Rate-Oriented Point-Wise Confidence Bounds for ROC Curves

Louise A.C. Millard[1,2], Meelis Kull[1], and Peter A. Flach[1,2]

[1] Intelligent Systems Laboratory, University of Bristol, United Kingdom
[2] MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, United Kingdom
{louise.millard,meelis.kull,peter.flach}@bristol.ac.uk

Abstract. Common approaches to generating confidence bounds around ROC curves have several shortcomings. We resolve these weaknesses with a new 'rate-oriented' approach. We generate confidence bounds composed of a series of confidence intervals for a consensus curve, each at a particular predicted positive rate (PPR), with the aim that each confidence interval contains new samples of this consensus curve with probability 95%. We propose two approaches; a parametric and a bootstrapping approach, which we base on a derivation from first principles. Our method is particularly appropriate with models used for a common type of task that we call rate-constrained, where a certain proportion of examples needs to be classified as positive by the model, such that the operating point will be set at a particular PPR value.

Keywords: Confidence bounds, rate-averaging, ROC curves, rate-constrained.

1 Introduction

ROC curves are informative visualisations of model performance that show the ranking performance at different regions of a ranking, or the performance of a scoring classifier at each possible choice of operating point. ROC curves are often used to determine if one model is better than other, and confidence bounds provide a measure of the uncertainty such that this can be determined, for a specified confidence level. In general when several independent sample ROC curves are generated, such as with m-fold cross validation, the variation between them can be used to estimate a confidence around the average (consensus) ROC curve. Several methods have been proposed to generate confidence bounds, mainly parametric approaches such as vertical [13] or threshold [5] averaging.

Vertical averaging is the most common approach, where the false positive rate is fixed and the mean and confidence interval across the true positive rate is calculated at each false positive rate value. Horizontal averaging is a similar approach that instead fixes the true positive rate and calculates the confidence interval across false positive rate values. However, these approaches have several shortcomings. Firstly, the false and true positive rates are metrics over which

T. Calders et al. (Eds.): ECML PKDD 2014, Part II, LNCS 8725, pp. 404–421, 2014.

we have little control, such that it is difficult to set a threshold at a particular value. It is therefore preferable to evaluate a ROC curve with respect to a metric with which setting the threshold is simple in practice. Furthermore, vertical and horizontal averaging are not invariant to swapping the classes, such that if the x-axis and y-axis of ROC space become the false and true negative rate respectively, equivalent points will have different confidence bounds. Finally, depending on the distributional assumptions of points at each false (or true) positive rate value, the confidence bounds may not be constrained to the bounds of ROC space, such that $tpr \in [0, 1]$ and $fpr \in [0, 1]$ (where tpr and fpr are the true and false positive rates respectively).

Threshold-averaging is similar to vertical (and horizontal) averaging but instead fixes the score and averages over each cloud of points in ROC space with the same score. This has the advantage that we can easily use thresholds set at a particular score, classifying each example by whether its score is below or above this threshold value. However, how best to generate confidence bounds for a set points that are not constrained to a single dimension is not obvious. Fawcett et al [5] suggest averaging separately across false and true positive rates, but this creates a rectangular shaped bound for each score where a smoother bound would seem more natural.

To address these shortcomings of existing methods, we specify a set of properties we would like our confidence bounds to satisfy. Firstly, the generated confidence bounds should be invariant to swapping the classes, by which we mean that if the positive and negative classes are swapped such that the x-axis and y-axis of ROC space refer to the false and true negative rate respectively, of the original class labels, then the confidence bounds of these two ROC curves should be symmetrical about the line $tpr = 1 - fpr$ (the descending diagonal). Secondly, the confidence bounds should be constrained to sit within the bounds of ROC space at all points along the lower and upper confidence bounds.

Furthermore, there is a specific type of task in which we are particularly interested. A task may be constrained to a certain proportion of examples that should be classified as positive by the model, the predicted positive rate (PPR). We call these tasks *rate-constrained*, and these are common in many fields. For example, screening a database of customers to decide who should be targeted in a direct sales campaign, where time and monetary budgets mean it is only possible to approach a proportion of the potential customers. Furthermore, the PPR value, hence also the operating point it infers, may not be known precisely, such as the task described by Millard et al. [12], of ranking research articles for rapid reviews in epidemiology.

We suggest that when a task is rate-constrained, the consensus curve should be generated by averaging a set of sample ROC curves while fixing the rate, which we call *rate-averaging*. Furthermore, the comparison of several models should use confidence intervals also created at each PPR value, which we call a *rate-oriented* approach, such that they can be compared with respect to the PPR. We illustrate this with Figure 1, which shows two ROC curves and their consensus curve, created by vertical- (left) and rate-averaging (right). Each point on the

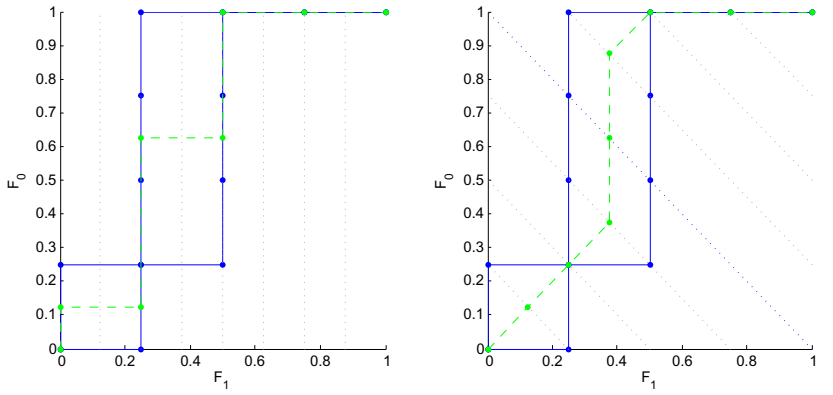


Fig. 1. Illustration of generating consensus curves (broken green curves) from two ROC curves. Left: vertical-averaging, right: rate-averaging. Dotted lines show false positive rate and PPR isometrics, in the left and right figures, respectively.

rate-averaged consensus curve gives the average performance of all sample ROC curves at a particular PPR value. In order for the confidence interval to give the uncertainty of a rate-averaged consensus curve, this should also be generated for each PPR value.

Our aim is to generate confidence intervals for a consensus curve at each PPR value, such that at significance level σ new samples generated from this consensus curve pass between the lower and upper confidence limits at a given PPR value, with probability $1 - \sigma$. The series of confidence intervals creates a *confidence bound* around the consensus curve. We call these *point-wise confidence bounds* in line with [10] in order to differentiate from the common meaning of ROC confidence bands, where the confidence refers to the proportion of whole curves sitting entirely inside the confidence band. Where we discuss methods that are solely used to generate a bound around the whole curve, we explicitly refer to these as bands.

Our main contribution is an approach to generate rate-oriented point-wise confidence bounds. We derive our approach from first principles and demonstrate its effectiveness experimentally.

2 Notation and Basic Definitions

We follow the notation of [7]. We assume a two-class classification problem with instance space $\mathcal{X}$. The positive and negative classes are denoted by 0 and 1, respectively. The learner outputs a score $s(x) \in [0, 1]$ for each instance $x \in \mathcal{X}$. The score densities (lower scores suggest positive class) and cumulative distributions are denoted by f_y and F_y for class $y \in \{0, 1\}$. Given a threshold at score t the true positive rate (also called sensitivity or positive recall) is $P(s(x) \leqslant t|y =$

$0) = F_0(t)$ and the false positive rate is $P(s(x) \leqslant t|y = 1) = F_1(t)$. The true negative rate, also called specificity or negative recall, is $1 - F_1(t)$.

The proportions of positives and negatives are denoted by π_0 and π_1 respectively. The score density of the mixed distribution is denoted by f and given by:

$$f(t) = \pi_0 \cdot f_0(t) + \pi_1 \cdot f_1(t) \tag{1}$$

The probability of a positive at score t is given by:

$$\pi_{0,t} = \frac{\pi_0 \cdot f_0(t)}{\pi_0 \cdot f_0(t) + \pi_1 \cdot f_1(t)} \tag{2}$$

The cumulative distribution of the mixed density distribution is denoted by F and given by:

$$F(t) = \pi_0 \cdot F_0(t) + \pi_1 \cdot F_1(t) \tag{3}$$

This is also the proportion of positive predictions at threshold t known as the predicted positive rate (PPR), which we abbreviate to the rate.

A ROC curve is a plot of true positive rate on the y-axis against false positive rate on the x-axis. A ROC table, such as that shown in Table 1, is a matrix with m rows and n columns, containing the results of independent tests using m samples, such as m-fold cross-validation.

Table 1. Example ROC table, with $m = 4$ samples and n columns, numbers of positive examples in each column pos_k

	k					
$y_{i,k}$	1	2	3	...	n-1	n
Sample 1	0	0	1	...	1	1
Sample 2	0	1	1	...	0	1
Sample 3	0	0	0	...	0	1
Sample 4	0	1	0	...	1	1
pos_k	4	2	2	...	2	0

Table 2. Example $S_{i,k}$ values (number of positive examples up to column k in a sample) for example ROC table(left)

$S_{i,k}$	$S_{i,1}$	$S_{i,2}$	$S_{i,3}$
Sample 1	1	2	2
Sample 2	1	1	1
Sample 3	1	2	3
Sample 4	1	1	2

Each cell contains the label $y_{i,k} \in [0, 1]$ of the example at position k along the ranking of sample i, where the examples of each sample are ranked by increasing score. A segment of consecutive positions in a ranking having the same score are assigned a fractional label to account for this – the average of the labels in this segment, calculated as $\frac{1}{1+q'-q} \sum_{j=q}^{q'} y_j$ where q and q' are, respectively, the start and end of the position range with equal score. The number of positives and negatives in a ranking are denoted by n_0 and n_1 respectively, such that $n = n_0 + n_1$.

The number of positives across samples at column k in the ROC table, denoted pos_k, is given in Equation 4 (and examples are given in Table 1). The number of positives up to position k of row i in the ROC table, which we refer to as the *true positive value* (as opposed to the true positive rate) and denote by $S_{i,k}$, is

given in Equation 5 (and examples are given in Table 2 for the ROC table shown in Table 1).

$$pos_k = \sum_{i=1}^{m} (1 - y_{i,k}) \quad (4) \qquad s_{i,k} = \sum_{j=1}^{k} (1 - y_{i,j}) \quad (5)$$

The number of positives up to position k across all samples in the ROC table, denoted s_k, is given by:

$$s_k = \sum_{j=1}^{k} pos_j = \sum_{i=1}^{m} s_{i,k} \quad (6)$$

Recall (of the positive class) is the proportion of positive examples, correctly classified as positive, at a given point on the ROC curve (also known as the true positive rate). We specify this in terms of rates. The recall $tpr_{i,k}$ of sample i with operating point at position k is given by:

$$tpr_{i,k} = \frac{s_{i,k}}{n_0} \quad (7)$$

We denote an unsorted list of n items as $a_1, a_2 \ldots a_n$ and a sorted list as $a_{(1)}, a_{(2)} \ldots a_{(n)}$.

3 Generating Confidence Bounds

In this section we give our approach to generating rate-oriented point-wise confidence bounds. This includes a new approach to generating samples that uses the ROC curve, rather than the common approach of sampling from the score probability density function of each class (Section 3.2). We begin by describing a simple approach of inferring confidence bounds, used as a baseline in our experiments (Section 4).

3.1 Baseline Method

We use a simple parametric approach as a baseline method. This method is similar to previous approaches such as vertical-averaging, but we fix the rate rather than the false positive rate, in line with our aims. We calculate the mean and variance of recall across samples and, after making an assumption of the underlying distribution across the ROC points of each sample at each rate, calculate the 95% confidence intervals. Here we use positive recall as a distance measure along rate isometrics in ROC space, but any metric that varies linearly along rate isometrics could also be used (such as negative recall or accuracy).

The variance of mean sample recall at each position k along the ranking is given by:

$$\sigma_k^2 = \frac{1}{m \cdot (m-1)} \sum_{i=1}^{m} (tpr_{i,k} - \overline{tpr}_k)^2 \quad (8)$$

where m is the number of samples, $tpr_{i,k}$ is the recall for sample i at position k and $\overline{tpr}_k$ is the mean recall across the samples, at position k. The additional m in the denominator is because we need the variance of the sample mean.

In order to infer a confidence interval we need to assume a particular distribution across the recall at each position k. Assuming a normal distribution the confidence intervals are given by $\overline{tpr}_k \pm 1.96 \cdot \sigma_k$.

We also test this method using a beta distribution, which is bounded by $[0, 1]$ such that we can constrain our confidence intervals to the bounds of ROC space. To use the beta distribution we rescale, at each position k, each sample recall value from the range $max\left(0, \frac{r-\pi_1}{\pi_0}\right) \ldots min\left(1, \frac{r}{\pi_0}\right)$, where $r = \frac{k}{n}$ is the rate at k, to the range $0 \ldots 1$. We calculate the mean and standard deviation of these scaled recall values at each position k and use these to calculate the α and β parameters of the beta distribution. We find the lower and upper limits of the 95% confidence interval of this distribution, and then rescale this back to the original range.

3.2 Generating Sample ROC Curves

Given the score densities of each class, sample rankings can be generated using this distribution. For instance, for each example in the new ranking we can sample a score from the mixed distribution and then sample a label using the probabilities of each class at this score (shown in Table 3 left).

However, the score distribution is not determined by the ROC curve and hence may not be known. In this case we can sample using the ROC curve instead of the score densities, by sampling across the rate. The gradient on the ROC curve is the class likelihood ratio, from which we can calculate the class probabilities at this point on the curve, and then sample the label using this.

We do not need to know the scores because the rate also determines the order of the examples in the ranking, and the ROC curve determines the class probabilities at each rate. We call this the 'rate-first' approach, given in Table 3 (right).

Table 3. Two sampling approaches. Left. Score-first approach. Right: Rate-first approach.

Score-first:	Rate-first
Repeat n times:	Repeat n times:
Sample score $s_j \sim f$	Sample rate $r_j \sim uniform(0, 1)$
Sample label $y_j \sim bernoulli(\pi_{0,s_j})$	$\pi_{0,r_j} \leftarrow$ calculated from gradient at r_j on ROC curve
Rank labels by score s_j	Sample label $y_j \sim bernoulli(\pi_{0,r_j})$
	Rank labels by rate r_j

3.3 Overview of Our Approaches

We assume a random process that generates ROC tables of size $n \cdot m$ from the usually unknown score densities. Let us denote by $S_{i,k}$ the random variable of

the sum of the number of positives at position k. Formally, for any fixed true positive value s at this position, with n_0 and n_1 all fixed, we want to estimate:

$$p\left(S_{i,k} = s \mid S_{i,n} = n_0\right) = \frac{p\left(S_{i,k} = s, S_{i,n} = n_0\right)}{\sum_{s'} p\left(S_{i,k} = s', S_{i,n} = n_0\right)} \tag{9}$$

We condition on the class distribution to reflect the fact that a data sample has a finite number of examples with a certain number of each class. This also corresponds to the fact that ROC curves must pass through the points $(0,0)$ and $(1,1)$. We present two alternative methods, a parametric and a bootstrap approach. We derive the probability distribution across the number of positives up to a position, k, in a sample, and use this to infer these two approaches. We develop bootstrap approaches for cases where the distributional assumptions of the parametric approach are invalid.

Importantly, our approach is naturally invariant to swapping the classes. In ROC space, swapping the classes means that the x-axis becomes the false negative rate $(1 - tpr)$, and the y-axis becomes the true negative rate $(1 - fpr)$. The corresponding ROC curve in this 'swapped' ROC space is simply a line mirroring of the original ROC curve along the descending diagonal ($tpr = 1 - fpr$). Furthermore, the rates are given by $r'(t) = \pi_0(1 - tpr) + \pi_1(1 - fpr)$. Therefore it follows that $r'(t) = 1 - r(t)$. Hence, for each set of points along a rate isometric in the original space, there is a corresponding rate isometric in the 'swapped' space along which this set of points also lie. The confidence bands along these corresponding rate isometrics will have equivalent confidence intervals.

3.4 Parametric Approach

We find the probability distribution across the number of positives from the first position to a position k in the ranking, $S_{i,k}$. We first derive an analytical solution (Theorem 1), and then provide an empirical version that can be used when only the ROC curve (and not the score densities) is available, as is usually the case. At this point we fix i as we refer only to a single sample, such that $S_{i,k}$ is denoted S_k and $S_{i,n}$ is denoted S_n.

Theorem 1. *Let the score densities, F_0 and F_1, and the number of examples of each class in the sample, n_0 and n_1, be fixed. Then:*

$$\begin{aligned} &p(S_{i,k} = s, S_{i,n} = n_0) \\ &\quad = \int_0^1 \left[binom(s, k-1, \pi_0^{<r}) \cdot (1 - \pi_0^{=r}) + binom(s-1, k-1, \pi_0^{<r}) \cdot (\pi_0^{=r})\right] \\ &\qquad\qquad \cdot binom(n_0 - s, n - k, \pi_0^{>r}) \cdot p(R_k = r) dr \end{aligned} \tag{10}$$

where

$$\pi_0^{<r} = \frac{\pi_0 F_0(t)}{\pi_0 F_0(t) + \pi_1 F_1(t)} \quad (11) \qquad \pi_0^{>r} = \frac{\pi_0(1 - F_0(t))}{\pi_0(1 - F_0(t)) + \pi_1(1 - F_1(t))} \quad (12)$$

$$\pi_0^{=r} = \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)} \quad (13)$$

$t = F^{-1}(r)$, $p(R_k = r) = beta(r, k, n - k + 1)$, R_k is the rate from which the example at position k was sampled and $binom(k_b, n_b, p_b)$ is the binomial distribution for k_b successes in n_b trials, with probability of success p_b, and $beta(x, a, b)$ is the probability of value x for beta distribution with $\alpha = a$ and $\beta = b$.

Proof. To compute the left hand side of Equation 9 it is sufficient to compute:

$$p\left(S_k = s, S_n = n_0\right) \quad (14)$$

The probability of $S = s$ and $S_n = n_0$ in the new sample depends on which rate it was sampled from, such that:

$$p\left(S_k = s, S_n = n_0\right) = \int_0^1 p\left(S_k = s, S_n = n_0 \mid R_k = r\right) \cdot p(R_k = r) dr \quad (15)$$

The order statistic states that when sampling n values uniformly within the range 0..1 and sorting these examples, the probability that an example at position k was sampled from a rate r is beta distributed with $\alpha = k$ and $\beta = n - k + 1$ [1]. Therefore, $p(R_k = r)$ of Equation 15 is the beta density.

The other component of Equation 15 is the probability of s positives up to a position k, given the example at this position is sampled from a particular rate r. There are two cases where value s is the number of positives up to a position k: 1) $s-1$ positives occur before position k and the example at k is a positive, or 2) s positives occur before position k and the example at position k is a negative. In either case there must also be $n_0 - s$ positives after position k to ensure that the class distribution is correct.

The examples before position k can be sampled independently, with probability of a positive given by Equation 11. The examples after position k can also be sampled independently, with probability of a positive given by Equation 12. The independence between samples is valid because we are sampling a set of *unordered* examples, and this means that the probabilities of the set of exam-

ples before and after position k are binomially distributed, which infers:

$$
\begin{aligned}
&p\left(S_k = s, S_n = n_0 | R_k = r\right) \\
&= \left[p\left(\sum_{i=1}^{k-1}(1-y_i) = s\right) p(y_k = 1) + p\left(\sum_{i=1}^{k-1}(1-y_i) = s-1\right) p(y_k = 0) \right] \\
&\qquad \cdot p\left(\sum_{i=k+1}^{n}(1-y_i) = n_0 - s\right) \\
&= \left[binom(s, k-1, \pi_0^{<r}) \cdot (1-\pi_0^{=r}) + binom(s-1, k-1, \pi_0^{<r}) \cdot (\pi_0^{=r}) \right] \\
&\qquad \cdot binom(n_0 - s, n-k, \pi_0^{>r})
\end{aligned}
\tag{16}
$$

Using Equation 16 in Equation 15 concludes the proof.

□

To reiterate, a key point - while an example at position k has rate $r = \frac{k}{n}$ for this ROC table, we can imagine this table is sampled from a ROC curve of all possible examples. The rate from which it is sampled from this 'true' ROC curve is probabilistic, corresponding to $p(R_k = r)$ in Equation 10. The class probabilities used to generate this example are determined by the class distribution at the rate from which this example was sampled.

An important aspect of Theorem 1 is that the sampling probabilities before, at and after rate r (Equations 11 - 13) can be computed solely using the ROC curve. Recall from Section 3.2 that Equation 13 can be calculated from the gradient at r on the ROC curve. We can also infer the values of Equations 11 and 12 from the ROC curve. Equation 11 is equivalent to the average probability of sampling a positive across all rates before r, and this can be inferred from the gradient of the straight line from point $(0, 0)$ to the point at r on the ROC curve. Similarly, Equation 12 can be inferred from the gradient of the straight line from the ROC curve point at r to the point $(1, 1)$.

Theorem 1 gives the analytical calculation but we cannot use this directly in practice, as we have empirical ROC curves / ROC tables rather than the score densities. Firstly, our empirical ROC tables have discrete rates such that in the discrete case the integral of Equation 15 is changed to a summation. We implement this as an average of the joint probability, for a set of rates of the CDF of the beta distribution (the sampling distribution for this k) at each 0.01 interval:

$$
p\left(S_k = s, S_n = n_0\right) = \sum_{t=1}^{99} p\left(S_k = s, S_n = n_0 \mid R_k = F_{beta}^{-1}(0.01 \cdot t)\right) \tag{17}
$$

such that we sample the rates at each 0.01 interval of the CDF of the beta distribution (with $\alpha = k$ and $\beta = n - k + 1$). This CDF models the probability

that an example at position k is sampled by each rate (according to the order statistic).

We also require discrete versions of Equations 11- 13 that can also be used with an empirical ROC table, and these are given in Equations 18- 20:

$$\pi_0^{<r} = \frac{1}{r \cdot n \cdot m} \left[S_{i,\lfloor r \cdot n \rfloor} + d \cdot pos_{\lceil r \cdot n \rceil} \right] \tag{18}$$

$$\pi_0^{=r} = \frac{1}{m} pos_{\lceil r \cdot n \rceil} \tag{19}$$

$$\pi_0^{>r} = \frac{1}{(1-r) \cdot n \cdot m} \left[n_0 - S_{i,\lfloor r \cdot n \rfloor} + (1-d) \cdot pos_{\lceil r \cdot n \rceil} \right] \tag{20}$$

where $d = r \cdot n - \lfloor r \cdot n \rfloor$ is the relative distance of the rate between positions $\lceil r \cdot n \rceil$ and $\lfloor r \cdot n \rfloor$.

The probabilities of each S_k value computed in Theorem 1, correspond to only a single row of the ROC table. We need the distribution across the number of positives up to position k of all samples in the ROC table. For each S_k value we need:

$$p\left(S_k = s \mid \forall i \in 1 \ldots m : \sum_{j=1}^{n} (1 - y_{i,j}) = n_0 \right) \tag{21}$$

Computing this exactly is computationally intractable, as for each possible s at a position k the probability is given as the summation of the probabilities of all possible combinations of values at position k that sum to this value. We instead approximate the confidence intervals using the estimated variance of this distribution. The mean and variance of the distribution of one sample up to position k are given by:

$$\mu_{1,k} = \sum_s p\left(S_k = s \mid S_n = n_0\right) \cdot s \tag{22}$$

$$\sigma_{1,k}^2 = \sum_s p\left(S_k = s \mid S_n = n_0\right) \cdot \left(s - \mu_{1,k}\right)^2 \tag{23}$$

where 1 denotes that these functions correspond to a single sample. We assume each row is identically distributed such that the mean and variance of s at position k of the ROC table are given by:

$$\mu_k = \sum_{i=1}^{m} \mu_{i,k} = m \cdot \mu_{1,k} \tag{24}$$

$$\sigma_k^2 = \sum_{i=1}^{m} \sigma_{i,k}^2 = m \cdot \sigma_{1,k}^2 \tag{25}$$

At each k we restrict to only the possible values of S_k, rescale these to between zero and one, and use a scaled beta distribution to model this distribution and estimate the confidence intervals. We calculate the mean and variance across S_k

values at each position k, where the S_k values have been rescaled to the range $[0, 1]$:

$$\mu_{k,\beta} = \frac{\mu_k - minS_k}{maxS_k - minS_k} \tag{26}$$

$$\sigma^2_{k,\beta} = \frac{\sigma^2_k}{(maxS_k - minS_k)^2} \tag{27}$$

where $maxS_k = m \cdot maxS_{1,k}$ and $minS_k = m \cdot minS_{1,k}$ and:

$$minS_{1,k} = max(0, n_0 - n + k) \tag{28}$$

$$maxS_{1,k} = min(k, n_0) \tag{29}$$

We use these to parameterise a beta distribution and infer a confidence interval, which we then rescale to the original scale.

3.5 Bootstrap Approach

We generate 2,000 bootstrapped ROC tables each with m samples. Each sample is generated independently using the rate-first sampling approach, as follows.

The rates are sampled uniformly and sorted:

$$r_1, r_2 \ldots r_n \xrightarrow{sort} r_{(1)}, r_{(2)} \ldots r_{(n)} \tag{30}$$

The probability distribution at each rate is found by:

$$\pi_{0,r} = \frac{1}{m} pos_{\lceil r \cdot n \rceil} \tag{31}$$

We then use this probability to generate a label at k:

$$l_k \sim binom(\pi_{0,r}) \tag{32}$$

In this way we generate a set of 2,000 bootstrap ROC tables (generating $2,000 \cdot m$ samples in total).

This sampling procedure does not ensure that each sample has the correct class distribution. This is needed so that the confidence intervals generated from these samples reflect that at rates 0 and 1 we are certain the curve passes through the points $(0, 0)$ and $(1, 1)$ in ROC space, respectively. A simple approach to restrict to a fixed class distribution discards all samples where the class distribution is not correct. However, this approach is only feasible when the number of examples is low, as otherwise samples are rarely generated with the correct class distribution and this method becomes too slow.

We propose another approach that can be used with a larger number of examples, where we adjust the rate and the number of true and false positives at each position in order to correct the class distribution. The rates of the bootstrap ROC tables are equally distributed along the ranking, as shown in Figure 2.

For each sample individually we adjust these rates and the true positive values at each position, by scaling each position according to a correction factor, a value for each sample and class that rescales the 'width' of each example in the ranking to correct the class distribution. This adjustment is illustrated in Figure 2, and shows how the effect is to stretch or narrow the examples along the ranking.

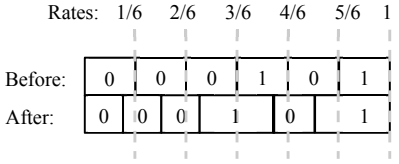


Fig. 2. Illustration of rate adjustment to correct class distribution

We use the bootstrapped ROC tables with the corrected true positive values, to estimate the confidence bound of the true ROC curve. For each ROC table, and at each position k along the ranking, we calculate the average recall across the samples:

$$\overline{tpr}_k = \frac{1}{m \cdot n_0} \sum_{i=1}^{m} s_{i,k} \tag{33}$$

Each position k in the ranking has a set of average recall values, one for each sample ROC table. This now corresponds to the probability density function we stated in Equation 9. The proportion of bootstrap ROC tables with recall value between $\overline{tpr}_k$ and $\overline{tpr'}_k$ gives an estimate of the probability that the recall at this position is between these values, given this sample has a particular class distribution.

The confidence interval for position k is obtained from the mean recall values, $\overline{tpr}_k$, of the bootstrapped ROC tables as follows. For each position k we take the $\overline{tpr}_k$ value of each ROC table, sort these values in ascending order, and select the 2.5% and 97.5% percentiles as the lower and upper endpoints of the 95% confidence interval. This gives a series of recall-rate pairs for the lower and upper limits of the confidence interval at each position k. A confidence bound can be created by interpolating between these points.

4 Experiments

Our experiments use a known ROC curve to generate samples for which we create confidence bounds, specified by normally distributed score density functions with mean 0 and 1 for the positive and negative class respectively, and a variance of 1. These score distributions, and the corresponding ROC curve are shown in Figure 3. Our tests use ROC tables with 10 samples and 50 examples per sample.

We evaluate whether the generated confidence intervals meet our aims, where at significance level σ new samples generated from this consensus curve pass

between the lower and upper confidence limits at a given PPR value, with probability $1 - \sigma$. Given a single sample ROC table and its confidence bounds, we generate 1,000 new sample ROC tables from this sample. We count, at each rate, the number of consensus curves (of these samples) the confidence interval contains. A true 95% confidence interval at a given rate, should contain the consensus curve of new samples 95% of the time.

The results are shown in Figure 4. The results of the basic parametric approaches (Figures 4a and 4b) are highly variable. Our parametric approach (Figure 4c) reliably generates confidence bounds with close to 95% confidence, except at the extremes. This indicates that the assumption that the number of positives up to a particular position in the ranking is beta distributed is not valid in these regions.

Our bootstrap approaches are also much more effective compared to the baseline results. They are a little conservative, particularly at the extremes of the distribution, due to the nature of bootstrap sampling, where the variation between bootstraps may be too low to calculate strict confidence intervals (for instance, where the lower and upper bounds of the 95% limits are the same as those for the 94 or 96% limits). For example if a bootstrap sample contained only one value then the values at the 95% bounds would also be the same values as for the 1% or 100% limits. This also justifies the shape of the graph in Figure 4f, as where rates have a probability of a positive near to 1, there is little variation across samples.

Figure 5a shows an example ROC curve generated using our analytical approach, and the equivalent rate-recall curve is shown in Figure 5b.

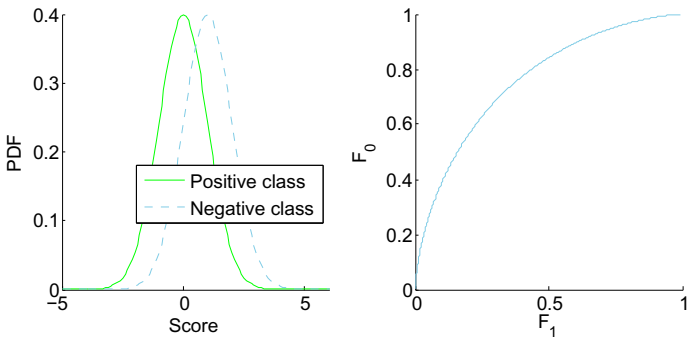
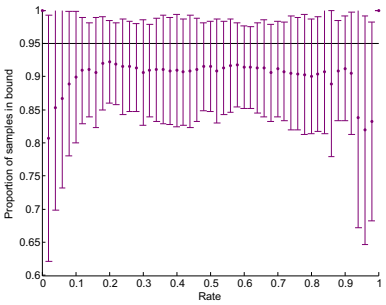
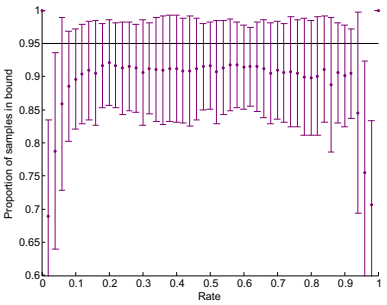


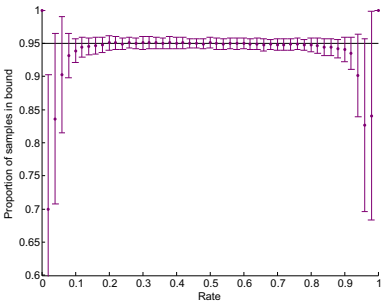
Fig. 3. Score probability densities for two classes (positive class: $\mu = 0$, $\sigma^2 = 1$; negative class: $\mu = 1$, $\sigma^2 = 1$), and corresponding ROC curve



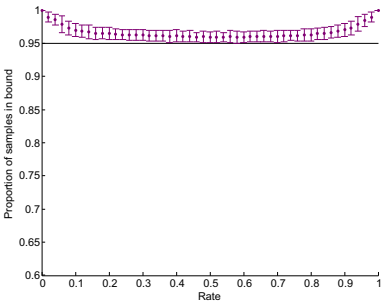
(a) Results of baseline with normal assumption



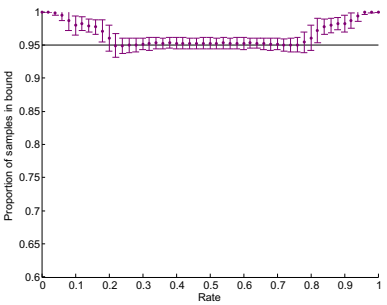
(b) Results of baseline with beta assumption



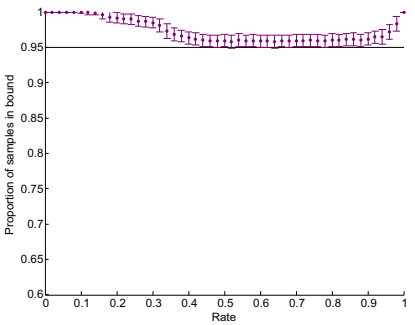
(c) Results of parametric approach



(d) Results of bootstrap approach (with discarding)

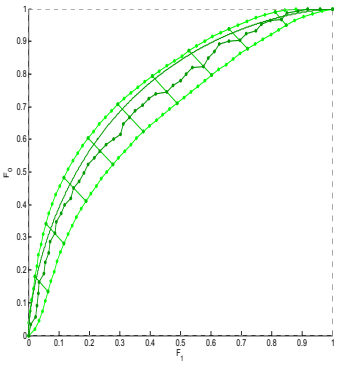


(e) Results of bootstrap approach (with adjustment)

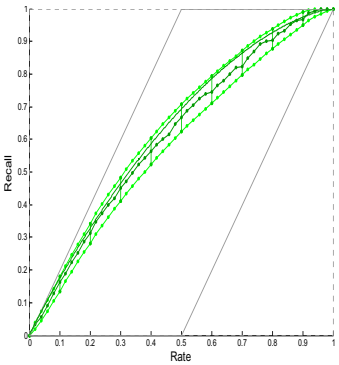


(f) Results of bootstrap approach (with discarding) for score distributions with: $\mu_0 = 0$, $\sigma_0^2 = 1$, $\mu_1 = 1$, $\sigma_1^2 = 0.2$

Fig. 4. Mean (variance) of the proportion of 1000 new samples (sampled from ROC table) within confidence interval at each rate, across 100 tests



(a) Example ROC curve and confidence bounds (confidence intervals at a selection of rates shown for illustration)



(b) Equivalent rate-recall curve [12] and confidence bounds for ROC curve shown in Figure 5a. Grey lines indicate bounds of rate-recall space.

Fig. 5. Example confidence bounds generated with our parametric approach, and the equivalent rate-recall curve. Also shown are two curves: 1) The smooth true curve specified by the score distributions and 2) The consensus curve from the original sample (also shown in Figure 3 (right)).

5 Related Work

In the introduction we discussed two parametric approaches to generating confidence bounds; vertical (horizontal) and threshold averaging. A non-parametric approach, called fixed width bands [4, 11] works by displacing the whole ROC curve up and left, and down and right, to create an upper and lower confidence band respectively. The curve is displaced along the gradient $-\sqrt{(}N^+/N^-)$ (chosen as an approximation of the standard deviation ratios of the two classes). Rate isometrics have a gradient $-N^+/N^-$ such that if we changed the displacement gradient to the gradient of the rate isometric this could be used as a rate-oriented approach. However, the size of displacement is constant along the ROC curve which does not constrain the confidence bounds to ROC space. Furthermore, this is an approach for calculating the confidence around the whole curve, but in this paper we are interested in point-wise confidence bounds instead.

Other approaches include a non-parametric approach by Tilbury et al., which they derived from first principles [16], and the use of kernel estimation to estimate the continuous probability density functions of the scores of each class [6]. We also refer the reader to comparisons of various approaches, performed by Macskassy et al. [9, 10].

Early retrieval tasks are those where the top ranked examples are of most interest [2], and metrics used for this task weight the importance of an example by its position in the ranking. For instance, the rate-weighted AUC (rAUC) [12] is a general measure where the distribution of weights along the ranking can be chosen for the specific task at hand. Other metrics that are restricted

to particular weight distributions include; discounted cumulative gain (DCG) and normalised discounted cumulative gain (NDCG) [8] in information retrieval, robust initial enhancement (RIE) [14], the Boltzmann-enhanced Discrimination of ROC (BEDROC) [17], concentrated ROC (CROC) [15] and sum of the log ranks (SLR) [18]. These metric all evaluate rankings with respect to the rate, such that when assessing tasks that use these metrics in ROC space, we suggest it is most appropriate to generate rate-averaged consensus curves with rate-oriented point-wise confidence bounds.

Rate-averaging has been previously used [3, 12] to generate consensus curves, referred to as pooling in [3]. To our knowledge there is no approach in the literature to infer rate-oriented confidence bounds. [9] claims that rate-averaging makes the strong assumption that the rates are estimating the same point in ROC space, and this is not appropriate. However, other approaches make this similar assumptions across a different metric, such as the false positive rate in vertical-averaging.

6 Conclusions

We have described a new approach to generate confidence bounds, which we call rate-oriented point-wise confidence bounds. Our main aim was to address some important weaknesses of other existing methods. Calculating the consensus and confidences bounds at each rate is practical as rate is a measure over which we have control in practice. On the other hand, vertical (or horizontal) averaging fix the false positive rate (true positive rate) and average across the true positive rate (false positive rate), but these metrics are not under our control so are of little use in practice. Score-averaging creates confidence bounds around clouds of points, and how best to do this is an open problem. Rate-averaging does not have this problem because it constrains to a single dimension.

Our approach is also invariant to swapping the classes, and we suggest that this property is sensible when generating confidence bounds. The confidence of a point on the ROC curve should not depend on which class is labelled as positive. Furthermore, our bounds have the advantage that they are smooth, due to the sampling across rates we perform as part of our method.

Our secondary aim was to find appropriate bounds for assessing models used specifically for rate-constrained tasks. Using a rate-oriented approach ensured that the performance (and confidence interval) shown at a rate is an estimate for this particular rate.

In this paper we analytically derived the probability distribution of the number of positives up to each position in the ranking, and then used this to develop two methods, a parametric and a bootstrap approach. The parametric approach gave confidence bounds having very close to the 95% confidence, except at the extremes. The bootstrap approach did generate satisfactory bounds at the extreme but also had greater variance around the 95% confidence level. Therefore, we suggest that when the performance at the extremes of the ROC curve are of little importance, the parametric approach should be used, but where this is not the case the bootstrap approach can be used instead.

Acknowledgments. This work is supported by the REFRAME project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Engineering and Physical Sciences Research Council in the UK. LACM is funded by a studentship from the UK Medical Research Council. This work was also supported by Medical Research Council grant MC_UU_12013/1-9.

References

1. Arnold, B.C., Balakrishnan, N., Nagaraja, H.N.: A first course in order statistics, vol. 54. SIAM (1992)
2. Berrar, D., Flach, P.: Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Briefings in Bioinformatics 13(1), 83–97 (2012)
3. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145–1159 (1997)
4. Campbell, G.: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statistics in Medicine 13(5-7), 499–508 (1994)
5. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Machine Learning 31, 1–38 (2004)
6. Hall, P., Hyndman, R.J., Fan, Y.: Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika 91(3), 743–750 (2004)
7. Hand, D.J.: Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning 77(1), 103–123 (2009)
8. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48. ACM (2000)
9. Macskassy, S., Provost, F.: Confidence bands for ROC curves: Methods and an empirical study. In: Proceedings of the First Workshop on ROC Analysis in AI (2004)
10. Macskassy, S., Provost, F., Rosset, S.: Pointwise ROC confidence bounds: An empirical evaluation. In: Proceedings of the Workshop on ROC Analysis in Machine Learning (2005)
11. Macskassy, S.A., Provost, F., Rosset, S.: ROC confidence bands: An empirical evaluation. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, New York, NY, USA, pp. 537–544 (2005)
12. Millard, L.A.C., Flach, P.A., Higgins, J.P.T.: Rate-constrained ranking and the rate-weighted AUC. In: Calders, T., Esposito, F., Hüllermeier, E. (eds.) ECML/PKDD 2014, vol. 8725, pp. 383–398. Springer, Heidelberg (2014)
13. Provost, F.J., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: ICML, vol. 98, pp. 445–453 (1998)
14. Sheridan, R.P., Singh, S.B., Fluder, E.M., Kearsley, S.K.: Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. Journal of Chemical Information and Computer Sciences 41(5), 1395–1406 (2001)
15. Joshua Swamidass, S., Azencott, C.-A., Daily, K., Baldi, P.: A CROC stronger than ROC: Measuring, visualizing and optimizing early retrieval. Bioinformatics 26(10), 1348–1356 (2010)

16. Tilbury, J.B., Van Eetvelt, W., Garibaldi, J.M., Curnsw, W.J., Ifeachor, E.C.: Receiver operating characteristic analysis for intelligent medical systems-a new approach for finding confidence intervals. IEEE Transactions on Biomedical Engineering 47(7), 952–963 (2000)
17. Truchon, J.-F., Bayly, C.I.: Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. Journal of Chemical Information and Modeling 47(2), 488–508 (2007)
18. Zhao, W., Hevener, K.E., White, S.W., Lee, R.E., Boyett, J.M.: A statistical framework to evaluate virtual screening. BMC Bioinformatics 10(1), 225 (2009)