(www.interscience.wiley.com) DOI: 10.1002/sim.3761 Published online in Wiley Interscience

Integrated discrimination improvement and probability-sensitive AUC variants

From: Ben Van Calster, Sabine Van Huffel
Department of Electrical Engineering (ESAT-SISTA),
Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10,
B-3001 Leuven,
Belgium

In 2008, Pencina *et al.* [1] suggested the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) as useful indices to assess the added discrimination offered by the addition of a marker to a prediction model. The NRI and IDI provide supplementary information over the difference in the areas under the receiver operating characteristic (ROC) curves (AUCs). The paper and the enlightening discussion instigated by it were highly interesting and relevant to read. The NRI uses risk categories to quantify the reclassification improvement of the new model over the old one. The IDI assesses the improvement in average sensitivity without sacrificing average specificity. This boils down to computing the difference of the average probability of event for events and non-events, and subtracting the obtained values for both models to determine how much difference in the average probabilities has increased in the new model compared with the old model.

In the reactions on the proposal of the NRI and IDI, several comments and caveats were raised. In this letter, we do not want to add to this discussion, but rather want to point at similar indices of the performance that have been suggested in the machine learning literature. It seems that statistical and machine learning researchers, each from their own perspective, have arrived at nearly identical measures.

The rationale behind IDI is that the usefulness of adding a new marker to a model is not adequately evaluated using the AUC because this measure is often sensitive only to huge independent effects of the added marker. In machine learning, researchers wanted to overcome the drawback of the AUC to focus only on the ranking of events and non-events with respect to the probability of event, herewith ignoring the actual probabilities and differences therein. The probability values contain useful information yet are only used to rank the samples. Therefore, two extensions to the traditional AUC were recently suggested that also take into account the actual probabilities [2, 3]. As written in [3], the 'AUC can remain unchanged even when the predicted scores change such that the AUC may fail to detect performance differences between models'.

A first measure is the 'probabilistic AUC', prAUC [2]. By labeling the probability of event as the score (p_e for events, p_n for non-events), the idea is to estimate the score difference between an event and a non-event rather than to estimate the probability that an event is given a higher score than a non-event. The score difference is estimated as the average score for events minus the average score for non-events. The resulting value is the same quantity used for computing the IDI. It falls in the $[-1,+1]$ range, and is linearly transformed so that prAUC yields a number between 0.5 and 1:

$$\text{prAUC} = 0.5 + \frac{1}{2N_eN_n}\sum_{i=1}^{N_e}\sum_{j=1}^{N_n}(p_{e,i} - p_{n,j})$$

with N_e and N_n as the number of events and non-events, respectively. The difference in prAUC between two models, here expressed as Δ, relates to the IDI as follows:

$$\text{IDI} = 2\Delta - 1 = 2(\text{prAUC}_2 - \text{prAUC}_1) - 1$$

The prAUC corresponds to the area under a probabilistic ROC (prROC) curve in which scores are considered as intervals to represent uncertainty in the obtained score. The simplest representation is to define a uniform distribution of width d centered at the obtained score. For a given threshold, the total area above the threshold for events is computed and divided by dN_e to obtain the true positive rate. In this way, a prROC curve is obtained as a smoothed version of the ROC curve. The parameter d is tuned so as to obtain a prROC curve whose area underneath equals prAUC (if more than one value for d satisfies this, the smallest is chosen). prROC is used as a graphical representation of prAUC but at the same time is derived using the prAUC (i.e. the tuning of d).

Copyright © 2010 John Wiley & Sons, Ltd.

LETTER TO THE EDITOR

A second measure [3], the 'scored AUC' (sAUC), is more directly related to the original AUC definition. The sAUC simply extends the AUC by weighting each pair by their score difference:

$$\text{sAUC} = \frac{1}{N_e N_n} \sum_{i=1}^{N_e} \sum_{j=1}^{N_n} C_{i,j}(p_{e,i} - p_{n,j}) \quad \text{with } C_{i,j} = \begin{cases} 1 & \text{if } p_{e,i} > p_{n,j} \\ 0 & \text{else} \end{cases}$$

Correctly ranked pairs ($p_e > p_n$) become more important or less important according to the score difference, whereas tied or incorrectly ranked pairs do not add to the sAUC because either $p_e - p_n$ or $C_{i,j}$ equal zero. The sAUC also corresponds to the area under a curve, called the scored ROC curve [4]. This curve plots a margin-based AUC measure versus τ, where τ is a number between 0 and 1 that is subtracted from p_e. Larger values of τ lower the discrimination between events and non-events. The sAUC then reflects the stability of the standard AUC with increasing margin τ.

It is interesting to know how prAUC and sAUC are correlated to the standard AUC, as well as to quantitative error measures such as the Brier score or cross-entropy (i.e. negative log-likelihood). One study investigated this [5]: the authors used six learning algorithms on 30 data sets (both binary and polytomous problems, and balanced and unbalanced outcomes), and evaluated the performance using cross-validation. The probability-sensitive AUC variants prAUC and sAUC were almost perfectly related (average Spearman rho = 0.97). In turn, these measures were correlated 0.47 and 0.58 with AUC, 0.61 and 0.72 with the Brier score and 0.03–0.14 with cross-entropy, respectively. When using nine learning algorithms to predict the outcome of pregnancies of unknown location (three-class problem, $n = 1003$) [6], we obtained similar correlations based on 100 train-test splits of the data (unpublished results): rho = 0.99 for prAUC versus sAUC; rho = 0.49 and 0.45 for prAUC and sAUC versus multi-class AUC measures such as the volume under the surface [7] and Hand and Till's *M*-index [8]; rho = 0.36 and 0.29 for prAUC and sAUC versus entropy. Multi-class versions of prAUC and sAUC were obtained as a natural extension of the *M*-index [5, 9].

In conclusion, the IDI is related to the difference in prAUC for both models under consideration. The prAUC is almost perfectly correlated with its alternative sAUC, but appears to capture different aspects of model performance when compared with AUC, The brier score or cross-entropy error.

Acknowledgements

Ben Van Calster is a post-doctoral researcher funded by the Research Foundation—Flanders (FWO). Research supported *by Research Council KUL*: GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC); *Flemish Government*: FWO project G.0302.07 (SVM); IWT: TBM-IOTA3; *Belgian Federal Science Policy Office*: IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007–2011).

References

1. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**:157–172. DOI: 10.1002/sim.2929.
2. Ferri C, Flach P, Hernández-Orallo J, Senad A. Modifying ROC curves to incorporate predicted probabilities. *Proceedings of the ICML Workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005. Available from: http://users.dsic.upv.es/~flip/ROCML2005/papers/ferriCRC.pdf.
3. Wu S, Flach P. A scored AUC metric for classifier evaluation and selection. *Proceedings of the ICML Workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005. Available from: http://users.dsic.upv.es/~flip/ROCML2005/papers/wuCRC.pdf.
4. Wu S, Flach P, Ferri C. *An Improved Model Selection Heuristic for AUC*. Lecture Notes in Computer Science, vol. 4701. Springer: Berlin, 2007; 478–489. DOI: 10.1007/978-3-540-74958-5_44.
5. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recognition Letters* 2009; **30**:27–38. DOI: 10.1016/j.patrec.2008.08.010.
6. Van Calster B, Condous G, Kirk E, Bourne T, Timmerman D, Van Huffel S. An application of methods for the probabilistic three-class classification of pregnancies of unknown location. *Artificial Intelligence in Medicine* 2009; **46**:139–154. DOI: 10.1016/j.artmed.2008.12.003.
7. Mossman D. Three-way ROCs. *Medical Decision Making* 1999; **19**:78–89. DOI: 10.1177/0272989X9901900110.
8. Hand DJ, Till RJ. A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning* 2001; **45**:171–186. DOI: 10.1023/A:1010920819831.
9. Van Calster B, Van Belle V, Condous G, Bourne T, Timmerman D, Van Huffel S. Multi-class AUC metrics and weighted alternatives. *Proceedings of the 21st International Joint Conference on Neural Networks (IJCNN)*, Hong Kong, China, 2008; 1391–1397. DOI: 10.1109/IJCNN.2008.4633979.

Copyright © 2010 John Wiley & Sons, Ltd.