Is the AUC the Best Measure?

Daniel M. Rice
Rice Analytics, St. Louis, MO (USA), www.riceanalytics.com
Sept 7, 2010

The area under the curve (AUC) that relates the hit rate to the false alarm rate has become a standard measure in tests of predictive modeling accuracy. The AUC is an estimate of the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. For this reason, the AUC is widely thought to be a better measure than a classification error rate based upon a single prior probability or KS statistic threshold.

When tested in a real world application, many models will still use a single threshold based upon the KS statistic or a prior probability rather than the range of thresholds in the AUC calculation. Hence, the AUC may not reflect the expected classification accuracy at this single threshold when these models are put to real world use. Another external validity problem is that the AUC will not assess the extent to which the model output is well calibrated to the target variable, as the AUC does not estimate the accuracy of the probabilities in the model output. In contrast, average squared error will directly reflect the error of the output probabilities. Indeed, there is a long and successful track record of average squared error to assess probability accuracy in areas of science such as weather forecasting back to Brier (1950). In many applications today, even a 1% reduction in classification error or average squared error would mean tens of millions of dollars or more of ROI. Yet, the AUC may miss these effects because of this lack of external validity.

What may be most troubling is that more published simulation data now show AUC estimates of error are less accurate than straight classification error estimates. The earliest simulations were the Huang and Ling (2005) work based upon up to 20 observations. Huang and Ling did report that AUC had better accuracy than classification rate. However, more recent simulations with much more data come to the opposite conclusion. Hanczar et al. (2010) report simulations at various sample sizes up to 1000 observations. They find that straight classification error is a better measure of actual error because the AUC predicted error can have much greater dispersion and is therefore less precise. These AUC inaccuracies were most apparent in imbalanced samples and smaller samples. Based upon these data, Hanczar et al. (2010) urge caution in the use of AUC measures unless the sample size is very large. Unfortunately, they also point out that while "it would be nice to have a simple rule of thumb to determine if a sample is sufficiently large ... no simple solution is possible" (p. 829).

What does this mean for the everyday practitioner? A more comprehensive study now suggests that the AUC may be noisier than previously thought (Hanczar et al. 2010). Other studies are needed, but this recent evidence does not support the superiority of the AUC as a measure of accuracy. Clearly, another problem is that a valid confidence interval for the AUC is not so simple to compute, so a valid repeated measures statistical test for AUC differences between two models built from the same data also would not be simple. In any event, given the apparently greater noise in the AUC, any practice of simply "eyeballing" AUC results might be equivalent to flipping a coin to determine the reliability of differences between models. In contrast to the AUC, well established and simple repeated measures statistical tests can be used to assess straight classification error rate differences or average squared error

differences (see Rice 2008, as an example). Thus, instead of picking a model winner in what could be a random AUC lottery, apparently more accurate measures - straight classification error rate and average squared error – with much better statistical and external validity should probably now be considered.

Postscript: A very nice critique that makes one of the same arguments as made here regarding external validity can be found in Lobo et al. (2008). Professor David Hand who has been doing research on the AUC for a very long time (see Hand and Till, 2001) was kind enough to send us his new articles on this subject on Sept 9, 2010. In the Hand (2009) article that we now reference below, he comes to the same conclusion as we do and as Lobo et al. do that the AUC is fundamentally flawed, although his remedy is different. In any event, Professor Hand's current position is that the AUC is only rarely an appropriate measure of classification performance.

Note added on Feb. 11, 2016. The Matthews Correlation Coefficient (MCC), along with the Brier score, is what we now use in our work for more imbalanced samples to judge the quality of a model, along with other measures like how well it is replicated in terms of its predictions and feature selection. This replication is measured when two modelers are blinded to each other (or one modeler uses a completely automated algorithm) and given independent development samples and the same algorithm is used to build independent models across these independent samples. With imbalanced samples, straight classification error along with hit rate and false alarm rate are used more descriptively to get a sense of when a model is accurate. We find the Brier score to be a useful measure of the accuracy of estimated probabilities across both balanced and imbalanced samples. With highly balanced models like in Rice (2008), we still believe that straight classification error can be useful to judge a best model. Yet, we always use more than one measure and always look at controlled replication measures, as they provide a measure of confidence that is predictive of external validity. We do not use the ROC AUC at all to avoid the artifacts best described now in the work of Professor Hand. Here is link that exemplifies how we measure replication:

http://www.skyrelr.com/replicating-predictions-using-public-uc-irvine-dataset/

REFERENCES

Brier (1950). Verificaton of forecasts expressed in terms of probability. *Monthly weather review* 78: 1–3.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. and Dougherty, E.R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics* 26 (6): 822-830.

Hand, D.J., & Till, R.J. (2001). A simple generalization of the area under the ROC curve to multiple class classification problems. Machine Learning, 45, 171-186.

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

Huang, J. and Ling, C.X. (2005): *Using AUC and Accuracy in Evaluating Learning Algorithms*. *IEEE Trans. Knowl. Data Eng.* 17(3): 299-310.

Lobo, J. M., Jiménez-Valverde, A. and Real, R. (2008), AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17: 145–151. doi: 10.1111/j.1466-8238.2007.00358.

Rice, D.M. (2008), Generalized Reduced Error Logistic Regression Machine, *Section on Statistical Computing - JSM Proceedings 2008*, pp. 3855-3862.