ROC Confidence Bands: 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 confidence bands around ROC curves. We first introduce to the machine learning community three band-generating methods from the medical field, and evaluate how well they perform. Such confidence bands represent the region where the "true" ROC curve is expected to reside, with the designated confidence level. To assess the containment of the bands we begin with a synthetic world where we know the true ROC curve—specifically, where the class-conditional model scores are normally distributed. The only method that attains reasonable containment out-of-the-box produces non-parametric, "fixed-width" bands (FWBs). Next we move to a context more appropriate for machine learning evaluations: bands that with a certain confidence level will bound the performance of the model on future data. We introduce a correction to account for the larger uncertainty, and the widened FWBs continue to have reasonable containment. Finally, we assess the bands on 10 relatively large benchmark data sets. We conclude by recommending these FWBs, noting that being non-parametric they are especially attractive for machine learning studies, where the score distributions (1) clearly are not normal, and (2) even for the same data set vary substantially from learning method to learning method.

1. Introduction

Many machine learning studies plot ROC curves to illustrate the possible tradeoffs of true-positive and false-positive rates that would be expected from a learned model. This paper addresses the problem of creating confidence bands around such ROC curves.[1] Confidence intervals generally are designed to contain (with probability $1-\delta$) the expectation of a function being estimated. For ROC curves this amounts to specifying a region of ROC space where some ROC curve of interest is expected to lie. For example, given a scoring model and a domain of interest, rather than simply plotting an ROC curve for a particular sample, it may be more informative to show the region expected to contain the "true" ROC curve—the ROC curve defined by the model and the distribution generating the data. We will call these "true-curve" confidence bands.

In machine learning research (and practice), confidence bands rarely are drawn on ROC curves, and the field generally is unaware of methods (introduced elsewhere) to produce such bands. There has been almost no research on the evaluation of confidence bands for ROC curves, and no research in a machine learning context (with the exception of the workshop paper that we extend here (Macskassy & Provost 2004)).[2] We first introduce the machine learning community to three existing methods from the medical literature for estimating confidence bands on ROC curves. We then assess the containment of these bands.

In a machine-learning setting with real data, we do not know the true ROC curve for a particular learned model, which stymies evaluations of true-curve containment on real data. However, being accustomed to estimating expected *future* performance, it is natural to evaluate whether confidence bands properly contain the ROC curves produced by a particular model on future data from the same domain. For this we will need to adjust the true-curve confidence bands to account for the added uncertainty in the composition of the future data. Furthermore, to generate these "future-curve" confidence bands, we also must take into account the size of the data set used to generate the ROC curve, because this influences the variance of the ROC curve (Macskassy & Provost, 2004). In sum, we want to generate a "future-curve" band that with a probability of $1-\delta$ will contain the ROC curve traced by the model on a future data set containing r examples.

Another issue is whether the bands are created for a specific, fixed model (perhaps a learned model), where variation comes only from the test data, or whether we are interested in bounding the performance of a learning algorithm, given different training data sets and different test data sets. While the latter problem is certainly important, we concentrate here on the simpler, more tractable problem of evaluating a fixed model.

[1] We are not considering *pointwise* confidence bounds in this paper. We discuss these elsewhere (Macskassy et al., 2005).

[2] Extending the prior workshop paper, here we clarify many details of the various methods, evaluate true-curve containment, introduce adjusted curves for future-curve evaluation, evaluate future-curve containment with a suite of real data sets, and clearly recommend one method.

Appearing in *Proceedings of the 22^{nd} International Conference on Machine Learning*, Bonn, Germany, 2005. Copyright 2005 by the author(s)/owner(s).

For ROC analysis, it is sufficient to represent a (learned) model simply by the class-conditional score distributions it produces (G^+ and G^-). We begin by adopting the conventional assumption that G^+ and G^- are normally distributed, and assess the containment of the true-curve bands. We show that one of the three methods, non-parametric fixed-width bands (FWBs), outperforms the others. We next introduce an adjustment to widen the FWBs so that they are appropriate as future-curve bands, and demonstrate that the widened FWBs continue to have reasonable containment. Finally, we assess the bands on 10 relatively large benchmark data sets.

We conclude by recommending FWBs, noting that being non-parametric they are especially attractive for machine learning studies, where the score distributions (1) clearly are not normal, and (2) even for the same data set vary substantially from learning method to learning method.

2. Confidence bands on ROC curves

Prior work in machine learning on creating confidence intervals for ROC curves for the most part has created one-dimensional, pointwise confidence intervals (cf. (Bradley, 1997; Provost et al., 1998; Fawcett, 2003)), which are not the focus of this paper. Many methods in the medical literature also generate pointwise intervals (cf. (Hilgers, 1991; Metz et al., 1998; Claeskens et al., 2003; Hall et al., 2004; Zou et al., 1997)) and are not considered here either. Connecting pointwise intervals to form confidence bands is a mistake: due in part to problems of multiple comparisons, these bands generally will be too narrow.

Medical researchers have examined the use of ROC curves extensively and have introduced techniques for creating confidence boundaries (pointwise intervals or bands). We consider three methods for generating ("simultaneous" or "joint" (Ma & Hall, 1993)) confidence bands on ROC curves. Working-Hotelling bands (WHB) are based on the Working-Hotelling hyperbolic confidence bands for simple regression lines (Working & Hotelling, 1929). *Simultaneous joint confidence regions* (SJR) use the distribution theory of Kolmogorov (Conover, 1980) to generate separate confidence intervals for TP and FP rates (Campbell, 1994), and use these to form bands. Finally, *fixed-width* simultaneous confidence bands (FWB) are non-parametric confidence bands created by displacing the entire ROC curve "northwest" and "southeast" a fixed amount (Campbell, 1994). FWBs require a set of ROC curves, which can be generated by evaluating the model on multiple testing sets or by resampling one test set. We resample with the bootstrap (Efron & Tibshirani, 1993), which also has been used in machine learning as a robust way to evaluate expected performance, for example for evaluating cost-sensitive classifiers (Margineantu & Dietterich, 2000).

2.1. Simultaneous Joint Confidence Regions (SJR)

The simultaneous joint confidence region (SJR) uses the Kolmogorov-Smirnov (KS) (Conover, 1980) test statistic to identify confidence intervals for TP and FP independently (Campbell, 1994). The KS statistic tests whether two samples come from the same underlying distribution by considering the maximal vertical distance in their respective estimated cumulative density functions. In our case, there are two relevant distributions we would like to test: for FP and TP. Thus, we can build a separate KS-based confidence band for FP, which would translate to a maximum horizontal distance allowed from the ROC curve, and a separate one for TP, which would translate to a maximum vertical distance allowed. The KS test identifies these two distances based on the number of instances in each sample—*i.e.*, the number of positives, m, and the number of negatives, n. To generate these distances, we look up d and e, the critical distances for a fixed TP and FP respectively, at confidence level $(1-\delta)$—Table 1 shows how these are calculated for sufficiently large set sizes (> 35).

	δ				
Set Size	0.20	0.15	0.10	0.05	0.01
> 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Table 1. Kolmogorov-Smirnov (KS) critical values for rejecting H_0 for set sizes > 35.

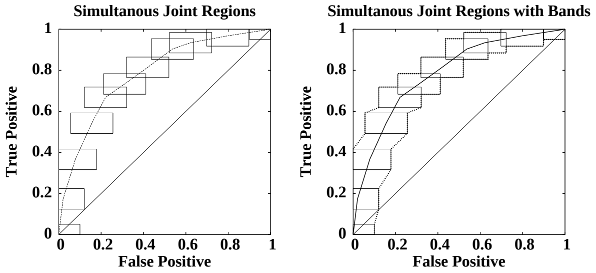

Figure 1. Transforming SJR into confidence bands.

The way confidence bands are generated using these regions is by generating a confidence region for each distinct point on the ROC curve constructed from the scored samples in $\mathcal{D}$. We trace the upper (lower) points of the confidence region to define the upper (lower) confidence band, cropped to stay within ROC space. Figure 1 illustrates this transformation.

Campbell (1994) argues that this procedure should give a $(1-\delta)^2$ confidence band for the true ROC curve. As we will see below, this is not the case, and in fact the procedure typically gives an implied confidence that is even bigger than $1-\delta$. To understand this, we should clarify that the horizontal and vertical bands we are building are using the model scores as the independent variable characterizing the distribution. Assume we build a separate box (fp ± d, tp ± e) around each point in our ROC curve, characterized by a threshold on the continuous scores. Then the two independent KS bands imply that with probability $(1-\delta)^2$ every threshold on the population score distribution would give a point in ROC space which falls within the box characterized by this score value. This rather com-

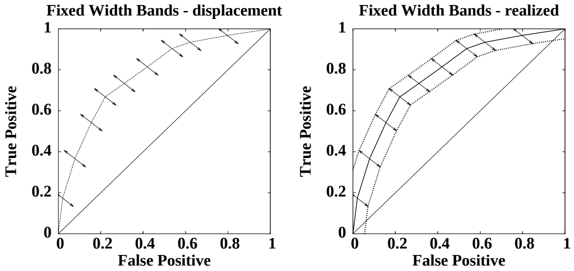

Figure 2. Displacing curve to generate FWB confidence bands.

plicated "score containment" characterization is not particularly useful for our case, since we are operating in (FP, TP) space, and ignoring the scores. The interesting thing about it is that "score containment" guarantees the "curve containment" we are interested in, but not the other way around. Hence we would expect the real confidence level to be higher than $(1-\delta)^2$.

2.2. Fixed-Width Bands (FWB)

To generate *fixed-width bands* (FWB) we start by identifying a slope, $b < 0$, along which to displace the original ROC curve (Campbell, 1994). The upper (lower) limit of the confidence band comprises each of the points of the observed ROC curve displaced "northwest" ("southeast") of its original location along an intersecting line of this slope. The resultant confidence band has a fixed width (along slope b) across the entire curve. Figure 2 illustrates this transformation.

Following Campbell (1994), we set $b = -\sqrt{(m/n)}$ (Campbell discusses how this is an approximation to the ideal, which would be to use the ratio of the standard deviations of TP and FP), and we use the bootstrap to identify the distance to displace the curve to generate the confidence bands. Given sample D, we generate bootstrap sample $D*$ (sample from D with replacement a set of the same size as D) and calculate the *maximum* distance along slope b from the ROC curve generated by D to the ROC curve generated by $D*$. We need the maximum distance because this is the width needed in order for $D*$ to be completely within the band. We sample 1000 $D*$'s, and find the distance needed in order to keep $1-\delta$ of all the curves completely within the generated bands. In our experiments below we observe that the FWBs attain containments of curves that for small sample sizes are smaller than the desired confidence level. This probably exposes one of the weaknesses of the bootstrap resampling methodology, when the sample from which we are resampling is not large enough to contain the full range of diversity of the population.

2.3. Simultaneous Working-Hotelling Bands (WHB)

Following Ma and Hall (1993) and Metz et al. (1998), we adapt a method for using Working-Hotelling hyperbolic bands (Working & Hotelling, 1929) to generate simultaneous confidence bands on an ROC curve. We use a publicly available implementation of the LABROC4 algorithm (Metz et al., 1998), which generates a "smooth" maximum likelihood (ML) estimation of an empirical ROC curve as well as pointwise confidence bounds.[3] The method is too complex to describe in detail here; we will give an intuitive overview and the interested reader is referred 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 (Ma & Hall, 1993) describe the construction of different sorts of confidence bands 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 bands are composed of points defined by the function l:

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

where k is a constant defined below, positive for the upper band and negative for the lower band, 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 band being generated. To generate confidence bands, we use Ma and Hall's simultaneous unrestricted Working-Hotelling bands, where, k_δ is determined using a chi-square distribution with 2 degrees of freedom:

$$k_\delta = \sqrt{-2\ln(\delta)} \qquad (2)$$

[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.

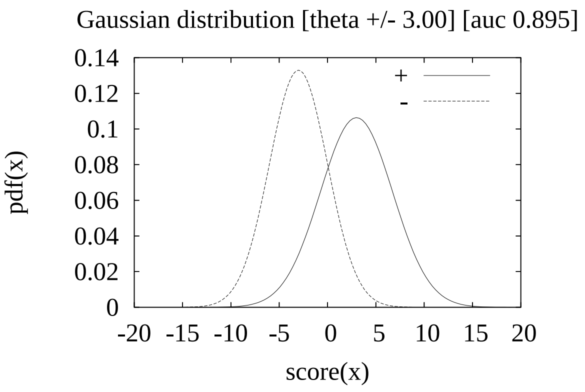

Figure 3. Example distribution used in study below.

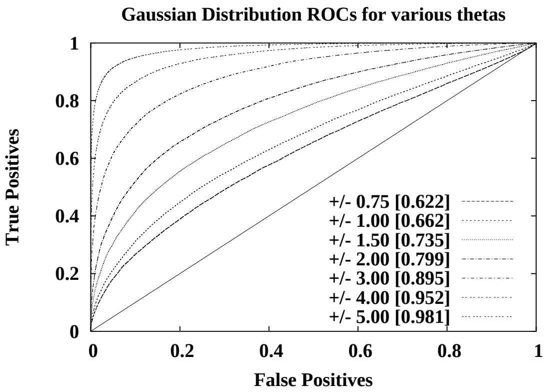

Figure 4. ROC curves generated for distribution as we vary θ.

3. Data Generation

To evaluate the different confidence bands, 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 $\theta^- = -\theta^+$. Figure 3 shows the distributions with $\theta = \pm 3.0$. Figure 4 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-curve" Evaluation

We expect the "true" ROC curve to fall completely within these bands with the specified probability (frequency)—in other words, were we to generate bands repeatedly from randomly drawn samples from $\mathcal{W}$, $1-\delta$ of the bands would contain the "true" ROC curve, where the "true" ROC curve is the curve generated directly from the cdf's (as above).

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 $(1-\delta)$ confidence bands, C_b, based on R as outlined in Section 2.

Table 2. Generating ROC Bands from Synthetic World.

We generate the bands using the simple methodology outlined in Table 2, with three parameters: (1) the synthetic world, which is defined by G^+, G^-, and $P(+)$, (2) the ROC-generation size, r, and (3) the confidence level δ.

4.1. Evaluation

To evaluate the bands, for each experiment we generate 1000 bands based on the method shown in Table 2, and count how many of them contain the true ROC curve. 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 parameters of the synthetic world. Ideally, $1-\delta$ of the calculated bands would contain the "truth".

4.2. Results

Figure 5 shows the containment for the 3 band methods for a subset of the values of r (horizontal axis) and θ (different curves).[4] We see very clear trends and interactions between these two parameters for each method. SJR is universally too wide, except for small values of r and θ. WHB seems to fail completely. This is due to the performance at small values of FP. By construction, the MLE curve fitting starts at $(0, 0)$ regardless of the empirical curve. This leads the WHBs to fail for this region of the ROC curve. If we modify the evaluation to start measuring containment at $FP > 0$, then the containment of WHB increases,[5] but it never performs as well as FWB and it always performs considerably worse at higher values of θ.

The FWBs clearly exhibit the best containment, close to $\delta = 0.1$ in all cases, with two exceptions: very small r, and the combination of low θ (low AUC) and large r (where it's still the best method of the three). Therefore, FWBs seem to be the method of choice, with caution taken for very small samples or extreme AUCs. For the rest of the paper, we will examine only FWBs.

5. "Future-curve" Evaluation

As described above, for machine-learning evaluations we don't know the true ROC curve, but often have sufficient data to answer a slightly different question. If the model

[4] We chose this subset for readability and to highlight the trends.

[5] We evaluate at starting values of $FP = \{0.01, 0.05, 0.10\}$

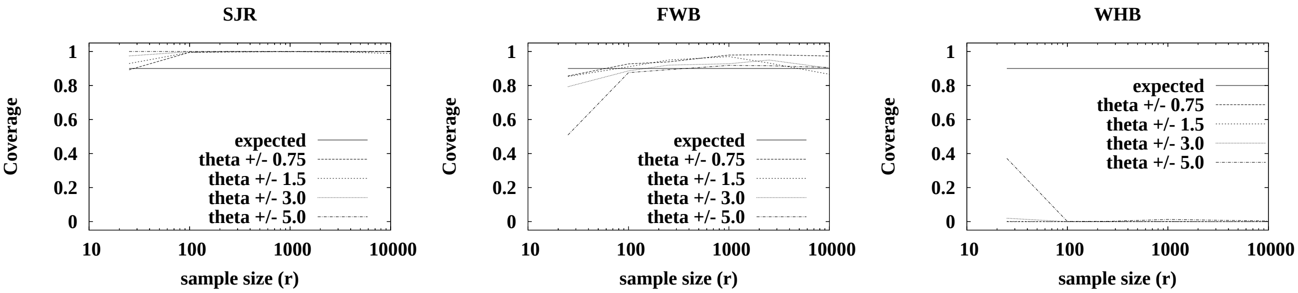

Figure 5. Containment of “true-curve” bands at $\delta = 0.1$. We show the containments for various values of r. As we can see, only FWB generates bands that are close to the expected containment.

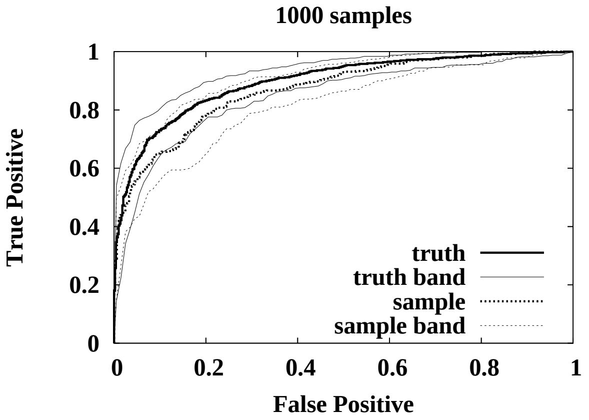

Figure 6. Variance problem with initial sample R. Variance about curves is correct, but the observed curve is off the true curve and the estimated bands are therefore off the proper region.

were to be used subsequently on the domain in question, is the resultant ROC curve likely to fall within the band?

To evaluate “future-curve” confidence bands, for each band we generate 1000 additional ROC curves, each based on r samples from $\mathcal{W}$, and count how many were completely contained by the band (where r is the same size as that used to generate the bands). Ideally, $1-\delta$ of the generated curves would fall within the bands.

Not surprisingly, all the methods fail. Each places a band about the observed curve. However, even if the methods are estimating the true variance correctly, future curves will be distributed about the true curve, not about the observed curve. Figure 6 illustrates the problem. The variances about the true and observed (sample) curves are very similar. However, because the sample is so far off from the true curve, the bands about it clearly are inappropriate for bounding the position of future curves.

5.1. Widening the band

One approach to addressing this problem is to widen the bands. Let us consider the true ROC curve (R$_T$), the sample ROC curve (R$_M$) from which we will calculate the bands (B$_M$) of width w, and an ROC curve sampled subsequently (R$_{M'}$), which with probability $(1–\delta)$ should lie within B$_M$.

Number of	Absolute values of θ						
Samples	0.75	1.0	1.5	2.0	3.0	4.0	5.0
25	0.86	0.91	0.96	0.83	0.76	0.69	0.86
100	0.94	0.88	0.96	0.95	0.95	0.88	0.86
250	0.93	0.93	0.96	0.89	0.93	0.92	0.93
1000	0.98	0.97	0.97	0.95	0.91	0.94	0.92
2500	0.97	0.95	0.95	0.92	0.96	0.94	0.92
10000	0.99	0.92	0.95	0.96	0.95	0.93	0.89

Table 3. Containments of FWB using calculated widths at $(1-\delta)$ and widen them by $\sqrt{2}$. As expected, these bands generally are slightly too wide (except at large values of θ or at $r \leq 100$).

Assume that we have a correct true-curve fixed-width band around R$_M$, calculated using the bootstrap approach or in any other way, and denote the chosen width parameter by w. This implies that the maximum distance between R$_T$ and R$_M$ in the chosen direction (slope $-\sqrt{m/n}$, see Section 2.2) 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 d$(\cdot)$ 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 band 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, change in the class proportions (m, n) dictating the direction in which w is chosen, etc.

Table 3 shows the containments we get from applying this technique, using $\delta = 0.1$. As suggested above, we see that in general these bands are slightly too wide. Nevertheless, we have not yet found an approach that performs better, either in terms of accuracy of containment or consistency.

5.2. Evaluation on Real Data

Now we are equipped to assess the containment of ROC confidence bands on real data, for which we do not know

Data set	**Size**	**Prior**
Adult	48842	0.761
Bacteria	40262	0.693
CalHous	20640	0.516
Coding	20000	0.500
Covertype	495141	0.572
Intcensor	18821	0.589
Intrusion	311025	0.805
Letter-A	20000	0.961
Letter-V	20000	0.806
Mailing	191779	0.949

Table 4. Data sets used in the real world setting.

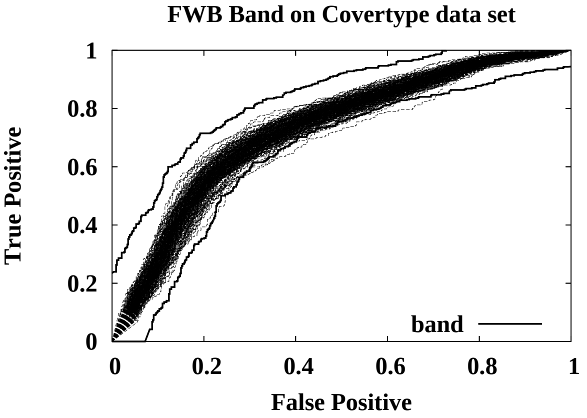

Figure 7. Example bands on the Covertype dataset. The model is a logistic regression model learned from 100 random samples, the bands were generated using $r = 1000$.

the true ROC curve. Note that we need relatively large data sets, in order to be able to do adequate bootstrap sampling for creating the curves *and* for the evaluation. We consider 10 relatively large data sets, used in prior machine learning work (Perlich et al., 2003) and listed in Table 4. See the original study for details on the data sets and the setup for binary classification.

We first draw a stratified random sample of 100 instances—the learning set—and build various learned models using Weka[6] (Witten & Frank, 2000)—logistic model trees (LMT) (Landwehr et al., 2003), J48, naive Bayes trees (NBT) (Kohavi, 1996), logistic regression (LR), and Naive Bayes (NB). We then generate prediction scores for the remaining instances. The log-odds scores, $\log \frac{P(+|x)}{P(-|x)}$, are used as the base population R from which to draw predictions. Using the same values of r as above, we sample r prediction scores from R to generate the confidence bands and sample 1000 scoring-sets of size r from the remaining prediction scores to evaluate the bands as "future" bands. We do this 10 times per R per data set to get containments for one learned model. We generated 10 models per learning algorithm by sampling 10 different learning sets.

Number of	Learning Method					
Samples	LMT	NBTREE	LR	J48	NB	average
25	0.81	0.79	0.84	0.88	0.87	0.84
100	0.87	0.85	0.89	0.84	0.91	0.87
250	0.88	0.87	0.89	0.82	0.92	0.88
1000	0.89	0.87	0.92	0.77	0.95	0.88
2500	0.89	0.88	0.94	0.69	0.94	0.87
10000	0.79	0.80	0.77	0.45	0.92	0.75
average	0.86	0.84	0.87	0.74	0.92	0.85

Table 5. Containments of FWB with a "widened" band to generate "future" bands based on prediction scores from the 5 machine learning methods. The models were learned from 100 randomly drawn samples. The scores reported are averages over 9 of the data sets.

Figure 7 shows one example band fitted to a logistic regression model, with $r = 1000$. The figure shows the confidence band and 250 of the 1000 verification ROCs. The figure clearly shows the variance problem—the observed ROC curve from which we generate the bands was obviously higher than the "true" curve, as the upper band is much higher than all the later drawn curves. We also see that FWB is much too wide at the extremes (due to its fixed width) and that when future curves fall outside the bands, they generally will do so in the middle.

Although there is considerable variance in the individual containment results on the real data, they generally are favorable with the exception of the Letter-A data set.[7] Table 5 shows the average containments (after removing Letter-A, which has a small but noticeable effect) for each of the five methods across various values of r. The overall average containment of the bands is 0.85, somewhat lower value than than desired. We increased the learning-set size to 2500 randomly drawn instances and repeated the evaluation outlined above. Table 6 shows the average containments (again after removing Letter-A) for each of the five learning methods across various values of r. The overall average containment of the bands increased to 0.87. The average containments clearly are dragged down by J48, which in many of the experiments—especially with 100 training examples—yielded poor containment. For these experiments we used the raw, class frequencies at the leaves of the trees (doing no smoothing), which are known to produce relatively poor ROC curves. Clearly by Table 6, with 2500 training examples the containments are substantially better.

[6] We use version 3.4.2. Weka is available at `http://www.cs.waikato.ac.nz/˜ml/weka/`

[7] For Letter-A for many cases (different learning techniques, different r values) the containment for Letter-A simply is 0. This deserves more investigation; we tentatively attribute the poor performance to the relatively small number of examples of the minority class. Letter-A has the most unbalanced class distribution and also is one of the smaller data sets. As Stein (2002) has shown, with a large class imbalance, the variance in ROC curves is extremely sensitive to the size of the minority class. Therefore, caution should be taken in extrapolating our results to data sets with relatively few examples of one class.

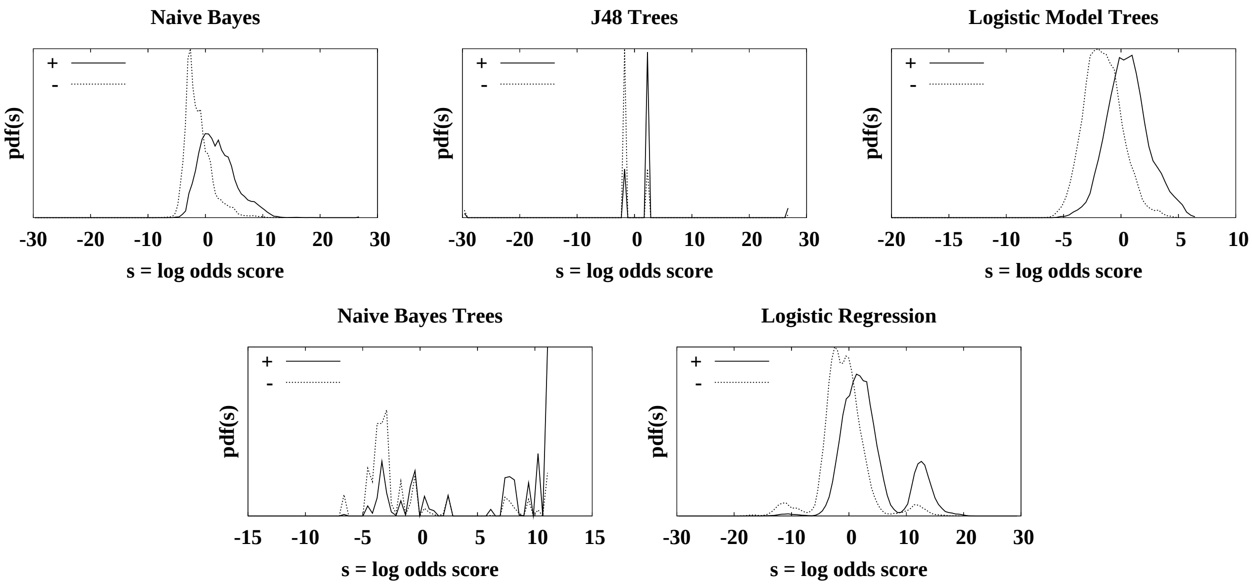

Figure 8. Score distributions of 5 Machine Learning methods on the Covertype data set.

Number of	Learning Method					
Samples	LMT	NBTREE	LR	J48	NB	average
25	0.71	0.78	0.74	0.78	0.83	0.77
100	0.83	0.87	0.86	0.85	0.88	0.86
250	0.88	0.91	0.91	0.89	0.91	0.90
1000	0.90	0.91	0.93	0.92	0.90	0.92
2500	0.92	0.94	0.95	0.95	0.94	0.94
10000	0.83	0.94	0.92	0.63	0.93	0.85
average	0.85	0.89	0.89	0.84	0.90	0.87

Table 6. Containments of FWB with a "widened" band to generate "future" bands based on prediction scores from the 5 machine learning methods. The models were learned from 2500 randomly drawn samples. The scores reported are averages over 9 of the data sets.

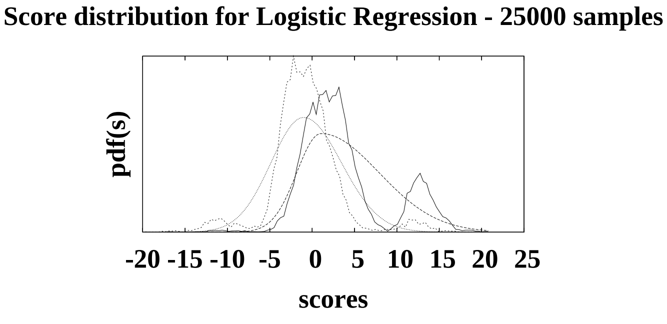

Figure 9. Sample logistic regression score distribution with $r = 250000$. These distributions are clearly not unimodal.

6. Discussion and Limitations

These results recommend the fixed-width bands, both for the true- and future-curve settings. The results may seem to justify applying the unwidened FWBs as true-curve confidence bands for the real-data setting—after all, the widened bands seem to give appropriate containment of future curves. This is an intuitively appealing heuristic, but one should keep in mind the assumptions on which the widening is based (see above), which may not hold for any particular data set and model.

A limitation of the fixed-width bands is their fixed width, which at least based on the cases we have looked at is too wide at toward the ends of the ROC curve. As we saw, the WHBs were too narrow at the ends of the ROC curve. Given that for many applications, the ends of the ROC curve are of particular interest, this leaves room for the design of better bands.

Non-parametric bands such as the FWBs have a special appeal for machine learning studies. Of course it may be that other bands, in particular bands based on binormal or other parametric models, also could be adjusted to perform well in the machine learning setting. However, machine-learned models generally do not produce binormal score distributions. A study by Bennett (2003) shows that standard ML methods do not induce models that generate Gaussian class-conditional score distributions; he shows distributions that have a closer fit to asymmetric Laplace distributions or asymmetric Gaussian distributions. Even for the same data set, different machine learning methods produce models with widely differing score distributions. Figure 8 shows the positive and negative score distributions generated by various types of learned model for the Covertype dataset. LMT has beautiful bell-shaped distributions, which itself may be be worth further investigation. Although LR and NB have fairly smooth distributions, they are clearly not binormal. The distributions of J48 and NBT are not even close to being bell-shaped. The naive Bayes distributions are more-or-less in line with observations made by Bennett (2003) (who studied naive Bayes for text classification): they are asymmetric, bell-shaped distributions where the positive distribution is fatter than the negative.

Finally, we reemphasize that this paper treated the problem of placing a confidence band around the ROC curve of a particular model for a particular domain (and testing-set size, for the future-curve bands). We have not addressed here the effect of using different values of r for creating the confidence bands and for testing them. We have shown previously that the variance of an ROC curve is directly related to r (Macskassy & Provost, 2004), which makes it crucial to ensure that these are equal. More importantly, we have not addressed at all the problem of assessing the confidence in the expected ROC performance of a learning algorithm for a particular domain, which also must account for the variance due to the choice of training data.

In conclusion, to produce confidence bands about ROC curves, our results recommend the non-parametric, fixed-width bands described by Campbell (1994), adjusted if necessary to produce future-curve bands. A promising avenue is to extend the bootstrap procedure to generate fixed-width confidence band for future curves, rather than use the heuristic $sqrt(2)$ correction. We hope eventually to offer a full bootstrap-based FWB solution both for true-curve and future-curve confidence bands.

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

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.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *7*, 1145–1159.

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.

Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley. 2nd edition.

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., & 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.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press/MIT Press.

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic Model Trees. *Proceedings of the 16th European Conference on Machine Learning*.

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 Analysis in AI (ROCAI-2004) at ECAI-2004*.

Macskassy, S., Provost, F., & Rosset, S. (2005). Pointwise ROC Confidence Bounds: An Empirical Evaluation. *Proceedings of the Workshop on ROC Analysis in Machine Learning (ROCML-2005) at ICML-2005*.

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

Perlich, C., Provost, F., & Simonoff, J. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, *4*, 211–255.

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.

Witten, I. H., & Frank, E. (2000). In *Data Mining: Practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann.

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.

Zou, K. H., Hall, W. J., & Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, *16*, 2143–2156.