = \Phi(\tfrac{\mu_1-\mu_0}{\sqrt{\sigma_1^2+\sigma_0^2}}) = \Phi(\Delta), \\
&AUC(x) = \Phi(\tfrac{\tilde{\mu}_1-\tilde{\mu}_0}{\sqrt{\tilde{\sigma}_1^2+\tilde{\sigma}_0^2}}) = \Phi(\tilde{\Delta}),
\end{aligned} \tag{2.1}
$$

where $\Delta = (\mu_1 - \mu_0)/\sqrt{\sigma_0^2 + \sigma_1^2}$ and $\tilde{\Delta} = (\tilde{\mu}_1 - \tilde{\mu}_0)/\sqrt{\tilde{\sigma}_0^2 + \tilde{\sigma}_1^2}$, which can be treated as effect size of marker x and $\tilde{x}$, respectively. For convenience, we give the following definition:

Definition 2.1. Let $ROC_\alpha(\tilde{x}) > ROC_\alpha(x)$ denotes that the ROC curve of $\tilde{x}$ is always higher than ROC curve of x, when false positive rate is not larger than α, i.e. $0 < FPR < \alpha$.

Definition 2.2. Let $PAUC_\alpha$ denotes that the partial area under ROC curve with $0 < FPR < \alpha$.

Note that $ROC_\alpha(\tilde{x}) > ROC_\alpha(x)$ implies $PAUC_c(\tilde{x}) > PAUC_c(x)$, for and c, $0 < c < \alpha$.

For convenience, we write $RSD(x) = \sigma_1/\sigma_0$, and $RSD(\tilde{x}) = \tilde{\sigma}_1/\tilde{\sigma}_0$. The following Lemma shows that the ratio of standard deviation of case group to standard deviation of control group, RSD, play a markable role when ROC curves cross.

Lemma 2.3. *Suppose there are two markers x and $\tilde{x}$. Cases and controls follow $N(\tilde{\mu}_1, \tilde{\sigma}_1^2)$ and $N(\tilde{\mu}_0, \tilde{\sigma}_0^2)$, respectively, for marker $\tilde{x}$, and $N(\mu_1, \sigma_1^2)$ and $N(\mu_0, \sigma_0^2)$ for marker x, respectively., we have*

a. The ROC curve of marker x will be cross with the ROC curve of marker $\tilde{x}$ if and only if

$$\frac{\sigma_1}{\sigma_0} \neq \frac{\tilde{\sigma}_1}{\tilde{\sigma}_0}, \quad (i.e, RSD(x) \neq RSD(\tilde{x})).$$

b. If $\sigma_1/\sigma_0 \neq \tilde{\sigma}_1/\tilde{\sigma}_0$, then $ROC_{\Phi(A)}(\tilde{x}) > ROC_{\Phi(A)}(x)$ holds if and only if

$$\frac{\tilde{\sigma}_1}{\tilde{\sigma}_0} > \frac{\sigma_1}{\sigma_0}, \quad (i.e, RSD(\tilde{x}) > RSD(x)$$

. where $\Phi(A)$ is the FPR of the cross point, and $A = (\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1} - \frac{\sigma_0}{\sigma_1})^{-1}(\frac{\mu_1-\mu_0}{\sigma_1} - \frac{\tilde{\mu}_1-\tilde{\mu}_0}{\tilde{\sigma}_1})$.

Proof. Since ROC curve of x is the plot of $(\alpha, TPR(FPR^{-1}(\alpha))), \alpha \in [0,1]$, where $TPR(c) = Prop(x_1 > c) = 1 - \Phi(\frac{c-\mu_1}{\sigma_1}) = \Phi(\frac{\mu_1-c}{\sigma_1})$ and $FPR(c) = Prop(x_0 > c) = 1 - \Phi(\frac{c-\mu_0}{\sigma_0}) = \Phi(\frac{\mu_0-c}{\sigma_0})$,

$$TPR(FPR^{-1}(\alpha)) = TPR(a(\alpha)) = \Phi(\frac{\mu_1 - \mu_0 + \sigma_0\Phi^{-1}(\alpha)}{\sigma_1}). \tag{2.2}$$

So, we can conclude that the point on ROC cuve $\tilde{x}$ with $FPR = \alpha$ is higher than the point on ROC cuve x with $FPR = \alpha$, $(0 < \alpha < 1)$, if and only if

$$
\begin{aligned}
&\frac{\tilde{\mu}_1 - \tilde{\mu}_0 + \tilde{\sigma}_0\Phi^{-1}(\alpha)}{\tilde{\sigma}_1} > \frac{\mu_1 - \mu_0 + \sigma_0\Phi^{-1}(\alpha)}{\sigma_1} \\
\Longleftrightarrow\ &(\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1} - \frac{\sigma_0}{\sigma_1})\Phi^{-1}(\alpha) > \frac{\mu_1 - \mu_0}{\sigma_1} - \frac{\tilde{\mu}_1 - \tilde{\mu}_0}{\tilde{\sigma}_1} \\
\Longleftrightarrow\ &\begin{cases} 0 < \alpha < \Phi(A), & \text{if } \tilde{\sigma}_0/\tilde{\sigma}_1 < \sigma_0/\sigma_1 \\ 1 > \alpha > \Phi(A), & \text{if } \tilde{\sigma}_0/\tilde{\sigma}_1 > \sigma_0/\sigma_1 \\ \text{trivial case,} & \text{if } \tilde{\sigma}_0/\tilde{\sigma}_1 = \sigma_0/\sigma_1. \end{cases}
\end{aligned}
$$

Here trivial case means ROC curve of one marker is always higher or always lower than that of another marker(except when $\alpha = 0$ or 1), then the conclusion follows immediately. □

Based on Lemma 2.3, the two ROC curves cross if and only if the RSDs are not equal. The ROC curve of the marker with large RSD will be higher on the left side of the cross point.

2.1.1 The range of FPR at the cross point

Following the above notation, the cross point of two ROC curves of marker x and $\tilde{x}$ is $(\Phi(A), \Phi((\mu_1 - \mu_0 + \sigma_0 A)/\sigma_1))$. Following simple calculation, we have

$$\begin{aligned} A &= \frac{\Delta\sqrt{1+(\frac{\sigma_0}{\sigma_1})^2}-\tilde{\Delta}\sqrt{1+(\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1})^2}}{\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1}-\frac{\sigma_0}{\sigma_1}} \\ &= \Delta\frac{\sqrt{1+(\frac{\sigma_0}{\sigma_1})^2}-\sqrt{1+(\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1})^2}}{\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1}-\frac{\sigma_0}{\sigma_1}} + \frac{\tau\sqrt{1+(\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1})^2}}{\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1}-\frac{\sigma_0}{\sigma_1}}, \end{aligned} \tag{2.3}$$

where $\tau = \Delta - \tilde{\Delta}$, the difference of effect size of the two markers. The FPR at the cross point curve is dependent on effect size of AUCs and RSDs. Denote $\kappa = \tilde{\sigma}_0/\tilde{\sigma}_1 - \sigma_0/\sigma_1$, without loss of generality, suppose $\kappa > 0$, then by Talor expansion theorem, we can get

$$-\Delta\frac{1}{\sqrt{1+(\frac{\tilde{\sigma}_1}{\tilde{\sigma}_0})^2}} + \frac{\tau}{\kappa}\sqrt{1+(\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1})^2} \leq A \leq -\Delta\frac{1}{\sqrt{1+(\frac{\sigma_1}{\sigma_0})^2}} + \frac{\tau}{\kappa}\sqrt{1+(\frac{\tilde{\sigma}_0}{\tilde{\sigma}_1})^2} \tag{2.4}$$

if $AUC(x) = AUC(\tilde{x})$, i.e., $\tau = 0$, then,

$$-\Delta\frac{1}{\sqrt{1+(\frac{\tilde{\sigma}_1}{\tilde{\sigma}_0})^2}} \leq A \leq -\Delta\frac{1}{\sqrt{1+(\frac{\sigma_1}{\sigma_0})^2}}. \tag{2.5}$$

Formulae (2.4) and (2.5) are not easy to use. Fortunately, a simple lower bound of A can be obtained, i.e., $A > -\Delta$ follows if $\tau \geq 0$ (i.e., $AUC(x) \geq AUC(\tilde{x})$). This lower bound is easy to use, because $\Phi(A) > \Phi(-\Delta) = 1 - AUC(x)$. For instance, if $AUC(x) \leq 0.9$, then $\Phi(A) > 0.1$, i.e., marker x with larger PAUC can be restricted at least within interval $(0, 0.1)$. Since in practice, AUC of a single maker can not be large (say less than 0.9), this lower bound is of practicable use.

We can conclude, not strictly, that if the difference of AUC is not large, $A > -\Delta$ holds, i.e., a lower bound FPR at the intersection point can be estimated by $1 - AUC$. We summarize this result in the following theorem:

Theorem 2.4. *Following the same distribution and notation of Lemma 2.3, let $\kappa = \tilde{\sigma}_0/\tilde{\sigma}_1 - \sigma_0/\sigma_1$ and $\tau = \Delta - \tilde{\Delta}$. Without loss of generality, suppose $RSD(x) > RSD(\tilde{x})$ i.e., $ROC_{\Phi(A)}(x) > ROC_{\Phi(A)}(\tilde{x})$ or equivalently, $PAUC_\alpha(x) > PAUC_\alpha(\tilde{x})$, for any α, $0 < \alpha \leq \Phi(A)$, then two following conclusions follows:*

a. if $AUC(x) \geq AUC(\tilde{x})$, i.e., $\tau \geq 0$, then, $\Phi(A) \geq 1 - AUC(x)$.

b. if $AUC(x) < AUC(\tilde{x})$, i.e., $\tau < 0$, and $|\tau|/\Delta \leq B$, then, $\Phi(A) \geq 1 - AUC(x)$.

where $B = \kappa(\sqrt{1 + (\tilde{\sigma}_0/\tilde{\sigma}_1)^2} - \tilde{\sigma}_0/\tilde{\sigma}_1)/(1 + (\tilde{\sigma}_0/\tilde{\sigma}_1)^2)$, which only dependents on RSD ($\tilde{\sigma}_0/\tilde{\sigma}_1$ and σ_0/σ_1).

By Theorem 2.4 a, if the classifier with larger RSD is also associated with larger AUC, t, then the lower bound of FPR at the cross point of two ROC curves is given by $1 - t$. Moreover, if the classifier with larger RSD is associated with less AUC, t, then by Theorem 2.4 b, if the relative change of effect size of AUC is bounded by B (which is only dependent on RSD), then the lower bound of FPR at the cross point of two ROC curves can be estimated by $1 - t$.

2.1.2 The range of TPR at the cross point

The TPR at the cross point is $\Phi((\mu_1 - \mu_0 + \sigma_0 A)/\sigma_1)$. If $\Phi(A) > 1 - AUC(x)$ holds, it is larger than $\Phi(\Delta(\sqrt{\sigma_0^2 + \sigma_1^2} - \sigma_0)/\sigma_1) > 0.5$. Especially, if AUCs of two compared markers are the same or their difference of AUC is bounded as Theorem 2.4 b, the TPR at the cross point of ROC curves is larger than 0.5.

2.2 A marker selection scheme by AUC and RSD

Since larger AUC is associated with larger PAUC (FPR in a low range), we propose an algorithm of marker selection by using AUC and RSD. Suppose there are p markers, our scheme goes as follows:

1. Rank markers by AUC and separate them into g groups. In each group, the markers have similar AUCs;
2. Rank markers by RSD in each group, larger RSD has higher rank.
3. Select the top p_i markers in the i-th group, $i = 1, ..., g$.

3 Numerical Study

3.1 Simulation

Here we generate 12 markers, following normal and mixture normal distributions. Assume the control group for all markers follow standard normal distribution, for case group, x_1-x_6 follow normal distribution, and x_7-x_{12} follow mixture normal distribution. The markers are generated as follows: Let AUC=0.7, 0.8, 0.9, for each AUC we generate 3 distributions $N(\mu_i, i^2)$, where μ_i is decided by AUC and standard deviation i, i=1, 2, 3. The mixture normal distribution for each AUC is constructed by the combination of these 3 distributions. We select 2 normal distributions and 2 mixture normal distributions for each AUC, then x_1-x_{12} are constructed.

To rank these markers, we compute the empirical AUC, $PAUC_{0.1}$ and $PAUC_{0.2}$ with repetition 100 times, where $PAUC_\alpha$ was defined as partial AUC as $0 < FPR < \alpha$. The sample sizes are 200 for both case and control groups. By our proposed scheme, we separate the 12 markers into 3 group, then rank each group by RSD of each marker. We list the distribution, mean and standard error of AUC, PAUC and RSD of each marker in Table 1.

We can see that in each group, both $PAUC_{0.1}$ and $PAUC_{0.2}$ increase strictly with RSD. By our theoretical results, for each group, the marker with larger RSD is associated with larger $PAUC_\alpha$, where $0 < \alpha < 1 - AUC$. In group 1 and 2, $AUC = 0.7$ and 0.8 respectively, then the upper bound of α can be at least as large as 0.3 and 0.2 respectively; In group 3, $AUC = 0.9$, so the upper bound of α can be at least as large as 0.1, which is of practicable use even 0.1 is a very conservative estimate, for the result is also true when $\alpha = 0.2$ by Table 1.

For the similar structure in the 3 groups, We plot the density function and average ROC curve for each marker in group 2 in Figure 1. The marker with higher RSD will have higher ROC curve in a reasonably low range of FPR. By numeric simulation, this result is also true for a kind of mixture normal distribution.

4 Conclusion

We found that among the markers with similar AUC, the one with larger RSD has larger PAUC, where FPR is in a low range. We have also estimated the corresponding range of FPR, which is simple and useful in a practice. We have introduced use of PAUC without requiring the specified threshold. Coupled with AUC, we also give a clue to do marker selection by RSD.

References

John Q.Su and Jun S.liu (1993). Linear Combinations Of Multiple Diagnostic Markers *Journal of the American statistical Assotiation* 88, 1350-1355

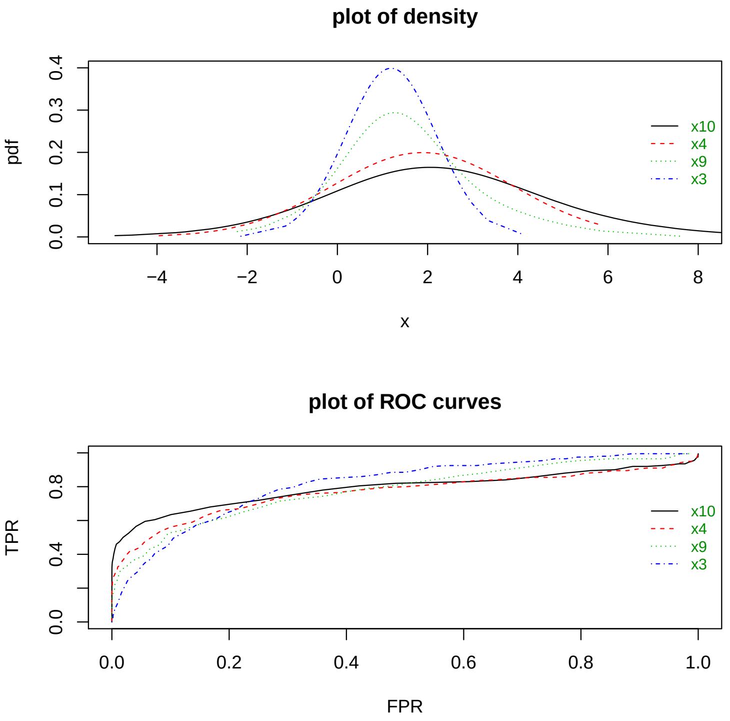

Figure 1: Plot of Group 2: density functions and ROC curves.

Zhangfeng Wang, Yuan-chin I. Chang etc (2007). A parisimonious threshold-independent protein feature selection method through the area under receive operating characteristic curve *Bioinformatics* 23, 2788-2794.

Osamu Komori and Shinto Eguchi (2010). A boosting method for maximizing the partial area under the ROC curve *BMC Bioinformatics* 11,

Bamber,D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph *Journal of Mathematical Psychology*, 12, 387-415.

Shuangge Ma, Jian Huang (2005). Regularized ROC method for disease classification and biomarker selection with microarray data *Bioinformatics* 21, 4356-4362.

Lori E. Dodd and Margaret S. Pepe(2003). Partial AUC estimation and regression. *Biometrics* 59, 614-623.

Table 1: *Summary of Simulation Study*

group	marker	case	AUC	$PAUC_{0.1}$	$PAUC_{0.2}$	RSD
1	x_1	$N(0.74, 1)$	0.697(0.026)	0.009(0.004)	0.031(0.012)	1.012(0.077)
	x_7	$0.5N(0.74, 1) + 0.5N(1.16, 2^2)$	0.697(0.026)	0.015(0.006)	0.045(0.013)	1.612(0.123)
	x_2	$N(1.16, 2^2)$	0.697(0.026)	0.024(0.008)	0.059(0.016)	2.040(0.137)
	x_8	$0.5N(1.64, 3^2) + 0.5N(1.16, 2^2)$	0.700(0.028)	0.028(0.008)	0.067(0.016)	2.543(0.167)
2	x_3	$N(1.19, 1)$	0.799(0.020)	0.016(0.006)	0.053(0.016)	1.000(0.065)
	x_9	$0.5N(1.19, 1) + 0.5N(1.89, 2^2)$	0.798(0.020)	0.024(0.008)	0.067(0.019)	1.622(0.128)
	x_4	$N(1.89, 2^2)$	0.802(0.023)	0.031(0.009)	0.080(0.018)	1.979(0.171)
	x_{10}	$0.5N(2.65, 3^2) + 0.5N(1.89, 2^2)$	0.801(0.021)	0.036(0.008)	0.087(0.020)	2.589(0.176)
3	x_5	$N(1.81, 1)$	0.902(0.014)	0.029(0.010)	0.083(0.021)	1.008(0.066)
	x_{11}	$0.5N(1.81, 1) + 0.5N(2.86, 2^2)$	0.902(0.014)	0.037(0.011)	0.091(0.019)	1.694(0.128)
	x_6	$N(2.86, 2^2)$	0.896(0.014)	0.046(0.012)	0.103(0.021)	2.014(0.127)
	x_{12}	$0.5N(4.05, 3^2) + 0.5N(2.86, 2^2)$	0.896(0.014)	0.048(0.012)	0.107(0.022)	2.629(0.183)

Pepe,M.s.,Janes,H.,LongTon,G.,Leisenring,W.and Newcomb,P.(2004). Limitation of the odds ratio in gauging the performance of a diagnostic, or screening marker. *Amenrican Journal of Epidemiology* 159, 882-890.

Zhanfeng Wang and Yuan-Chin Ivan Chang(2010). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* 0, 1-17.

RÉSUMÉ (ABSTRACT)

Area under ROC curve (AUC) and partial area under ROC curve (PAUC) are two popular measures based on ROC curve. AUC is threshold independent while PAUC is threshold dependent in the sense that predetermined specificity or false positive rate (FPR) needs to be specified. In this work, we give a scheme to automatically select the marker with larger PAUC, while maintaining specificity in a practicably low range and without specifying the specificity. The marker with larger ratio of the standard deviation of measurement of case group to the standard deviation of measurement of control group (RSD) will have larger PAUC with FPR in a low range. It is proven under normal distribution assumption, while numerical study is performed under non-normal distribution. The range of the corresponding FPR of PAUC was also estimated. This indicates that RSD can be a good additional measure to AUC to marker selection. We give a scheme to select a marker with AUC.