A Critical Analysis of Variants of the AUC*

Stijn Vanderlooy[1] and Eyke Hüllermeier[2]

[1] MICC, Department of Computer Science, Maastricht University, The Netherlands
s.vanderlooy@micc.unimaas.nl
[2] Department of Mathematics and Computer Science, Marburg University, Germany
eyke@mathematik.uni-marburg.de

The area under the ROC curve, or AUC, has been widely used to assess the ranking performance of binary scoring classifiers. Given a sample of labeled instances, the metric considers the number of correctly ordered pairs of instances with different class label. Thus, its value only depends on the ordering of the scores but not on the "margin" between them. Consequently, it can happen that a small change in scores leads to a considerable change in AUC value. Such an effect is especially apparent when the number of instances used to calculate the AUC is small. On the other hand, two classifiers can have the same AUC value, even though one of them is a "better separator" in the sense that it increases the difference between scores of positive and negative instances, respectively. It has been argued that this insensitivity toward score differences is disadvantageous for model evaluation and selection. For this reason, three variants of the AUC metric that take the score differences into account have recently been proposed, along with first experimental results.

We present a unifying framework in which the conventional AUC and its variants can be modeled as special cases of a generalized AUC metric. Within this framework, we provide a formal analysis showing that the AUC variants produce estimates of the true AUC with a non-constant, model-specific bias, while the variance can decrease as well as increase. All things considered, the net effect on the quality of the estimations is thus not clear and, hereby, there is no solid theoretical foundation for the variants. Our analysis leads us to conjecture that actually none of the variants should be able to perform better in model selection than conventional AUC. This conjecture is corroborated by extensive experiments with synthetic data as well as real benchmark data, showing that the conventional AUC cannot be outperformed systematically by any variant, not in an ideal setting according to the theoretical analysis, and not in real model selection scenarios. Finally, our contribution also sheds light on recent research dealing with the construction of classifiers that (approximately) optimize the AUC directly, rather than accuracy or another performance metric.

References

1. Vanderlooy, S., Hüllermeier, E.: A critical analysis of variants of the AUC. Machine Learning 72(3), 247–262 (September 2008)

* This is an extended abstract of an article published in the Machine Learning Journal [1].

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, p. 13, 2008.