Pointwise ROC Confidence Bounds: An Empirical Evaluation

Sofus A. Macskassy SMACSKAS@STERN.NYU.EDU
Foster Provost FPROVOST@STERN.NYU.EDU
New York University, Stern School of Business, 44 W. 4th Street, New York, NY 10012

Saharon Rosset SROSSET@US.IBM.COM
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Abstract

This paper is about constructing and evaluating pointwise confidence bounds on an ROC curve. We describe four confidence-bound methods, two from the medical field and two used previously in machine learning research. We evaluate whether the bounds indeed contain the relevant operating point on the "true" ROC curve with a confidence of $1-\delta$. We then evaluate pointwise confidence bounds on the region where the future performance of a model is expected to lie. For evaluation we use a synthetic world representing "binormal" distributions–the classification scores for positive and negative instances are drawn from (separate) normal distributions. For the "true-curve" bounds, all methods are sensitive to how well the distributions are separated, which corresponds directly to the area under the ROC curve. One method produces bounds that are universally too loose, another universally too tight, and the remaining two are close to the desired containment although containment breaks down at the extremes of the ROC curve. As would be expected, all methods fail when used to contain "future" ROC curves. Widening the bounds to account for the increased uncertainty yields identical qualitative results to the "true-curve" evaluation. We conclude by recommending a simple, very efficient method (vertical averaging) for large sample sizes and a more computationally expensive method (kernel estimation) for small sample sizes.

1. Introduction

In this paper we address the problem of creating pointwise confidence bounds on ROC curves. Increasingly, machine learning studies plot ROC curves to assess possible trade-offs of true-positive and false-positive rates to be expected from a learned model. In machine learning, confidence bounds rarely are drawn on ROC curves, and the field generally is unaware of methods (introduced elsewhere) to produce such bounds. There has been almost no research on the assessment of confidence bounds on ROC curves, and no research in a machine learning context (with the exception of a previous workshop paper (Macskassy & Provost 2004)).

Pointwise confidence bounds generally are designed to contain (with probability $1-\delta$) the expectation of the point being estimated. The analog for ROC curves is that the expected or "true" ROC curve should be contained within the confidence bounds, with the designated probability. To contrast with what follows, we call these "true-curve" confidence bounds. When the points in question lie on a curve in multidimensional space, it is important to consider the dimension(s) along which the points are bounded, and associated semantics of the bounds. In ROC space, most existing methods compute bounds parallel to the axes, which correspond to assessing the confidence that for a given FP rate, the model's TP rate is expected to be contained within the bound (or vice versa).[1]

In machine learning settings, it may be desireable to place a bound on the future performance of a scoring model: where should we expect the relevant operating point on an ROC curve to lie given a future sample. To generate such "future" confidence bounds, we need (1) to adjust the bounds to take into account the increased uncertainty, and (2) to take into account the size of the future sample. The latter is important because the variance of an ROC curve depends on the number of data on which it is based (Macskassy & Provost, 2004).[2] Specifically, we will generate bounds that are expected to contain $(1-\delta)$% of the appropriate points

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

[1] As pointed out by Provost et al. (1998), these pointwise bounds may not be appropriate for common machine learning evaluations, e.g., choosing the operating point corresponding to the minimum expected-cost calculation (Provost & Fawcett, 2001). It may be more appropriate to compute confidence *bands* about the entire curve. We discuss ROC confidence bands in a companion paper at this workshop—a paper that also appears in the main ICML conference (Macskassy et al., 2005).

[2] For highly unbalanced class distributions, it also is critically dependent on the number of minority class instances (Stein, 2002).

on ROC curves produced from data sets containing r examples.

There has been very little research on the assessment of confidence regions for ROC curves, even for their designed purpose, although a few pieces of work did perform empirical evaluations of the efficacy of their methods (e.g. (Claeskens et al., 2003; Hall et al., 2004)). For ROC analysis, it is sufficient to represent a (learned) model simply by the class-conditional score distributions it produces (G^+ and G^-). For this paper, we adopt the conventional ("binormal") assumption that G^+ and G^- are normally distributed, and assess the containment of the bounds. We further assume that it is desired to compute an ROC curve for a learned scoring model, rather than for a learning algorithm. The latter also is important, but we treat the simpler question here.

For the "true-curve" bounds, all methods are sensitive to how well the distributions are separated, which corresponds directly to the area under the ROC curve. One method produces bounds that are universally too loose, another universally too tight, and the remaining two are close to the desired containment although containment breaks down at the extremes of the ROC curve. As would be expected, all methods fail when used to contain "future" ROC curves. Widening the bounds to account for the increased uncertainty yields identical qualitative results to the "true-curve" evaluation. We conclude by recommending a simple, very efficient method (vertical averaging) for large sample sizes and a more complex and computationally expensive method (kernel estimation) for small sample sizes.

2. Generation of Pointwise ROC Confidence Bounds

In machine learning research, prior work on creating confidence intervals for ROC curves has for the most part been in the context of creating pointwise confidence intervals. *Vertical averaging* (VA) looks at successive FP rates and averages the TPs of multiple ROC curves at that FP rate (Provost et al., 1998). A potential weakness of this method is the practical lack of independent control over a model's false-positive rates (Fawcett, 2003). *Threshold averaging* (TA) chooses a set of decision thresholds and, for each, identifies the mean and standard deviation (along both FP and TP) of the set of ROC points that would be generated using the threshold. Below, we select a uniformly distributed subset from the sorted set of all scores observed across the set of ROC curves in the sample (Fawcett, 2003).

VA and TA both require a set of ROC curves to generate their confidence regions. These can be generated by evaluating the model on multiple, sampled fitting sets or by resampling one fitting set. In this paper, we take an observed sample set R and repeatedly resample with replacement sets $R*$. The resulting ROC curves from R and $R*$ will be used to generate confidence regions about an average curve. Use of such resampling (the bootstrap (Efron & Tibshirani, 1993)) as a robust way to evaluate expected performance has been suggested for evaluating cost-sensitive classifiers (Margineantu & Dietterich, 2000).

Medical researchers have examined the use of ROC curves extensively and have introduced many techniques for creating confidence boundaries (intervals or bands) (Beck & Shultz, 1986; Zweig & Campbell, 1993; Hilgers, 1991; Ma & Hall, 1993; Campbell, 1994; Metz et al., 1998; Tilbury et al., 2000; Claeskens et al., 2003; Hall et al., 2004). We focus here on two methods that are directly applicable to the generation of pointwise confidence bounds for continuous score distributions. One method (Ma & Hall, 1993) is based on the Working-Hotelling hyperbolic confidence bounds for regression lines (Working & Hotelling, 1929). The other method is based on kernel estimation, which recently has seen increasing use to estimate points and their variances in ROC space (cf. (Hall et al., 2004)).

2.1. Vertical Averaging (VA)

Given a false-positive rate (FP), the *vertical averaging* (VA) method works as follows: sample the distribution of true-positive rates (TPs) from a collection of ROC curves at the given FP. Given this distribution a TP confidence interval can be created using either the empirical distribution, a fitted Gaussian distribution, or the binomial distribution. We use the empirical distribution in this paper—the fitted Gaussian bounds had similar containments in preliminary experiments, but we do not report on those here. Using the binomial distribution yielded bounds that were far too narrow and had close to 0 containment; we therefore do not report results using it. We generate VA bounds for a range of false-positive values between 0 and 1 and evaluate the bound for each FP value separately.

2.2. Threshold Averaging (TA)

Given a threshold, the confidence bound calculation for the *threshold averaging* (TA) method works by sampling the distribution of ROC points generated at the given threshold. It then generates the mean (FP,TP) point for the sampled threshold and finds the confidence intervals of the FPs and TPs. We use the empirical distribution for the same reason as for the vertical average method. We generate TA bounds for 100 thresholds, uniformly distributed among the observed thresholds and evaluate the bound for each threshold separately.

2.3. Pointwise Working-Hotelling Bounds (WHB-p)

Following Ma and Hall (1993) and Metz et al. (1998), we adapt a method for using Working-Hotelling hyperbolic bounds (Working & Hotelling, 1929) to generate pointwise confidence bounds on an ROC curve. We use a modified version a publicly available implementation of the LABROC4 algorithm (Metz et al., 1998)[3] The method is too complex to describe in detail here; we will give an intuitive overview and the interested reader is refereed to the original sources.

Previously, much work on generating ROC curves in the medical literature dealt with ordinal decision categories, notably estimating ROC curves using maximum likelihood (ML) estimation based on an assumed parametric form for the ROC curve. However, we are interested in continuous decision scores (e.g., estimates of the probability of class membership). Metz et al. observed that ML estimation of an ROC curve from continuous scores is equivalent to ML estimation from ordinal scores if runs of positives/negatives (as well as equal-scored cases) in the rank-ordered data are interpreted as ordinal categories. LABROC4 first groups the data into such runs. Then assuming a binormal score distribution it uses an ordinal ('rating method') algorithm (Dorfman & Alf, 1969) to fit a smooth ROC curve. Two different notions of binormality are taken by this approach. One, which we use later, is that the class-conditional score distributions G^+ and G^- are normally distributed. The second is that the ROC curve is a straight line using "normal-deviate" axes—the so-called "probit" space; that is, $\Phi^{-1}(TP) = a + b\Phi^{-1}(FP)$, where $\Phi(\cdot)$ represents the cumulative normal distribution function and TP and FP are the true- and false-positive rates. This straight line in probit space corresponds to a smooth curve in ROC space.

Ma and Hall (1993) describe the construction of different sorts of confidence bounds for such ROC curves. Following their line of reasoning, the LABROC4 program generates pointwise confidence bounds via the ROC regression line in probit space, which is fit using maximum-likelihood estimation (MLE). Specifically, the bounds are composed of points defined by the function l:

$$l(x, k) = a - b \cdot x + k \cdot \sigma(x), \tag{1}$$

where k is a constant defined below, positive for the upper bound and negative for the lower bound, x is a probit-transformed false-positive rate, and $\sigma(x)$ is the estimated variance of the prediction at x, using the standard linear regression inference methodology.

The constants $\pm k$ are determined by the confidence level $(1 - \delta)$ and the type of bound being generated. To generate pointwise confidence bounds, we use Ma and Hall's pointwise Working-Hotelling bounds, where, $k_\delta = \Phi^{-1}(1 - \delta/2) = z_{\delta/2}$ (the z-score for a two-sided bound of δ confidence).

[3]We acquired the LABROC4 FORTRAN source code from a public web-site and modified its I/O to work with our ROC analysis toolkit. Our Java 1.5 toolkit will be released to the public later this year.

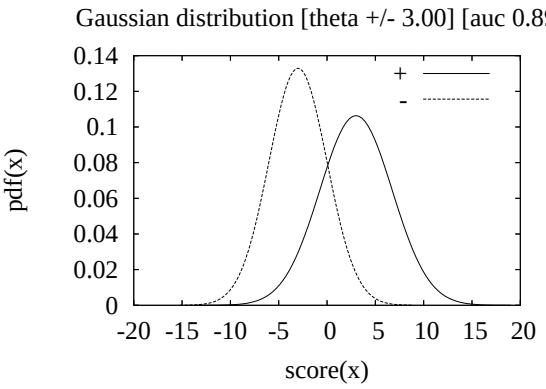

Figure 1. Example distribution used in study below.

2.4. Kernel Estimation (KE)

Recently there has been an increased focus on the use of kernel estimation to generate ROC curves and confidence bounds around them (cf. (Hall et al., 2004)). Kernel estimation (KE) is used to estimate a continuous density function from a discrete observed score distribution using some kernel function $K(x)$. We refer the reader to the original paper for the details of this method.[4]

3. Data Generator

To evaluate the different confidence bounds, we generate G^+ and G^- as two normal distributions, only differing in their parameters. Our synthetic world $\mathcal{W}$ is defined by five parameters:

1. $P(+)$, the probability that an instance is from G^+;
2. the two model parameters for G^+: θ^+ and σ^+;
3. the two model parameters for G^-: θ^- and σ^-.

For the study below, we fix $P(+) = 0.5$, $\sigma^+ = 3.75$, and $\sigma^- = 3.0$, making G^+ "fatter" than G^- (following an observation of Bennett (2003), discussed below). We used a range of values of θ, setting $\theta^+ = \{0.75, 1.00, 1.50, 2.00, 3.00, 4.00, 5.00\}$, and $\theta^- = -\theta^+$.

[4]We use the R script written by the authors to generate the bounds. This script is publicly available from Hyndman's website. We wrote an R script wrapper to tailor the I/O to our ROC toolkit.

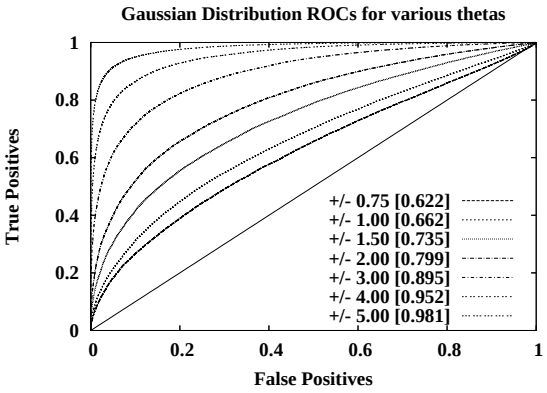

Figure 2. ROC curves generated for distribution as we vary θ.

Figure 1 shows the distributions with $\theta = \pm 3.0$. Figure 2 shows the resulting ROC curves for all values of θ, generated by plotting the points $(\text{cdf}_{G^-}(x), \text{cdf}_{G^+}(x))$, for x ranging from ∞ down to $-\infty$. The smaller θ, the closer the true ROC curve will be to the random line ($x = y$); these choices of θ yield a range of AUCs from 0.62 to 0.98.

4. "True" Bound Evaluation

The semantics for the "true" confidence bound is that we would expect the relevant operating point on the "true" ROC curve to lie within the bound with the specified probability (frequency). In other words, were we to generate a large number of bounds, we would expect that $(1-\delta)$ of them contain the "true" operating point.

We generate the bounds using the straightforward methodology outlined in Table 1.

1. Build a synthetic world, $\mathcal{W}$, consisting of two distributions, G^+ and G^- with means θ and $-\theta$ respectively.
2. Fix a sampling size, r, and sample from $\mathcal{W}$ a confidence-generation set, R, of size r.
3. Generate confidence bounds, C_b, based on R as outlined in Section 2.

Table 1. Generating ROC Bounds from Synthetic World.

This methodology has three parameters: (1) the synthetic world, which is defined by G^+, G^-, and $P(+)$, (2) the ROC-generation size, r, and (3) the confidence δ.

4.1. Evaluation

We described the synthetic worlds we will use in Section 3. We fix $\delta = 0.1$ and examine the sensitivity of the confidence calculations to the ROC-generation size, $r \in \{25, 100, 250, 1000, 2500, 10000\}$ and the synthetic world used. To evaluate the efficacy of the bounds, we generate 1000 bounds based on the method shown in Table 1, and count how many of those contained the true operating point. We should expect that $(1-\delta)\%$ of the calculated bounds contain the "truth".

4.2. Results

Figure 3 shows the containment for the 4 pointwise bound methods across various values of θ for $r = 10000$. For low thresholds (which correspond to high values of FP) the TA bounds are universally too wide, whereas for high thresholds (low FP) the bounds are sensitive to how well the distributions are separated—*i.e.*, the TA bounds fall below the desired containment in direct relation to the increasing values of θ. The containments of WHB are very erratic.

KE and VA have almost identical patterns. Both methods yield, on average, bounds with close to the desired containments. Exceptions include dips around FP$= 0.15$ for $\theta \in \{4, 5\}$, at FP=$(0.6, 0.7)$ for $\theta = 1.5$ and a very large dip at FP=0.90 for $\theta = 5$. For $\theta \in \{0.75, 1\}$, both methods have containments around 0.95 for all FP. The bounds are for the most part slightly too wide at any given point. These results indicate that with a large data set using the VA bounds generally would be preferable, given the computational expense required to generate the KE bounds.

Exploring the sensitivity to data set size, Figure 4 shows the containments for each of the 4 methods across various values of r at $\theta = 3.0$.[5] We see the same pattern for TA as before—TA bounds are for the most part too wide, then drop below the desired containment at high threshold values. WHB shows more consistency across r values, with somewhat better containment at lower FP values, but shows consistently poor performance for larger FP values.

KE and VA exhibit different containment patterns for small data sets and larger FP values. The VA bounds become far too narrow for small data sets. The KE bounds become too inclusive, except for very large values of FP. This is in line with containment patterns reported on the original KE work (Hall et al., 2004). Assuming that one generally would prefer to be overly conservative in creating confidence bounds (avoiding Type I errors in inference), the KE bounds would be recommended for small data sets.

5. "Future" Bound Evaluation

For machine-learning evaluations we may want to bound the future performance of the model. Moreover, to evaluate the bounds on (real) data for which the score distri-

[5] We see similar containment patterns for other values of θ.

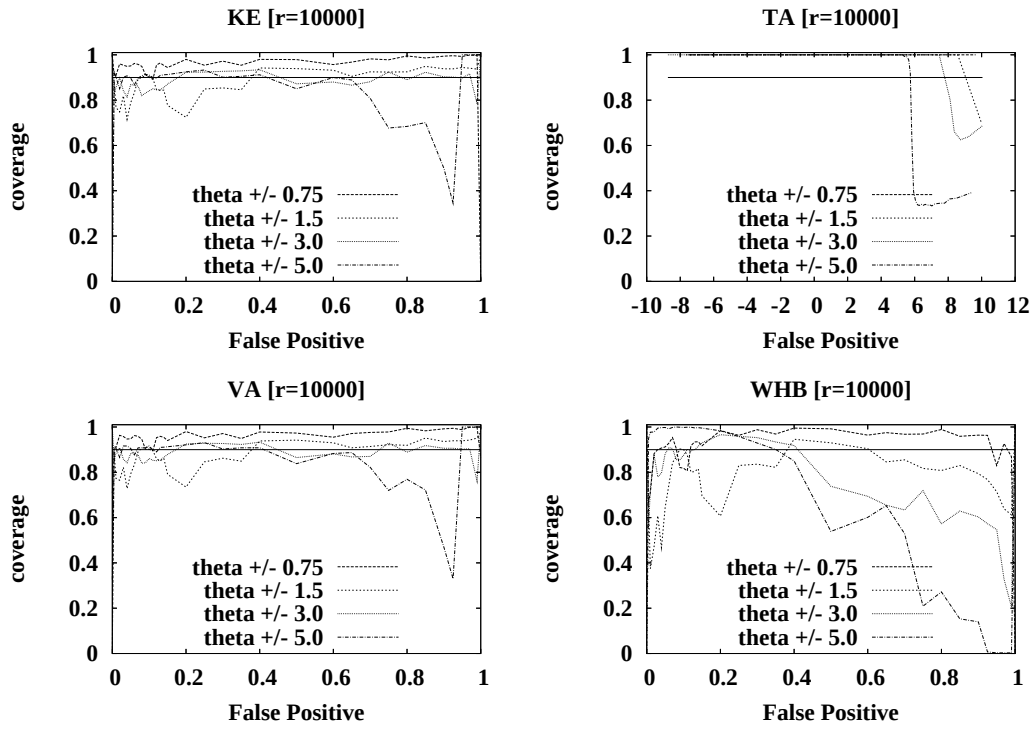

Figure 3. Coverage of pointwise "truth" bounds at $\delta = 0.1$ and $r = 10000$. The horizontal line shows the expected coverage. We show the coverages for various values of θ.

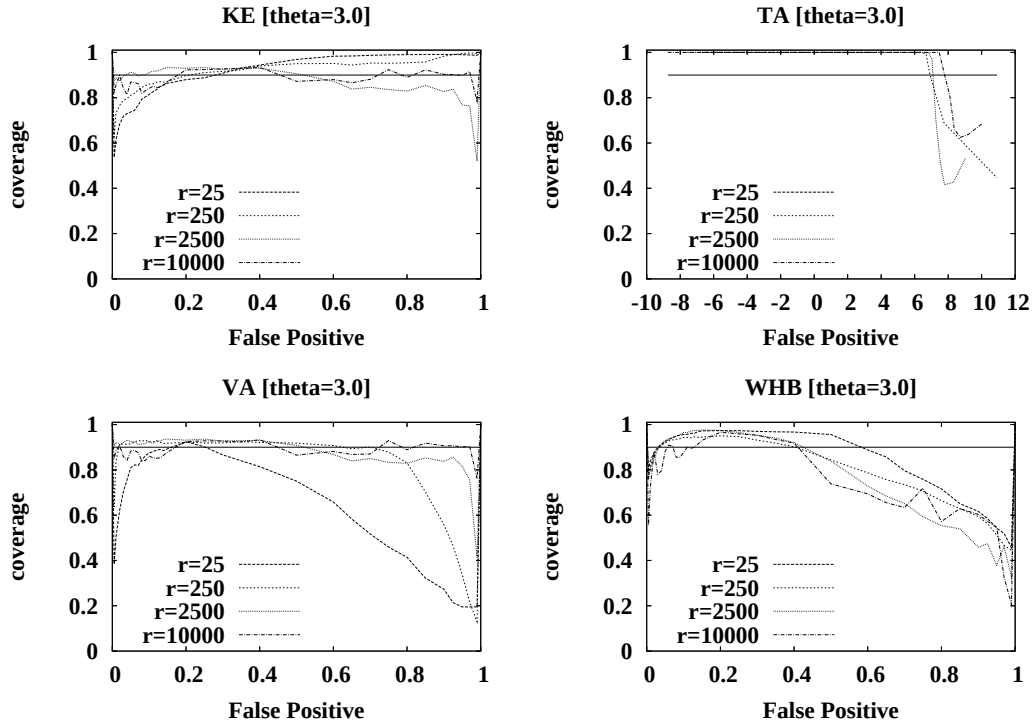

Figure 4. Coverage of pointwise "truth" bounds at $\delta = 0.1$ and $\theta = 3.0$. The horizontal line shows the expected coverage. We show the coverages for various values of r.

butions are not binormal,[6] we need an alternative strategy. We don't know the true ROC curve, but we often have sufficient data to assess the containment of future ROC curves. To evaluate "future-curve" confidence bounds, for each bounded curve we generate 1000 additional ROC curves based on r samples from $\mathcal{W}$ and count how many of the relevant operating points were contained by the bounds, where r is the same size as that used to generate the bounds. Ideally, $1-\delta$ of the generated operating points would fall within the bounds.

Not surprisingly KE, VA and WHB all fail. TA is still too wide. Each method places bounds about the observed curve. Even if the methods are estimating the true variance correctly, future curves will be distributed about the true curve, not about the observed curve. See Macskassy et al. (2005) for further explanation.

5.1. Widening the bounds

Following Macskassy et al. (2005), we address this problem by widening the bounds. Let us consider the true ROC curve (R_T), the sample ROC curve (R_M) from which we will calculate the bounds (B_M) of width w, and an ROC curve sampled subsequently ($R_{M'}$) the appropriate points of which should with probability $(1-\delta)$ lie within B_M.

Assume that we have a correct true-curve bound around R_M, and denote the width by w. The distance between R_T and R_M in the chosen direction (TP for three of our bounds) has probability $(1-\delta)$ of being smaller than w. Denote this distance by d(R_T, R_M). The distance measure for R_T and $R_{M'}$, d(R_T, $R_{M'}$), follows the same distribution and is independent. Now, if we assume that this distance has a Gaussian distribution, then it is easy to verify that:

$$P(d(R_T, R_M) + d(R_T, R_{M'}) \leq \sqrt{2}w) = 1 - \delta.$$

With non-Gaussian, but "reasonable" distributions, this should still hold approximately. Since $d(R_M, R_{M'}) \leq d(R_T, R_M) + d(R_T, R_{M'})$ we expect the resulting bounds to be a little too wide, but this could be offset somewhat by additional uncertainties not accounted for by our methodology, such as non-Gaussianity, etc.

5.2. Results

Figure 5 shows the containments we get from applying this technique using various values of θ for $r = 10000$. We see a close correspondence to the results for the "true-curve" bounds. KE and VA have similar containments and TA and WHB fail in the same manner as before. Notably, KE and VA are more robust than they were for the "true-curve" bounds, not having as many dips. They are both generally a little too wide, as suggested above. Note that both here and above, even the better methods are too inclusive for very-difficult-to-separate score distributions (low AUCs).

We performed the same sensitivity analysis to r as before. Figure 6 shows the containments we get at $\theta = 3.0$ across various values of r. Again we see close correspondence to the performances of the "true-curve" bounds. KE has very good containment across the board and VA is worse for smaller values of r, but for $r \geq 500$ it performs comparably to KE. We also see that TA is again too wide and WHB performs even worse than before, being everywhere too narrow.

Figures 5 and 6 clearly show that only one of the methods, KE, shows consistently appropriate containment. However, for larger data sets VA performs comparably and therefore it may be recommended—for computational reasons—if r is large enough (≥ 500 for these experiments).

6. Discussion and Limitations

In this paper we assessed various methods for generating pointwise confidence bounds for ROC curves. We evaluated two types of bounds: (1) containment of points on the "true" ROC curve, and (2) containment of points on future ROC curves produced by the same model (for a test set of a particular size). The former evaluation is based on what the bounds were designed to do. The latter may be appropriate for real settings, when a practitioner may need to bound the performance expected of a model in the future (see also the evaluation by Macskassy et al. (2005)).

For computing a confidence bound on where the operating points on the "true" ROC curve for a model lies, we saw that a resampling technique (VA) and a semi-parametric technique (KE) both performed comparably for large sample sizes, and that the kernel estimation technique (KE) was much more robust for smaller sampling sizes. Because KE is computationally expensive, VA should be the method of choice unless r is small. The two other techniques, TA and WHB, were either far too conservative (TA) or too tight (WHB). This is somewhat surprising for the parametric WHB method, since the underlying parametric assumptions seem to hold, and deserves further analysis.

We proposed a method to widen the bounds to account for the added uncertainty incorporated in "future" bounds. Using the widened bounds, KE was always very close to the proper containment regardless of θ and r, VA broke down for small r but was otherwise comparable to KE, TA was always too conservative and WHB was always too tight. Therefore, KE is the method of choice unless computational expensive is an issue, in which case VA can be used unless r is small.

[6] Which we do not do in this paper for pointwise bounds, but see Macskassy et al. (2005) for a similar evaluation for confidence bands.

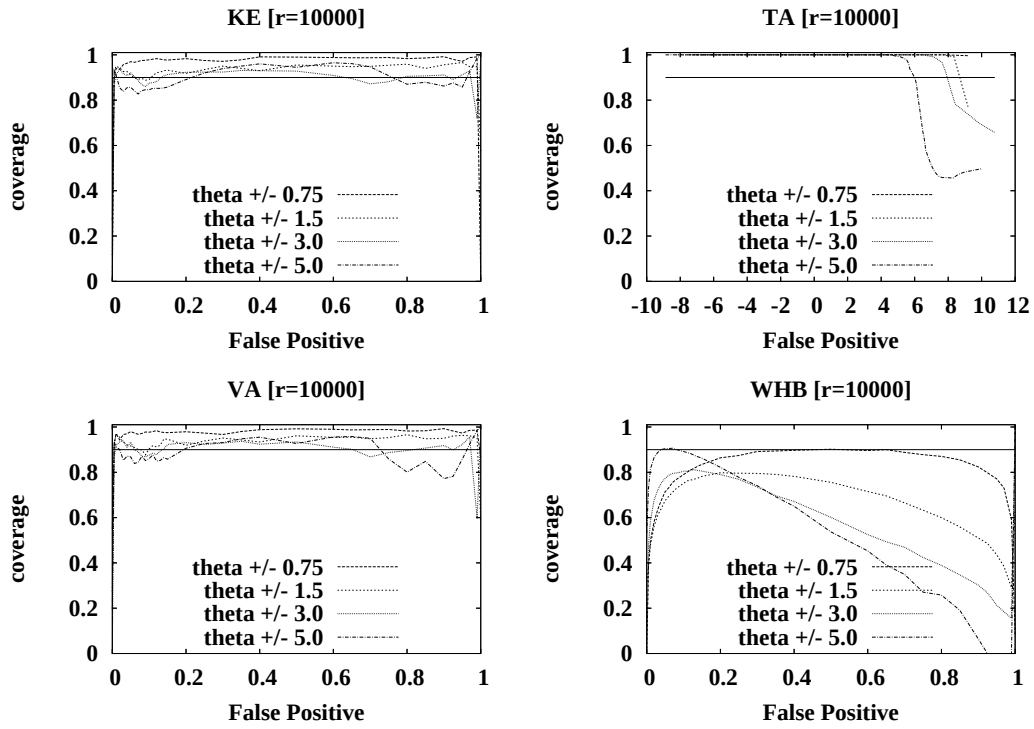

Figure 5. Containment of pointwise "future" bounds at $\delta = 0.1$ and $r = 10000$. The horizontal line shows the expected containment. We show the containments for various values of θ.

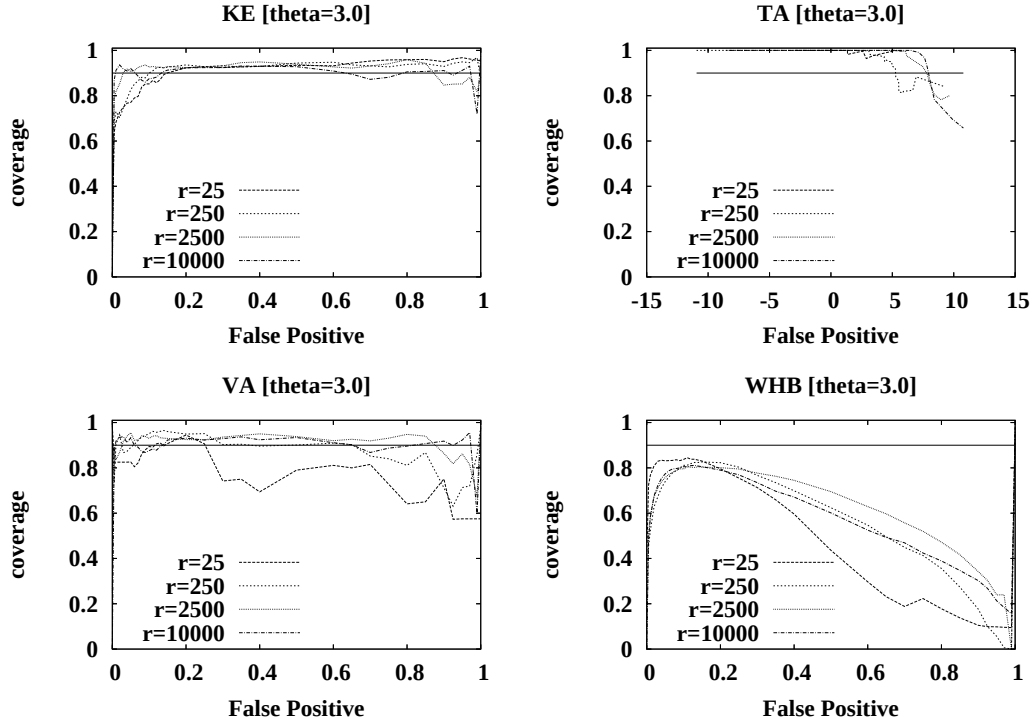

Figure 6. Containment of pointwise "future" bounds at $\delta = 0.1$ and $\theta = 3.0$. The horizontal line shows the expected containment. We show the containments for various values of r.

Our study is based only on synthetic data in an ideal setting. Prior research (Bennett, 2003; Macskassy et al., 2005) has shown that scoring distributions from learned models *do not* follow Gaussian distributions. Whether the current results will carry over to real data the subject of our ongoing investigation—however, given that they do carry over for confidence bands (Macskassy et al., 2005), it is not unreasonable to expect a similar result for pointwise confidence bounds.

One issue which we did not cover here is the effect of class skew, which requires investigation. Stein (2002) has shown that for data sets with a large class imbalance, the variance in ROC curves is extremely sensitive to the size of the minority class. Macskassy et al. (2005) show that for the highly unbalanced Letter-A data set, confidence band containment tends to fail. Therefore, caution should be taken in extrapolating our results to data sets with relatively few examples of one class.

Acknowledgments

We would like to thank Tom Fawcett for his pointers to related work and for many discussions about ROC curves, an anonymous early reviewer for directing us to additional medical literature we were unaware of, Michael Littman for initial discussions on ROC evaluations, Haym Hirsh for his feedback early in the design stages and Matthew Stone who initially suggested using the bootstrap for evaluating ROC curves.

This work is sponsored in part by the National Science Foundation under award number IIS-0329135. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Science Foundation or the U.S. Government.

References

Beck, J. R., & Shultz, E. K. (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology and Laboratory Medicine*, *110*, 13–20.

Bennett, P. N. (2003). Using Asymmetric Distributions to Improve Text Classifier Probability Estimates. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada: ACM Press.

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, *13*, 499–508.

Claeskens, G., Jing, B.-Y., Peng, L., & Zhou, W. (2003). Empirical likelihood confidence regions for comparison distributions and roc curves. *The Canadian Journal of Statistics*, *31*, 173–190.

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–rating method data. *Journal of Mathematical Psychology*, *6*, 487–496.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fawcett, T. (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers* (Technical Report HPL-2003-4). HP Labs.

Hall, P. G., & Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating roc curves. *Statistics and Probability Letters*, *64*, 181–189.

Hall, P. G., Hyndman, R. J., & Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, *91*, 743–750.

Hilgers, R. A. (1991). Distribution-free confidence bounds for ROC curves. *Methods of Information in Medicine*, *30*, 96–101.

Ma, G., & Hall, W. J. (1993). Confidence bands for receiver operating characteristic curves. *Medical Decision Making*, *13*, 191–197.

Macskassy, S., & Provost, F. (2004). Confidence Bands for ROC Curves: Methods and an Empirical Study. *Proceedings of the First Workshop on ROC Analasis in AI (ROCAI-2004) at ECAI-2004*.

Macskassy, S., Provost, F., & Rosset, S. (2005). ROC Confidence Bands: An Empirical Evaluation. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. Bonn, Germany.

Margineantu, D. D., & Dietterich, T. G. (2000). Bootstrap methods for the cost-sensitive evaluation of classifiers. *International Conference on Machine Learning, ICML-2000* (pp. 582–590).

Metz, C. E., Herman, B. A., & Roe, C. A. (1998). Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets. *Medical Decision Making*, *18*, 110–121.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*, 203–231.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453). San Francisco, CA: Morgan Kaufman.

Stein, R. (2002). *Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation* (Technical Report #030124). Moody's KMV.

Tilbury, J., Eetvelt, P. V., Garibaldi, J., Curnow, J., & Ifeachor, E. (2000). Receiver operating characteristic analysis for intelligent medical systems – a new approach for finding non-parametric confidence intervals. *IEEE Transactions on Biomedical Engineering*, *47*, 952–963.

Working, H., & Hotelling, H. (1929). Application of the theory of error to the interpreation of trends. *Journal of the American Statistical Assocation*, *24*, 73–85.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561–577.