Quaderni di Statistica
Vol. 14, 2012

Making classifier performance comparisons when Receiver Operating Characteristic curves intersect

Silvia Figini
Department of Economics, Statistics and Laws, University of Pavia
E-mail: silvia.figini@unipv.it

Chiara Gigliarano
Department of Economics and Social Sciences, Marche Polytechnic University
E-mail: c.gigliarano@univpm.it

Pietro Muliere
Department of Decision Sciences, Bocconi University
E-mail: pietro.muliere@unibocconi.it

Summary: The main objective of this paper is to propose a novel approach for model comparisons when ROC curves show intersections. We investigate in a theoretical framework the relationship between ROC orderings and stochastic dominance and we propose alternative indicators that could substitute the common AUC measure.

Keywords: ROC curve, Classification, Stochastic dominance.

1. Introduction

The receiver operating characteristic (ROC) curve describes the performance of a classification or diagnostic rule, while the area under this curve (AUC) is a common measure for the evaluation of discriminative power; see e.g. Krzanowski et al. (2009). When ROC curves cross each other, the AUC measure can lead to biased results and we are not able to select the best model; see e.g. Hand (2009). Common practise is to compare crossing ROC curves by restricting the performance evaluation to proper subregions of scores (see e.g. Thomas, 2009). In our opinion, however, this issue should be more adequately handled in the statistical literature.

The main objective of this paper is, therefore, to propose a novel approach - based on stochastic dominance - for model comparisons, when ROC curves show intersections.

2. ROC curve and stochastic dominance

Consider a classification tool that gives a real-valued score to classify items into two categories: good or bad. Let the random variable X with c.d.f. F represent the score and $x = (x_1, x_2, ..., x_n)$ be a score profile from X with mean $\mu(x)$ and variance $\sigma^2(x)$. Let $\mathcal{X} = \{x : \mu(x) = \mu\}$ be the set of n-dimensional score profiles with mean μ.

Suppose that for a prespecified cut-off c, item i is labeled as *bad* if $x_i \leq c$ and as *good* otherwise. The true positive rate, or sensitivity, is $F_B(c) = Pr(X \leq c|Bad)$, while the false positive rate, or (1 - specificity), is $F_G(c) = Pr(X \leq c|Good)$.[1]

The ROC curve is obtained representing, for any fixed cut-off value, a point in the cartesian plane having as x-value the false positive rate and as y-value the true positive rate. The best curve is the one that is leftmost, the ideal one coinciding with the y-axis. Then the ROC curve is defined as a plot of $\{(u, ROC_X(u)), u \in (0, 1)\}$, where $ROC_X(u) = F_B(F_G^{-1}(u))$.

For sake of model comparisons, performance indicators based on the ROC curve have been proposed, such as the AUC, which is defined as the integrated sensitivity over all specificity ranges: $AUC = \int_{-\infty}^{+\infty} F_B(s)dF_G(s)$.

If the ROC curves do not cross each other, there is an unambiguous comparison of two diagnostic tests in terms of discriminative power and the AUC index provides consistent results. The ordering induced by the ROC curves is equivalent to the first stochastic dominance: $ROC_X(u) \leq ROC_Y(u)$ if and only if $F_B(F_G^{-1}(u)) \leq H_B(H_G^{-1}(u))$, $\forall u \in (0, 1)$, where X and Y represent the score of two different classifiers, with c.d.f. F and H, respectively. In symbols, we write that $X \geq_{FSD} Y$.

In comparing two score distributions, it is of interest to investigate the transformations by which one distribution is obtained from the other. Saying that $X \geq_{FSD} Y$ means that Y is obtained from X by a *first order performance increasing (FOPI) transfer*, according to which the cumulative proportion of bad individuals, increasingly ordered according to their scores, is always higher in Y than in X.

Let us denote *discriminative power index* any function $I : \mathcal{X} \to \mathbf{R}$. The function I satisfies the $FOPI$ principle of transfers if $I(X) \leq I(Y)$ whenever (X, Y) is a $FOPI$ transfer. Obviously, AUC satisfies this principle.

3. Comparing crossing ROC curves

If two ROC curves intersect each other, the first order stochastic dominance fails and it is not possible to employ the AUC index. Thus we move to the second order stochastic dominance (SSD), according to which X dominates Y (in symbols, $X \geq_{SSD} Y$) if $\int_0^z ROC_X(u)du \leq \int_0^z ROC_Y(u)du \ \forall z \in [0, 1]$.

[1] The sensitivity is the probability of correctly classifying a bad item, while the specificity is the probability of correctly classifying a good item.

The SSD can be obtained from a *second order performance increasing (SOPI) transfer*, according to which Y assigns to bad individuals the smallest scores with higher frequency and the highest scores with smaller frequency than X.[2]

Here we focus on the scenario of one crossing and we say that the ROC curve of distribution X *intersects* that of Y *once from below* if and only if there exists $u^* \in (0,1)$ such that $ROC_X(u) \leq ROC_Y(u)\ \forall\, u \leq u^*$ and $<$ for some $u \leq u^*$, and $ROC_X(u) \geq ROC_Y(u)\ \forall\, u \geq u^*$ and $>$ for some $u \geq u^*$. Figure 1 illustrates an example of intersecting ROC curves.

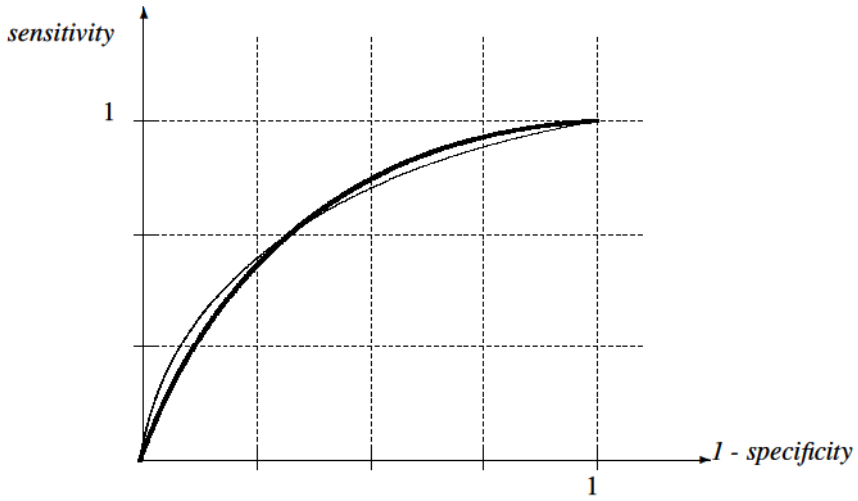

Figure 1. Intersecting ROC curves

Note that if $ROC_X(u)$ intersects once from below $ROC_Y(u)$ and if $\int_0^{u^*}(ROC_Y(u) - ROC_X(u))du \geq \int_{u^*}^1 (ROC_X(u) - ROC_Y(u))du$, then $X \geq_{SSD} Y$.

Since the AUC index may contradict with the criterion of the SSD, alternative measures are required. From Fishburn (1980), we have that the class of indices $I(X) = \int \psi(x) dF_B(x)$, with ψ nondecreasing and concave, is consistent with the SSD. This class of measures provides, therefore, a coherent alternative to the AUC.

If also the SSD is violated, we refer to the third order stochastic dominance: $X \geq_{TSD} Y$ if $\int_0^z \left(\int_0^x ROC_X(u)du\right) dx \leq \int_0^z \left(\int_0^x ROC_Y(u)du\right) dx\ \forall z \in [0,1]$.

The TSD can be obtained from a *third order performance increasing (TOPI) transfer*, according to which in Y a $SOPI$ transfer happens at a higher level of specificity than in X; this criterion thus puts more weigh to smaller false positive rates.

[2] In the income distribution literature, this transfer is called *regressive transfer*.

A discriminative power index I is consistent with the $TOPI$ transfer if and only if $I(Y) \geq I(X)$ with (X, Y) being a $TOPI$ transfer. Note that the AUC index does not satisfy this property.

In case of violation of SSD, it is still possible to compare two crossing ROC curves, provided that the ROC curve corresponding to the score distribution with lower variance intersects once from below the other curve; in particular, if $ROC_X(u)$ intersects once from below $ROC_Y(u)$ and if $\int_0^{u^*}(ROC_Y(u) - ROC_X(u))du \leq \int_{u^*}^1(ROC_X(u) - ROC_Y(u))du$, then $I(Y) > I(X)$ for all $TOPI$ consistent discriminative power indices $I(\cdot)$ if and only if $\sigma^2(y) \geq \sigma^2(x)$.

Following Fishburn (1980), we propose then a class of indices that are consistent with the TSD. More precisely, the class of indicators $I(X) = \int \psi(x)dF_B(x)$, where the function ψ is non-decreasing and concave with a non-negative third derivative, provides an alternative to the AUC measure that is coherent with the $TOPI$ principle of transfers.

4. Concluding remarks

We have provided a novel method for checking for unanimous classifier performance rankings when the ROC curve dominance fails. Our result does not resolve all the ambiguous rankings associated with single crossing ROC curves; it will, however, assist a large number of pairwise comparisons for which the AUC index is not applicable.

Next step of further research will be focused on (i) applying the inverse stochastic dominance theory within the ROC curve framework, (ii) extending the class of discriminative power indices on the basis of the Fishburn (1980)'s results, and finally (iii) providing empirical applications of our methodologies.

References

Fishburn P. (1980), Continua of stochastic dominance relations for unbounded probability distributions, *Journal of Mathematical Economics*, 7, 271–285.

Hand D.(2009), Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning*, 77, 103–123.

Krzanowski, W.J. and Hand, D.J. (2009), *ROC curves for continuous data*, CRC/Chapman and Hall.

Lee W. (1999), Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine*, 18, 455–471.

Thomas L.C. (2009), *Consumer credit models: pricing, profit, and portfolios*, Oxford University Press.