Some results on stochastic comparisons of ROC curves

Silvia Figini, Chiara Gigliarano and Pietro Muliere

Abstract The main objective of this paper is to propose a novel approach for model comparisons when ROC curves show one or more intersections. We investigate in a theoretical framework the relationship between ROC orderings and stochastic dominance, and we propose alternative indicators for performance evaluation.

Key words: Model selection, ROC (Receiver Operating Characteristic), AUC (Area Under ROC curve), Gini Index, 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. [3]. 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. [2]. Common practise is to compare crossing ROC curves by restricting the performance evaluation to proper subregions of scores (see e.g. [7]). 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 for model comparisons, when ROC curves show intersections. Referring to the literature on stochastic dominance, we provide a novel method for checking for unanimous rankings when the ROC curve dominance fails.

Silvia Figini
University of Pavia, e-mail: silvia.figini@unipv.it

Chiara Gigliarano
Marche Polytechnic University, Ancona, e-mail: c.gigliarano@univpm.it

Pietro Muliere
Bocconi University, Milan, e-mail: pietro.muliere@unibocconi.it

2 ROC curve, AUC and Gini index

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 $\mathscr{X} = \{x : \mu(x) = \mu\}$ 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: $AUC = \int_{-\infty}^{+\infty} F_B(s) dF_G(s)$.

Since the ROC curve measures the inequality between the good and the bad score distributions, it seems reasonable to see a connection between the ROC curve and the Lorenz curve; see [3]. Twice the area between the Lorenz curve and the straight line at 45 degree corresponds to the well-known Gini concentration index. This leads to an interesting interpretation of the AUC measure in terms of the Gini coefficient G; more precisely: $G = 2 \cdot AUC - 1$; see [4].

3 ROC ordering and stochastic dominance

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. More specifically, Y has more discriminative power than X, if Y is obtained from X by some kind of performance-increasing transfers.

We say that $X \geq_{FSD} Y$ if and only if 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. In the empirical case, (X,Y) is a *FOPI transfer* if $p_j \leq q_j\ \forall j = 1, ..., m$,

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

where $(p_1,...,p_m)$ and $(q_1,...,q_m)$ are the true positive rates for X and Y in correspondence to each of the m given cut-offs.

Let us denote *discriminative power index* any function $I:\mathscr{X}\rightarrow\mathbf{R}$. 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.

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 $\int_0^z ROC_X(u)du\leq\int_0^z ROC_Y(u)du\,\forall z\in[0,1]$.

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]

Although multiple crossings of ROC curves can occur, in practise they are less common than single intersections. Here we focus on the scenario of one crossing.

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 $u\geq u^*$.

Proposition 1. *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 [1], 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, according to which $X\geq_{TSD}Y$ if $\int_0^z(\int_0^x ROC_X(u)du)\,dx\leq\int_0^z(\int_0^x ROC_Y(u)du)\,dx$ $\forall z\in[0,1]$.

$X\geq_{TSD}Y$ if and only if Y is obtained from X through 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. In [7], indeed, it is stated that "if the curves of the two scorecards overlap then one scorecard is better in one region of scores an the other in another region of scores. (...) Normally one is anxious to accept a large proportion of the goods and so the cut-off scores would tend to be in the area nearer the left of the graph." (page 116).

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 does not satisfy this property.

Proposition 2. *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 $\sigma^2(y)\geq\sigma^2(x)$.*

Proposition 2 states that we can still compare two crossing ROC curves, in case of violation of SSD, provided that (i) the score means are equal and (ii) the ROC

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

curve of the score distribution with lower variance intersects once from below the other curve; for the proof of this result we refer to Theorem 3 in [6]. 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.

Following [1], we propose then a class of indices that are consistent with the TSD, and thus can be used when the ROC curves intersect each other. More precisely, the class of indicators $I(X) = \int \psi(x) dF_B(x)$, where the function ψ is nondecreasing 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 Conclusions

We have provided a novel method for checking for unanimous classifier performance rankings when the ROC curve dominance fails. Our method has the main advantage of establishing whether one distribution can be ranked superior to another according to discriminative power, by looking at the entire score distribution.

Furthermore, we have tested our theoretical proposal on a real data set provided by a rating agency. The application (here not shown for sake of space, but available from the authors upon request) shows that the indices proposed in Section 3 provide coherent results in terms of model selection when ROC curves cross.

Next steps of further research will be focused on (i) applying the inverse stochastic dominance theory within the ROC curve framework (see e.g. [5]), and (ii) extending the class of discriminative power indices on the basis of the results provided in [1].

References

1. Fishburn, P.: Continua of stochastic dominance relations for unbounded probability distributions. J. Math. Econ. **7**, 271–285(1980)
2. Hand, D.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**, 103–123 (2009)
3. Krzanowski, W.J. and Hand, D.J.: ROC curves for continuous data. CRC/Chapman and Hall (2009)
4. Lee, W.: Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. Stat. Med. **18**, 455–471 (1999)
5. Muliere, P. and Scarsini, M.: A note on stochastic dominance and inequality measures. J. Econ. Theory **49**, 314–323 (1989)
6. Shorrocks, A. F. and Foster, J. E.: Transfer sensitive inequality measures. Rev. Econ. Stud. **54**, 485–497 (1987)
7. Thomas, L. C. : *Consumer credit models: pricing, profit, and portfolios*. Oxford University Press.(2009)