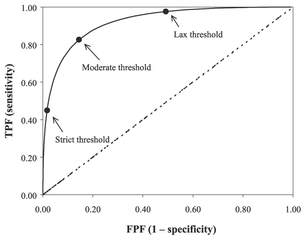
Lies, Damned Lies, and AUC Confidence Intervals

Imran S. Haque[1] and Vijay S. Pande[1,2]

[1]Department of Computer Science and [2] Department of Chemistry, Stanford University, Stanford, CA

BACKGROUND

The receiver operating characteristic (ROC) curve is a useful visualization of the performance of a classifier. Such a curve plots the true positive rate of a classifer as a function of its false positive rate, graphically illustrating the sensitivity-specificity tradeoff in its parameters.



ROC curve, illustrating how different points on the curve represent different trade-offs in the classifier.

As it is a function, the ROC is an unwieldy tool with which to compare different classifiers. For comparisons, the ROC is often reduced to a single number, the "area under the curve", or ROC AUC. The AUC is the integral of the ROC function from FP=0 to FP=1, and is bounded within [0,1]. A classifier performing no better than random will exhibit an AUC of 0.5; a perfect classifier will have AUC=1, and a perfectly wrong classifier will have AUC=0. A classifier with higher AUC than another is considered superior.

However, AUCs are evaluated on a set of test data, which are samples of some underlying data distribution. As such, it is **improper to consider AUCs as point estimates**; a confidence interval must be associated with an AUC to assess whether changes in AUC between classifiers are statistically significant.

What is the best way to assess our confidence in an AUC estimate?

PRIOR METHODS

Analytic Confidence Interval Estimation

Several formulas exist to estimate the variance of the AUC, given various assumptions.

Fixed distribution of positive and negative scores P_x and P_y

$$\sigma_A^2 = \frac{A(1-A) + (m-1)(P_{xxy} - A^2) + (n-1)(P_{xyy} - A^2)}{mn}$$

Any distribution

$$\sigma_{max}^2 = \frac{A(1-A)}{\min(m,n)} \leq \frac{1}{4\min(m,n)}$$

Fixed classifier error rate

It's long and annoying to LaTeX...look up the Cortes paper.

Neither the first nor the third assumptions are applicable to VS: it is not safe to assume a particular distribution of scores, nor to assume a fixed error rate for the method! The second method produces CIs too loose to be useful.

Central Limit Theorem and Standard Deviation

Typical test sets for ligand-based VS will include multiple active queries with which to search a pool of decoys. One simple option is to take the mean and standard deviation of AUCs across each VS experiment (each query).
This approach is statistically unsound:

1. The distribution of the AUC depends on the distribution of the underlying data, and is not inherently normal
2. Each AUC sample (each query) is not an iid sample: in particular, the molecule pool does not change (except for the query, so the samples are not independent).

There is no reason to expect the AUC to be normally distributed; it is neither inherently normal nor the result of a CLT process.

A MODEST PROPOSAL

Ideally, one would be able to test a VS method on the universe of all chemicals, with appropriate labels. Since this is not possible, test sets are used which are assumed to represent a sample from an appropriate distribution of chemical space.

Rather than using an analytic or CLT-based approximation of the CI from a single sample of the distribution, **we propose a nonparametric bootstrap-based estimator of the AUC CI.**

The key benefit of such an estimator is that it **resamples from the distribution, better emulating an independent VS test**; it does assume that the test set is representative of the distribution sampled in real VS, but this is a weak condition (it is equivalent to stating that the test set is useful).

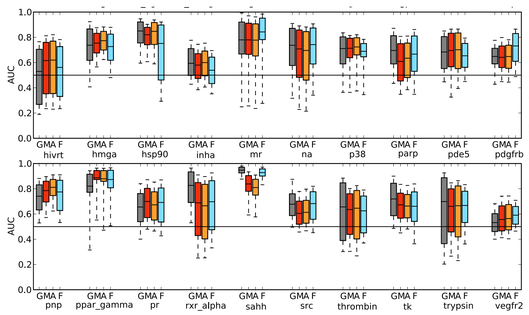
Algorithm for Bootstrapped CIs

Given a system with Na actives and Nd decoys:

1. Select an active uniformly at random
2. Select Na + Nd -1 molecules at random, excluding the molecule from step 1, with replacement
3. Calculate an AUC for screening set 2 against query 1
4. Append AUC to list; repeat until convergence

Calculating 68% and 95% CIs can be done trivially by sorting and counting AUC values; we define convergence for each CI bound (68/95% upper and lower) as the most recent 25 estimates of the bound having stdev < 0.5% of the mean of the same estimates; complete convergence is achieved when all bounds converge.

These bounds on the AUC are often much wider than those predicted by a standard deviation over actives; this is because the standard deviation method inappropriately assumes independence.



AUC comparison for various classifiers on several cases in DUD test set, with bootstrapped CIs.

Note that CIs in general need not be symmetric!

Citations

Cortes C, Mohri M. Confidence Intervals for the Area under the ROC Curve. In NIPS 2004, vol 17 (2005).

Haque IS, Pande VS. PAPER - Accelerating Parallel Evaluations of ROCS. J. Comp. Chem 31(1), 117-132 (2010)

Acknowledgments

Thanks go to John Chodera and Kim Branson for helping work out the bootstrap method used in this poster.