# Confidence Bands for ROC Curves: Methods and an Empirical Study

**Sofus A. Macskassy** and **Foster Provost**[1]

**Abstract.** In this paper we study techniques for generating and evaluating confidence bands on ROC curves. ROC curve evaluation is rapidly becoming a commonly used evaluation metric in machine learning, although evaluating ROC curves has thus far been limited to studying the area under the curve (AUC) or generation of one-dimensional confidence intervals by freezing one variable—the false-positive rate, or threshold on the classification scoring function. Researchers in the medical field have long been using ROC curves and have many well-studied methods for analyzing such curves, including generating confidence intervals as well as simultaneous confidence bands. In this paper we introduce these techniques to the machine learning community and show their empirical fitness on the Covertype data set—a standard machine learning benchmark from the UCI repository. We show how some of these methods work remarkably well, others are too loose, and that existing machine learning methods for generation of 1-dimensional confidence intervals do not translate well to generation of simultanous bands—their bands are too tight.

## 1 Motivation

Receiver-Operator Characteristic (ROC) analysis is an evaluation technique used in signal detection theory, which in recent years has seen an increasing use for diagnostic, machine-learning, and information-retrieval systems [26, 22, 19, 23, 14]. ROC graphs plot false-positive (FP) rates on the x-axis and true-positive (TP) rates on the y-axis. ROC curves are generated in a similar fashion to precision/recall curves, by varying a threshold across the output range of a scoring model, and observing the corresponding classification performances. Although ROC curves are isomorphic to precision/recall curves, they have the added benefits that they are insensitive to changes in marginal class distribution. Often the comparison of two or more ROC curves consists of either looking at the Area Under the Curve (AUC) or focusing on a particular part of the curves and identifying which curve dominates the other in order to select the best-performing algorithm.

Much less attention has been given to robust statistical comparisons of ROC curves. This paper addresses the creation and evaluation of confidence bands on ROC curves. We ask whether, assuming test examples are drawn from the same, fixed distribution, one should expect that the model's ROC curves will fall completely within the bands with probability $1-\delta$. Prior work in machine learning has considered sweeping across thresholds on the classification scoring function, creating confidence intervals around the TP/FP points for various thresholds [9], or sweeping across the FP rates and creating vertical confidence intervals around averaged TP levels [24]. Much more work has been done in the medical field, but so far has not penetrated into the machine learning community. Prior work in this field includes generating confidence intervals around TP/FP points based on the exact binomial distribution [10], and generation of confidence intervals based on the binormal distribution[2] [17]. Confidence bands could be created by connecting any of these confidence intervals (as we will show). More relevant work to that of creating confidence bands is the creation of simultaneous joint confidence regions based on the Kolmogorov-Smirnov test statistic [5], simultanenous confidence bands based upon the Working-Hotelling hyperbolic confidence bands around simple regression lines [13], and use of the bootstrap to generate empirical fixed-width confidence bands [5]. None of these prior studies on simultaneous confidence bounds, however, have asked whether the created bands actually hold empirically.

In this paper we examine these methods for creating such confidence bands for a given learned model. As we will show, the bands created by many of these techniques are too tight. To these ends, we describe a framework for evaluating the fit of ROC confidence bands. Specifically, we examine $1-\delta$ confidence bands on a model's ROC curve.

The main contributions of this paper are the introduction of relevant techniques from the medical field to the machine learning community and an empirical study of these techniques as well as the techniques already used by machine learning researchers.

The rest of the paper is organized as follows. The next section discusses related work on creating confidence intervals for ROC curves, followed by a section describing the methods we use in this paper for generating ROC confidence bands. We then describe our evaluation methodology and a case study showing that two of these methods—fixed-width bands and simultaneous joint confidence regions—perform close to expectation in most cases, whereas the rest do not.

[1] NYU Stern School of Business, 44 W. 4th Street, New York, NY 10012, email: {smacskas,fprovost}@stern.nyu.edu

[2] Binormal distributions, or bivariate normal distributions, are joint distributions over $x$ and $y$, two independent variables which are normally distributed.

## 2 Overview of Existing Relevant Techniques

Within the machine learning field, prior work on creating confidence intervals for ROC curves has for the most part been in the context of creating one-dimensional confidence intervals.

*Pooling* is a technique in which the $i$-th points from all the ROC curves in the sample are averaged [4]. This makes a strong assumption that the $i$-th points from all these curves are actually estimating the same point in ROC space, which is at best a doubtful assumption.

*Vertical averaging* looks at successive FP rates and averages the TPs of multiple ROC curves at that FP rate [24]. By freezing the FP rate, it is possible to generate a (parametric) confidence interval for the TP rate based on the mean and variance; multiple curves are generated using cross-validation or other sampling techniques. A potential weakness of this method is the practical lack of independent control over a model's false-positive rates [9].

*Threshold averaging* seeks to overcome the potential weakness of the vertical averaging by freezing the thresholds of the scoring model rather than the FP rate [9]. It chooses a uniformly distributed subset of thresholds among the sorted set of all thresholds seen across the set of ROC curves in the sample. For each of these thresholds, it identifies the set of ROC points that would be generated using that threshold on each of the ROC curves. From these ROC points, the mean and standard deviations are generated for the FP and TP rates, giving the mean ROC point as well as vertical and horizontal confidence intervals.

Use of the bootstrap [8] as a more robust way to evaluate expected performance has previously been used for evaluating cost-sensitive classifiers [15]. In this work, bootstrapping was used to repeatedly draw predictions $p(i, j)$, where $p(i, j)$ is the probability that an instance of class $j$ was predicted to be in class $i$. Using these sample predictions, it was possible to generate a final cost based on a cost-matrix. They did this repeatedly to generate a set of estimated costs, which they then used to generate confidence bounds on expected cost.

Medical researchers also have examined the use of ROC curves extensively and have introduced many techniques for creating confidence boundaries. The problem domains and tasks in medical research are generally different from those of machine learning in that they often consider only small data sets, where one instance is the test result from a patient. Further, it is often assumed that these data are ordinal in nature—e.g., that it is 'ratings' data with a small scale such as 'definitely diseased','probably diseased','possibly diseased',' 'possibly non-diseased','probably non-diseased','definitely non-diseased' [1, 26, 30].

One technique, similar to that of threshold averages, creates a confidence boundary around each of the $N$ ROC points associated with $N$ discrete events based on an underlying model [27]. It does this by considering each axis as independent and considering an $N$-dimensional vector along each axis, where the $i$-th element in the vectors represent the $i$-th point on the ROC curve. Discretizing the values and assuming a binomial distribution, it then generates a probability distribution of the likelihood that the $j$-th value lies in each discretized cell. It maps this probability density back into ROC space thereby generating confidence boundaries for each point in the ROC curve. These models are very complex and are not tractable for even small $N$ larger than about 10, and would currently be intractable for large sets of ROC points as is typically found in machine learning studies.

Other work has created a joint confidence region (or "local confidence rectangle") for a given fixed threshold $t$ under the assumption of a binomial distribution [10]. This region is constructed by generating separate $(1-\delta)$ confidence intervals for TP and FP rates independently at the given threshold. The resulting region should then contain the (FP,TP) point at threshold $t$ with confidence $(1-\delta)^2$. This is equivalent to the threshold averaging method described above, using the binomial distribution rather than the normal distribution.

*Simultaneous joint confidence regions* uses the distribution theory of Kolmogorov [6] to generate separate confidence intervals for TP and FP rates [5]. This is done by finding the Kolmogorov $(1-\delta)$ confidence band for TP (tp$\pm d$) and FP (fp$\pm e$). By an independence assumption, the rectangle with width $2e$ and height $2d$, centered at a given point, should contain points at the given threshold with confidence $(1-\delta)^2$. Unlike Hilgers's approach above, all rectangles using this method will be of the same size. We describe our use of this method in Section 3.6.

Creating a confidence region in ROC space restricts both FP and TP rates to the region $(0, 1)$. This restriction can cause difficulties when using intervals based on normal distributions. One solution is to transform the points to logit space[3], generate the confidence intervals in that space, and then convert them back into ROC space [29]. An alternative transformation also used is that of converting to and from probit space[4] as done in the ROCKIT/LABROC4 algorithms [17, 16]. Both of these bodies of work assume an underlying binormal distribution and focus on creating either one-dimensional confidence intervals, or joint confidence regions. We use our own implementation of ROCKIT to generate confidence bounds under the binormal distribution, as described in Section 3.8.

One method for generating simultaneous confidence bands on ROC curves [13] makes use of Working-Hotelling hyperbolic confidence bands for simple regression lines [28]. Under the binormal model, an ROC curve can be parameterized as $\text{TP} = \Phi(a - b\Phi^{-1}(\text{FP}))$, where $\Phi(z)$ is the standard-normal cumulative distribution function [7]. Using this parametrization, the Working-Hotelling bands can then be applied to ROC curves to generate simultaneous confidence bands. We describe our use of this method in Section 3.8.

The *fixed-width* simultaneous confidence bands method is a non-parametric method, which generates simultaneous confidence bands by displacing the entire ROC curve "northwest" and "southeast" along lines with slope $b = -\sqrt{(m/n)}$, where $m$ is the number of true positives and $n$ is the number of true negatives [5]. This slope is an approximation of the ratio of the standard deviations for TP and FP—a property which tries to take into account the curvature of the ROC plot rather than using a displacement along one of the two axes as is done by the majority of methods described above. They use the bootstrap to identify the distance the curve should be displaced,

[3] $logit(p) = \log(\frac{p}{(1-p)}); logit^{-1}(p') = \frac{1}{1+exp(-p')}$.

[4] $probit(p) = \Phi(p); probit^{-1}(p') = \Phi^{-1}(p')$, where $\Phi(z)$ is the cumulative normal distribution function.

thereby generating a fixed-width band across the complete curve. We describe how we use this method in Section 3.7.

## 3 Generating Confidence Bands

In this section we describe our methodology for generating confidence bands for a classification model or modeling algorithm. We adapt two existing methods from machine learning: vertical averaging (VA) and threshold averaging (TA) for generating confidence intervals, and three methods from the medical field: simultaneous joint confidence regions (SJR), Working-Hotelling based bands (WHB), and fixed-width confidence bands (FWB).

Three of the methods (VA, TA and FWB) work based on the assumption that we can generate (or are given) a set of ROC curves. These can be generated by running a learning algorithm on multiple training sets, testing on multiple testing sets, or resampling the same data. These ROC curves will be used to generate confidence bands about an average curve.

While all these methods generate different types of confidence bounds—VA and WHB generate 1-dimensional intervals, while TA and SJR generate intervals on TP and FP axes both, and lastly FWB generates a complete curve—we use the following general methodolgy that can be applied to each of them to generate confidence bands. This general methodology, which we use throughout this paper, consists of the following steps:

1. Create a distribution of ROC Curves, if the method needs it (VA, TA and FWB).
2. Generate points for the confidence bands.
   (a) Choose an underlying distribution, if applicable (VA and TA, see below).
   (b) Sweep across the ROC curves to calculate, on a point by point basis, where the respective confidence boundaries are. We use one of the five methods mentioned above in this step.
3. Create confidence bands by considering all upper (lower) interval points found in step 2(b) to make up the upper (lower) confidence band.

### 3.1 Creating the Distribution of ROC Curves

There exist various ways of generating a distribution of instances from which to generate a confidence interval. The most common methods, including cross-validation [11], repeatedly split a data set into training and test sets. Each such split gives rise to a learned model, which can be evaluated against the test set—thereby generating one ROC curve per split. The bootstrap [8] is a standard statistical technique that creates multiple samples by randomly drawing instances, with replacement, from a host sample (the host sample is a surrogate for the true population). Each such set of samples can then be used to generate an ROC curve. We can repeatedly draw $N$ samples to generate a distribution of ROC curves. To our knowledge there is only one previous body of work which has applied the bootstrap to generate multiple ROC curves to use for fitting confidence bands [5]. See Section 5.3 for details on how we use bootstrapping in our study.

### 3.2 Distribution Assumption

ROC methodologies have historically assumed a binormal distribution [30, 29]. However, it may be that other distributions are more appropriate or work equally well. For example, for a given x-value (FP rate) the y-value (TP rate) is a proportion. So a binomial distribution may be appropriate [1, 10]. We consider four distributions for creating confidence intervals: binormal, normal, binomial and empirical. Let us assume that we are given a sample distribution $\mathcal{D}$ of points along some dimension and a confidence threshold of $\delta$.

We generate confidence intervals and bands under the assumption of a *binormal distribution* when using Working-Hotelling confidence bands. See Section 3.8.

We generate confidence intervals under the assumption of a *normal distribution* by calculating the mean $\mu$ and standard deviation $\sigma$ of $\mathcal{D}$. We then look up the statistical constant, $z$, for a two-sided bound of $\delta$ confidence on a distribution size of $|\mathcal{D}|$ giving us a confidence interval of $\mu \pm z \cdot \sigma$.

For the *binomial distribution*, we calculate the variance as $V = \mu \cdot (1-\mu)$, thus giving confidence interval $\mu \pm z \cdot \sqrt{V/|\mathcal{D}|}$.

For an *empirical distribution*, we create empirical bounds as follows: we sort the values of $\mathcal{D}$ and choose $v_l$ and $v_u$, such that $v_l$ is smaller than $1-\frac{\delta}{2}$ of all values and $v_u$ is larger than $1-\frac{\delta}{2}$ of all values. Thus $1-\delta$ of all values lie between $v_l$ and $v_u$.

### 3.3 Sweep Methodology

So what are the dimensions along which the confidence intervals will be created? These are defined by how one "sweeps" across the ROC space to generate these intervals. A sweep samples the observed ROC point (or average ROC point for a set of curves) and the confidence boundary about it. These boundary points are then used to generate the upper/lower confidence bands. We describe how the sweep methodology is used in the following sections.

For the TA and VA sweeps, we use one of three distribution assumptions (normal, binomial, empirical). Some methods (VA, TA, and FWB) require a sampling of points. In this case, the sweep uses a distribution of ROC curves which is used by the respective methods. For the other methods (WHB and SJR), the sweep uses just one ROC curve.

All of our sweep methods require two parameters:

1. The confidence $\delta$, which we set to 0.05 for a 95% confidence bound throughout this paper. We did preliminary tests with other $\delta$'s (0.10 and 0.01) with similar results as those presented below.
2. The number and distribution of points to sample along the sweep, which we set to a uniformly distributed 100 points along the sweep orientation axis. This number can be changed depending on how fine-grained a curve is needed.[5]

### 3.4 Vertical Averaging (VA)

Sweeping the ROC curve using the *vertical averaging* (VA) method works as follows: sweep a vertical line from FP = 0 to FP = 1, sampling the distribution of TPs from the collection

[5] While this is a free variable that will have some effect on the overall fit of the bands, we do not investigate its effect in this paper.

of ROC curves at regular points along the sweep. For each such sampling at a fixed FP, TP confidence intervals can be created using any of the distribution assumptions mentioned above.

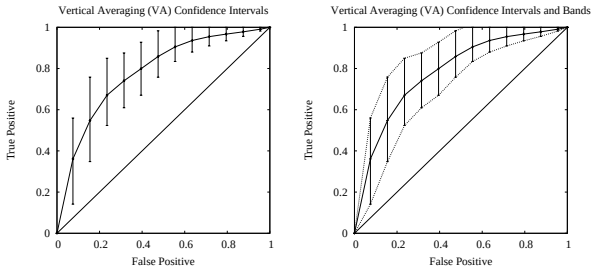


**Figure 1.** Transforming vertical averaging into confidence bands.

We generate confidence bands by considering all the upper (lower) interval points as the points making up the upper (lower) band. Figure 1 illustrates this methodology. For each FP (0.0 through 0.99 since FP=1.0 always has a TP of 1.00), we generate a distribution of possible TPs across all the sampled ROC curves and generate the bands based on this distribution.

## 3.5 Threshold Averaging (TA)

The sweep for the *threshold averaging* (TA) method works a little differently than that for the VA method. It sweeps along the thresholds on the model scores from the smallest to the largest observed threshold, sampling the distribution of ROC points generated with each threshold. It then generates the mean (FP,TP) point for each sampled threshold and finds the confidence intervals of the FPs and TPs, using any of the distribution assumptions mentioned above.

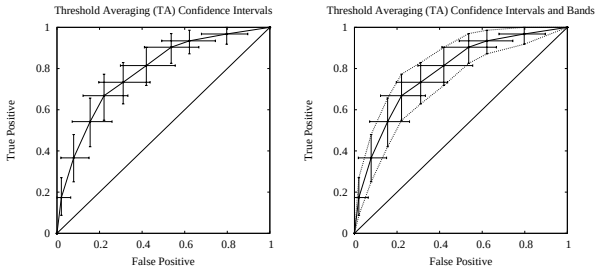


**Figure 2.** Transforming threshold averaging into confidence bands.

This method is less straightforward to adapt to our framework as there are various ways to deal with two confidence intervals. In this paper we chose the simplest approach: discount the confidence interval for FP and only use the confidence interval for TP. Because of this, the bands we generate turn out to be somewhat conservative and containment probably is underestimated. Figure 2 illustrates the transformation as well as the drawback. In the figure, we clearly see that some FP intervals reach outside the confidence bands (opposite to the vertical intervals, the horizontal intervals will tend to be larger for higher FP rates). While there are alternative methods for generating the bands, such as considering the bounding box, or the diamond made up by the interval boundaries, we do not consider them for this study.

## 3.6 Simultaneous Joint Confidence Regions (SJR)

The simultaneous joint confidence region (SJR) works differently than either of VA and TA. It uses the Kolmogorov-Smirnov (KS) [6] one-sample test statistic to identify a global confidence interval for TP and FP independently [5]. The KS statistic is used to test whether two sampled sets come from the same underlying normal distribution by considering the maximal vertical distance in their respective estimated cumulative density functions. For our purpose, that means the maximal vertical (horizontal) distance allowed from the given ROC curve to another ROC curve without rejecting $H_0$—*i.e.*, the confidence interval along FP (TP). Using the KS one-sample test allows us to identify these two distances, using the number of instances in each sample—*i.e.*, the number of true positives, $m$, and the number of true negatives, $n$. For sufficiently large set sizes ($> 35$), these distances are defined as follows.

| | $\delta$ | | | | |
|---|---|---|---|---|---|
| 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 $> 35$.

We look up $d$ and $e$, the critical distances along TP and FP respectively, at confidence level $(1-\delta)$. These identify the simultaneous joint confidence region for a given observed point $(\text{fp}, \text{tp})$ to be $(\text{fp} \pm d, \text{tp} \pm e)$ at confidence level $(1-\delta)^2$. Note that while the confidence level is theoretically $(1-\delta)^2$, we empirically test it as though it is at the $(1-\delta)$ level. Surprisingly, we show that it generally achieves this tighter bound—*i.e.*, $(1-\delta)^2$ would be too loose.

The way we generate the confidence bands using these regions is by sweeping along FP in a similar fashion as was done with VA. At regular intervals, we freeze FP and identify the respective TP. We use the upper left (lower right) corners of the confidence region to define the upper (lower) confidence band, cropped to stay within ROC space. Figure 3 illustrates this transformation.

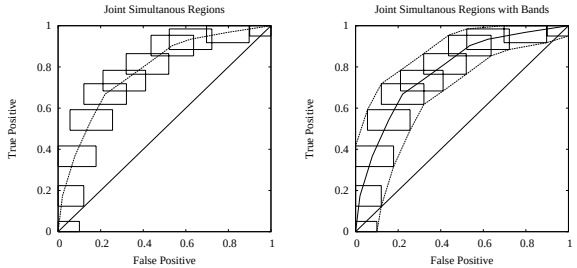


**Figure 3.** Transforming SJR into confidence bands.

### 3.7 Fixed-Width Bands (FWB)

The *fixed-width bands* (FWB) method works by identifying a slope, $b < 0$, along which to displace the original ROC curve to generate the confidence bands [5]. In other words, the upper (lower) confidence band would consist of all the points of the original observed ROC curve displaced "northwest" ("southeast") of their original location. This creates a confidence band with a fixed width across the entire curve. The question is what slope to use and what distance to displace the curve. While the ideal slope would be the ratio of the standard deviations associated, respectively, with TP and FP, we here adopt the same approximation as that used in the original work and use the slope $b = -\sqrt{(m/n)}$. The way

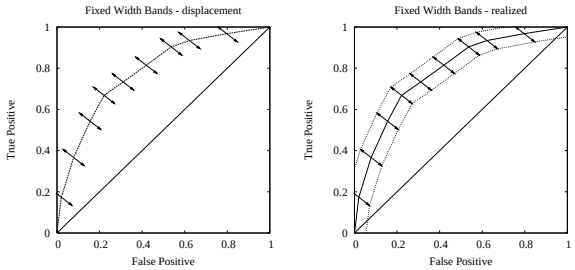


**Figure 4.** Displacing curve to generate FWB confidence bands.

we generate the confidence bands using this method is similar to that of SJR. We sweep along the FP axis, freezing FP at regular intervals and identify the TP value at that FP. We add the upper (lower) boundary points by moving a distance $d$ in each direction along the line with slope $b = -\sqrt{(m/n)}$. Figure 4 illustrates this transformation.

As with our study, the original work used the bootstrap to identify the distance to displace the curve to generate the confidence bands. Using this distribution of ROC curves, we identify the distance needed to have $(1-\delta)$ of all the curves be completely contained within the confidence bands. This is notably different from the approaches taken by TA and VA which generated intervals on a per-point basis.

### 3.8 Working-Hotelling Bands (WHB)

We adapt a method for using Working-Hotelling hyperbolic bands [28] to generate simultaneous confidence bands on an ROC curve [13]. The confidence bands are fitted to a regression line, $y = a - b \cdot x$, and are of the form:

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

where $k \geq 0$ is a constant which we define below, and $\sigma(x) = \sqrt{\sigma_a^2 - 2\rho\sigma_a\sigma_b \cdot x + \sigma_b^2 \cdot x^2}$, as defined by the covariance matrix $\Sigma$:

$$\Sigma = \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \tag{2}$$

We use maximum-likelihood estimation (MLE) to generate a regression line to fit the ROC curve. We use our own implementation of the ROCKIT algorithm [16] to do so.[6] The ROCKIT algorithm works by first grouping continuous data into 'bins' or 'runs' of instances either with the same model score and/or same label. Then it uses an ordinal ('rating method') algorithm [7] to create a smooth binormal ROC curve. The covariance matrix is also calculated as part of the algorithm.

There are various constants, $k$, available at confidence level $(1-\delta)$, depending upon the type of band being generated. For the purpose of our study, we use two types of bands: two-sided *pointwise* confidence bands (WHB-p) and *simultaneous unrestricted* confidence bands (WHB-s). The pointwise confidence bands are analogous to the vertical averaging confidence intervals under the binormal distribution. As such, WHB-p will generate tighter bands than WHB-s, as we will show later. For WHB-p, the constant $k_\delta$, for confidence level $(1-\delta)$ is $2\Phi(x) - 1$.[7] For WHB-s, $k_\delta$ is determined by the chi-square distribution with 2 degrees of freedom: $k_\delta = \sqrt{-2ln(\delta)}$.

## 4 Evaluation

The key question we ask in this paper is how good are these bands? As with confidence intervals on a single variable, we would like to be able to say that given a $\delta$, the bands generated can be expected to fully contain the curve from a given model with a probability of $1-\delta$ (assuming that new test instances come from the same distribution). As we will show, for only one of the methods proposed above does this hold.

## 5 Case Study

### 5.1 Data

We now present a case study using the Covertype data set from the UCI repository [2]. We chose this data set because its large size enables in-depth testing across a wide range of model-generation and ROC-generation set sizes. The Covertype data set consists of 581,012 instances having 54 features, 10 being numerical and the rest being ordinal or binary. While it has seven classes, there is a large variation in class membership sizes. To study the ROC curves, we chose examples of the two classes with the most instances, giving us a data set of 495,141 instances (57.2% base error rate).

### 5.2 Scoring Model

We use a modified C4.5R8 [25] that generates probabilities of class membership [21]. If a leaf matches $p$ positive examples and $n$ negative examples, we apply a simple Laplace correction [20] giving us a probability estimate of $\frac{p+1}{p+n+2}$, as we have 2 classes. Further, we do no pruning of the tree, as standard pruning does not consider differences in scores that do not affect 0/1 loss (but may deflate the ROC curve) [21].

Any classifier–even a fixed function—would suffice for this step, as the final fixed model is used only as a score generator for our ROC analysis.

[6] This is part of our ROC analysis toolkit, which we plan on releasing to the public later this year. This toolkit is written in Java.

[7] This, it turns out, is equivalent to $k_\delta = z_{\delta/2}$, where $z_{\delta/2}$ is the statistical constant for a two-sided bound of $\delta$ confidence.

## 5.3 Bootstrap-based Evaluation

To generate and evaluate confidence bands, we use the following method based on a bootstrapped empirical sampling distribution.

1. Randomly split the complete data set into a model-generation ($\mathcal{MG}$) set of 256,000 instances and a ROC-generation ($\mathcal{RG}$) set of 125,000 instances, keeping these two sets disjoint.
2. Fix the model-generation size, $m$, and sample with replacement from $\mathcal{MG}$ a model-generation set, $M$, of size $m$.
3. Learn a classifier based on $M$.
4. Fix the ROC-generation size, $r$, and sample with replacement from $\mathcal{RG}$ a ROC-generation set, $R$, of size $r$.
5. Generate multiple "fitting" sets $F_i$ multiple verification sets, $V_i$:
   (a) Generate $r_{fit}$ "fitting sets", $F_i$ of size $r$ by repeated sampling with replacement from $R$. For each $F_i$, generate an ROC curve, $\text{roc}(F_i)$, for the model. The result is a set of ROC curves, $\text{roc}_F = \{\text{roc}(F_i)\}$.
   (b) Generate confidence bands, $C$, based on $\text{roc}_F$.
   (c) Generate $r_{eval}$ "verification" sets, $V_j$, of size $r$ by repeated sampling with replacement from $\mathcal{RG}$. For each such sample, generate a verification ROC curve, $\text{roc}(V_j)$. The result is a set of ROC curves, $\text{roc}_V = \{\text{roc}(V_j)\}$.
   (d) Calculate the percentage of ROC curves in $\text{roc}_V$ that fall completely within the generated confidence bands, $C$.
6. Repeat steps (4)–(5) 10 times to account for variability in the generated confidence bands.

This methodology has four parameters: the model-generation size, the ROC-generation size, the number of sampling runs, $r_{fit}$ , used to generate $\text{roc}_F$ in step 5(b) to generate the confidence curves, and the number of sampling runs, $r_{eval}$, used to generate $\text{roc}_V$. We fix this latter number of sampling runs to 1000. We examine the sensitivity to each of the remaining parameters in the next section. Note that for this paper, we do not consider variance in curves due to the model-generation set—only confidence bands on the ROC curve of a particular (learned) classifier. However, a similar methodology would apply to the generation of confidence bands for a learning algorithm.

## 5.4 Trends in Confidence Bands

In this section we examine the experimental parameters identified above. Unless stated otherwise, we will use the FWB method for the figures presented as this method is the best performer among the methods used in this study.

### *5.4.1 Model-Generation Set Size*

This parameter is the least interesting for this particular case study. As the model-generation set size increases, the ROC curves become higher as would be expected. However, while this has some effect on the width of the confidence bands, it is more a matter of considering different learned models than of how to generate good bands for a given model. As such, we do not consider this to be an important dimension for further discussion here and fix the size to 1000 instances.

### *5.4.2 ROC-Generation Set Size*

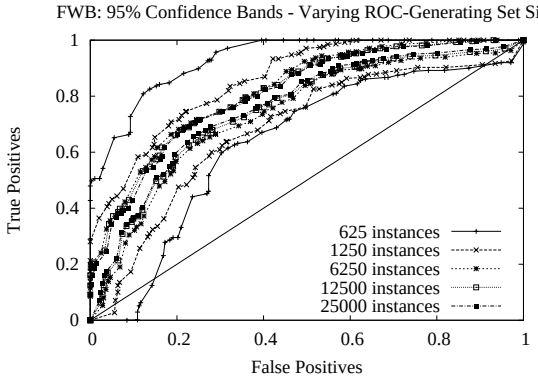


**Figure 5.** ROC bands using various ROC-generation sizes.

The ROC-generation set size should have an obvious effect on the bands generated. We varied the ROC-generation set size between $625, 1250, 6250, 12500$ and 25000 instances (0.5%, 1%, 5%, 10% and 20%, respectively, of $\mathcal{RG}$). As the set size increases, as expected the approximate confidence intervals generated by any of our methods become narrower and therefore so do our confidence bands. With too few samples, the estimate of the confidence interval tends to be inaccurate and biased to be too wide. Figure 5 illustrates this effect clearly.

### *5.4.3 Number of Fitting Curves*

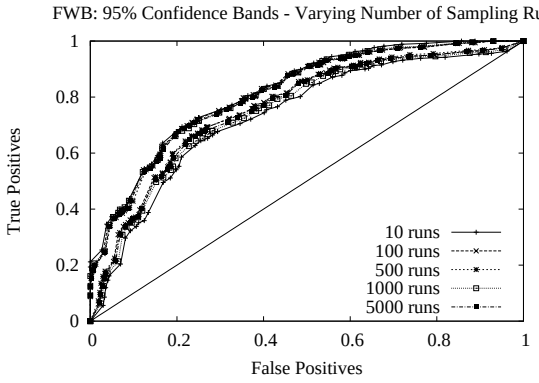


**Figure 6.** ROC Bands using varying number of sampling runs, using a ROC-generation size of 1000.

The number of samplings runs, $r_{\text{fit}}$, used to create the empirical distribution (step 5(a) in Section 5.3) is the last free parameter that we consider. In order to generate the ROC bands using VA, TA and FWB, we need to have a sample of ROC curves from which to generate these bands. The question to answer is how many such fitting ROC curves—the number of sampling runs—are needed to generate reasonable bands. While the effect of this variable is not as intuitive as the ROC-generation or model-generation set sizes, it still does have an effect as can be seen in Figure 6. While the upper band is

fairly stable we see that the lower band tightens with more sampling runs.

### 5.5 How Good Are The Bands?

Having considered our experimental parameters, let us now ask our main question: do the $1-\delta$ confidence bands actually contain $1-\delta$ of the empirical distribution? For an initial answer, we first fix the number of fitting curves to 1000 and the ROC-generation size to 12,500.

As per our bootstrap-based methodology, we randomly sampled 1000 ROC-generation sets of size 12,500 with replacement from $\mathcal{RG}$ and counted the percentage of curves in $\text{roc}_V$ that fell completely within the confidence band. We did this for each of our methods using each of the applicable distribution assumptions. Table 2 shows the coverage achieved by each of the methods tested.

| | distribution assumption | | | |
|---|---|---|---|---|
| **Method** | empirical | normal | binomial | binormal |
| FWB | 95.6(0.7) | — | — | — |
| SJR | — | 97.0(4.3) | — | — |
| WHB-p | — | — | — | 7.6(6.7) |
| WHB-s | — | — | — | 86.6(13.1) |
| VA | 28.5(10.1) | 34.5(1.0) | 43.5(1.1) | — |
| TA | 0.2(0.2) | 0.2(0.2) | 42.2(1.0) | — |

**Table 2.** How many verification ROC curves fall within the bands of each method using a given distribution for generating bands? Each cell shows the percentage and standard deviation of curves completely contained within the created confidence bands.

As we can see in the table, only two of the bands (FWB and SJR) achieve the 95% bound that we would expect, although WHB-s gets close with a coverage of 86, 6%. Not surprisingly, neither TA nor VA get anywhere near the bound.[8] It was suggested to us that the failure of these bounds are to be expected for TA, VA and WHB-p due to the multiple comparisons problem and that we should in fact be doing a Bonferroni correction [3, 18], which in a nutshell states that the probablity of a Type I error is $1-(1-\delta)^k$, where $k$ is the number of comparisons. The problem with this correction is that it is overly conservative $1-(1-\delta)^{100} = 0.994$ for $\delta = 0.05$—*i.e.*,, the probability of falsely rejecting $H_0$ is 99.4%! In fact, we'd need to set $\delta = 0.0005$ in order to generate a 95% confidence bound with 100 points on the ROC curve. While there are less conservative alternatives to the Bonferroni correction [12], they are still too conservative for the number of comparisons done in our study (the different points on ROC curves are very far from being independent).

Based on Table 2, it seems that VA, TA and WHB-p are not good methods for generating confidence bands, while any of the remaining three methods are plausible. In order to verify these findings, we tested all the methods in a wider range of ROC-generation set sizes $(625, 1250, 6250, 12500, 25000)$ and number of sampling runs $(10, 100, 500, 1000, 5000)$ to verify that these findings would hold. For the most part they do, though there were some notable surprises. Except for TA and VA under the binomial distribution, all methods had their performance be relatively consistent for a given ROC-generation set size regardless of the number of sampling runs. Under the binomial distribution, the fewer the sampling runs, the wider the bands and thus the more containment. When given only 10 runs (and 100 as well for TA), the methods had 100% containment for ROC-generation size $\geq 6250$. However, this quickly went to 0% containment as the number of sampling runs increased or the ROC-generation size decreased. TA otherwise generally had containment of less than 10% and VA had containments from 10% to 40%. The remaining four methods showed interesting containment curves, however, as shown in Table 3.

Some immediate patterns emerge. All methods performed equally "badly" when given smaller ROC-generation sets (and this got worse if we made the ROC-generation sets even smaller). Second, we see that both FWB and SJR are very consistent. Interestingly, looking at different number of sampling runs further strengthens the case for FWB, while SJR has a higher variance in coverages, although that are all above 89% for ROC-generation sizes $\geq 1250$. Interestingly, we see that WHB-s actually starts to perform worse as we increase the ROC-generation size, suggesting it might have a performance curve similar to that of WHB-p but with a wider peak.

## 6 Discussion and Future Work

In this paper we evaluated various methods for generating confidence bands for ROC curves. We adapted two methods from the machine learning literature and introduced three methods from the medical field. We described our general framework, based on the bootstrap, for generating confidence boundaries for ROC curves and empirically evaluating whether they hold at their given confidence level.

Not surprisingly, methods that generate confidence intervals (1-dimensional boundaries) did not translate well to confidence bands and generated bands that were too tight. These included vertical averaging (VA), threshold averaging (TA) and pointwise Working-Hotelling bands. Surprisingly, the simultaneous Working-Hotelling bands, while at first seemingly robust, did not hold up as we varied the parameters for generating the confidence bands. In all fairness, the failure of VA and TA probably can be contributed to our naive methodology for converting them into bands.

Two of the methods used in medical literature for generation of simultanous confidence boundaries did turn out to be relatively robust to changes in the number of samples used for generating the confidence boundaries and the number of instances making up each sample. The simultaneous joint confidence region method, while having higher overall variance, is easy to use and does not require any samples in order to generate the confidence bands. The fixed-width confidence bands, while requiring the bootstrap to empirically determine the proper width, turned out to be very stable and consistently achieve the desired containment of ROC curves used for the verification.

Surprisingly, all of the methods were robust and did not change performance markedly when we varied the number of sampling curves used to generate the confidence bands. More surprising, none of the methods were able to generate confidence bands with the desired coverage as we lowered the

[8] Recall that the bands generated by the TA method are overly conservative and that better bands may be found with a better connecting method.

| Method | verification set size | | | | |
|---|---|---|---|---|---|
| | 625 | 1250 | 6250 | 12500 | 25000 |
| FWB | 56.2( 1.2) | 82.3(28.8) | 95.7( 0.9) | 95.6( 0.7) | 95.1( 0.9) |
| SJR | 62.4( 3.6) | 91.8( 3.0) | 96.2( 7.7) | 97.0( 4.3) | 96.5( 4.0) |
| WHB-p | 54.4( 3.0) | 78.4(28.0) | 65.5(15.6) | 7.6( 6.7) | 0.0( 0.0) |
| WHB-s | 56.2( 1.2) | 73.0(38.6) | 99.6( 0.5) | 86.6(13.1) | 35.9(16.8) |

**Table 3.** Containments of FWB, SJR, WHB-p and WHB-s with 1000 verification curves with varying ROC-generation set sizes. Each cell shows the percentage and standard deviation of verification curves completely contained within the created confidence bands.

number of instances drawn in each sampling run. When we lowered this size to 625, the coverage fell to below 50% even for the best performing methods.

One important question to ask, however, is whether these findings hold across a wider variety of classifiers and data sets. Our results were based on one classifier (C4.5R8) and one data set (covertype). The question we would like to address in future work is what are the conditions under which these methods perform well (or poorly). For this, we plan to move away from the classifier and specific data set and characterize the problem based on class skew and how well the model can separate one class from another. Work is currently under way to address this issue.

One promising method recently proposed to us involves using the AUCs from the set of "fitting curves", choosing the curves with the middle $1-\delta$ AUCs and using their upper and lower hulls. We did not have time to implement this method before submission, but plan on incorporating it in future work.

## 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. Roger Stein proposed using the upper and lower hulls from the middle $1-\delta$ AUCs—a strategy we plan to use in future work.

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

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

[2] C. L. Blake and C. J. Merz. UCI Repository of machine learning databases, 1998.

[3] C. E. Bonferroni, 'Teoria statistica delle classi e calcolo delle probabilità', *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62, (1936).

[4] A. P. Bradley, 'The use of the area under the ROC curve in the evaluation of machine learning algorithms', *Pattern Recognition*, **7**(30), 1145–1159, (1997).

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

[6] W. J. Conover, *Practical Nonparametric Statistics*, Wiley, New York, 2nd edn., 1980.

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

[8] Brad Efron and Rob Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1993.

[9] Tom Fawcett, 'ROC Graphs: Notes and Practical Considerations for Data Mining Researchers', Technical Report HPL-2003-4, HP Labs, (2003).

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

[11] Ron Kohavi, 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1137–1143, San Francisco, CA, (1995).

[12] P. Legendre and L. Legendre, *Numerical Ecology*, Elsevier, 2nd English edn., 1998.

[13] Guangqin Ma and W. J. Hall, 'Confidence bands for receiver operating characteristic curves', *Medical Decision Making*, **13**, 191–197, (1993).

[14] Sofus Attila Macskassy, Haym Hirsh, Foster Provost, Ramesh Sankaranarayanan, and Vasant Dhar, 'Intelligent information triage', in *The 24th Annual Internationl Conference on Research and Development in Information Retrieval (SIGIR)*, New Orleans, LA, (September 2001).

[15] D. D. Margineantu and T. G. Dietterich, 'Bootstrap methods for the cost-sensitive evaluation of classifiers', in *International Conference on Machine Learning, ICML-2000*, pp. 582–590, (2000).

[16] Charles E. Metz, Benjamin A. Herman, and Cheryl A. Roe, 'Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets', *Medical Decision Making*, **18**, 110–121, (1998).

[17] Charles E. Metz, Benjamin A. Herman, and Jong-Her Shen, 'Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data', *Statistics in Medicine*, **17**, 1033–1051, (1998).

[18] Rupert S. Miller, *Simultaneous statistical inference*, Springer Verlag, 2nd edn., 1981.

[19] Kwong-Bor Ng and Paul Kantor, 'Predicting the effectiveness of naive data fusion on the basis of system characteristics', *Journal of American Society for Information Science*, **51**(13), 1177–1189, (2000).

[20] T. Niblett, 'Constructing decision trees in noisy domains', in *Proceedings of the Second European Working Session on Learning*, pp. 67–78, Sigma, Bled, Yugoslavia, (1987).

[21] Foster Provost and Pedro Domingos, 'Tree induction for probability-based rankings', *Machine Learning*, (2002). to appear.

[22] Foster Provost and Tom Fawcett, 'Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions', in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*,

pp. 445–453, (1997).
[23] Foster Provost and Tom Fawcett, 'Robust classification for imprecise environments', *Machine Learning*, **42**(3), 203–231, (2001).
[24] Foster Provost, Tom Fawcett, and Ron Kohavi, 'The case against accuracy estimation for comparing induction algorithms', in *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 445–453. Morgan Kaufman, (1998).
[25] J. Ross Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.
[26] John A. Swets, 'Measuring the accuracy of diagnostic systems', *Science*, **240**, 1285–1293, (1988).
[27] Julian Tilbury, Peter Van Eetvelt, Jonathan Garibaldi, John Curnow, and Emmanuel Ifeachor, 'Receiver operating characteristic analysis for intelligent medical systems – a new approach for finding non-parametric confidence intervals', *IEEE Transactions on Biomedical Engineering*, **47**(7), 952–963, (2000).
[28] H. Working and H. Hotelling, 'Application of the theory of error to the interpreation of trends', *Journal of the American Statistical Assocation*, **24**, 73–85, (1929).
[29] Kelly H. Zou, W. J. Hall, and David E. Shapiro, 'Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests', *Statistics in Medicine*, **16**, 2143–2156, (1997).
[30] M. H. Zweig and G. Campbell, 'Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine', *Clinical Chemistry*, **39**, 561–577, (1993).