Mach Learn (2010) 78: 103–136
DOI 10.1007/s10994-009-5134-6

Pointwise exact bootstrap distributions of ROC curves

Charles Dugas · David Gadoury

Received: 22 November 2007 / Revised: 27 April 2009 / Accepted: 29 June 2009 /
Published online: 16 July 2009
Springer Science+Business Media, LLC 2009

Abstract We derive pointwise exact bootstrap distributions of ROC curves and the difference between ROC curves for threshold and vertical averaging. From these distributions, pointwise confidence intervals are derived and their performance is measured in terms of coverage accuracy. Improvements over techniques currently in use are obtained, in particular in the extremes of ROC curves where we show that typical drastic falls in coverage accuracy can be avoided.

Keywords Receiver operating characteristics · Bootstrap · Coverage probabilities · Model selection

1 Introduction

Literature on receiver operating characteristics (ROC) curves is abundant and scattered across different communities and this breadth of applicability has been reported in Swets and Pickett (1982) and more recently in Swets et al. (2000). In the machine learning community, ROC curves have recently gained popularity as a tool to measure and visualize the performance of binary classifiers. See (Fawcett 2006) for an excellent introduction to ROC curves along with descriptions of the essential elements of ROC graph analysis, within the context of machine learning. ROC graph analysis can be enhanced if confidence intervals for the curve are provided along with the curve itself as this allows the user to assess the reliability of the estimated performance of a model considered for implementation. When comparing two models through their respective ROC curves, confidence intervals for the difference in performance can be used to assess the significance of the superiority, which we

Editor: Johannes Fürnkranz.

C. Dugas (✉) · D. Gadoury
Department of Mathematics and Statistic, Université de Montréal, 2920 chemin de la Tour, Montréal,
Québec H3T 1J4, Canada
e-mail: dugas@dms.umontreal.ca

D. Gadoury
e-mail: gadoury@dms.umontreal.ca

hereafter refer to as *dominance*, of one model over the other. Unfortunately, in some cases, coverage of confidence intervals can be far from its intended target (usually, 90%, 95%, or 99%). The goal of this paper is to propose confidence intervals, for ROC curves and the difference between two ROC curves, with better and more reliable coverage accuracy than what is obtained using currently available techniques.

When performing model selection (or model comparisons) with ROC curves, this issue becomes paramount: confidence intervals with coverage above target are too conservative (too wide) and will more likely fail to identify significant differences where they should (type II error). Conversely, confidence intervals with coverage below target are too assertive (too narrow) and are more likely to conclude to the dominance of one model over another in cases where this is not true (type I error). When considering a single model, confidence intervals with coverage above (below) target will exhibit an unduly wide (narrow) range of potential performances which may hamper decision-making. There have been numerous suggestions on how to obtain confidence intervals for ROC curves. On the other hand, confidence intervals for the difference between two ROC curves, although of great interest to the machine learning community, have received very little attention. In this paper, this later issue is treated in great details, thereby providing new analysis tools for machine learning research.

Bootstrap resampling of the test set in order to obtain an out-of-sample distribution of a ROC curve or the difference between two ROC curves has previously been used. Confidence intervals obtained through this procedure may vary, depending on the actual resamples drawn and in order to reduce this resampling noise, a larger number of resamples can be drawn. In the limit, an infinite number of resamples will bring resampling noise to zero, thus leading to deterministic confidence intervals, given a specific test set. Although infinite resampling is not feasible, the distribution of the statistic of interest (here, a ROC curve or the difference between two ROC curves) associated to an infinite number of resamples can, in some cases, be derived analytically. Such a distribution is often referred to as an *exact bootstrap* distribution. Computing this exact bootstrap distribution therefore provides the same information as would have been obtained through an infinite number of resamples, without performing any. In this paper we derive exact bootstrap distributions of ROC curves and the difference between two ROC curves. From these distributions, we obtain confidence intervals.

Deriving pointwise distributions for curves requires a specification of how the points of different curves are to be associated and averaged. In this paper, we consider two averaging techniques that have previously been used in the machine learning community: threshold and vertical averaging. Statistical and medical literatures on the subject of ROC curve confidence intervals have exclusively addressed the case of vertical averaging although, as described in Sect. 2, threshold averaging is of obvious practical relevance. So far, it seems that only the fields of machine learning and data mining have considered the issue of threshold averaging.[1]

Combining the two statistics (ROC curves and differences between ROC curves) with the two averaging techniques (threshold and vertical) leads to the four problems addressed in this paper. We provide contributions to each of these four issues by proposing confidence intervals that alleviate the problem of coverage accuracy drops in the extremes of the ROC curve, as observed in particular in Macskassy et al. (2005). The paper presents numerous different theoretical and numerical results. In order to make things clearer for the reader,

[1] A recent exception in engineering is Kerekes (2008).

Table 1 Exact bootstrap distributions of ROC curves and their approximations. Design: some results apply to the distribution of a single curve (Single). These results can also be used to estimate the distribution of the difference between two ROC curves for unpaired designs. Other results apply to the distribution of the difference between two ROC curves when the design is paired (Paired). Approximation: results are for the exact bootstrap distribution itself (Exact) or an approximation of it (Approx.). Sections: sections where relevant theory and numerical results appear

Design (theory)	Values	Approx.	Times[a]	Algorithms, Equations	Sections Theory/Simul.
Threshold averaging					
Single	C.I.	Approx.	$O(n \ln n)$	Algorithm 1	3.1/5.1
Paired	f_{t_1,t_2}	Exact	$O(n \cdot h)$	Algorithm 2, (2)	3.2/–
	C.I.	Approx.	$O(n \cdot h)$	Algorithm 3	3.2/5.2
Vertical averaging					
Single	C.I.	Approx.	$O(n \cdot h)$	Algorithm 4, (4), (6), (7)	4.1/5.3
Paired	C.I.	Exact	$O(n^4)$	Algorithm 5	4.2/5.4

[a] n: sample size, h: number of evaluation points

Table 1 classifies algorithms, equations, and subsections (where theory and numerical results appear) with respect to the problem to which they apply. Also, for convenience, Table 2 provides a summary of the notation used throughout the paper.

The rest of the paper is as follows: in Sect. 2, we discuss in more details the issues related to the computation of confidence intervals for ROC curves and present existing work. Theoretical developments for threshold and vertical averaging are presented in Sects. 3 and 4, respectively. In Sect. 5, we report numerical results. We conclude in Sect. 6.

2 Confidence intervals for ROC curves

This paper considers models that have been trained to discriminate between two classes. We assume each model's output is a score with unknown continuous distribution. Instances for which a certain condition is present (e.g. a tumor is cancerous, a credit card transaction is fraudulous) will be referred to as *positive* instances. Also, without loss of generality, we assume higher scores indicate that an instance is more likely to be positive. Our interest lies in evaluating pointwise confidence intervals, i.e., the dispersion of individual points of the curve. Confidence bands, that define regions within which a certain portion of the curve should lie, are not treated here.

2.1 Experiment designs

We assume model selection is performed by comparing each model's out-of-sample performance. An *unpaired design* refers to the situation where the first model's test set is disjoint from the other model's test set. This may occur in medicine if a patient can only be given one of two possible treatments. In *paired design* experiments, both models are evaluated using the same test set. Obviously, most of the machine learning literature focuses on paired rather than unpaired designs in order to perform model selection. For convenience, we define a *single* design as the situation where only one model is considered.

With unpaired designs, the distribution for the difference between two ROC curves can be obtained by assuming the two individual curves are independent, since the test sets are

Table 2 Notation

Threshold averaging, single design	
n	number of instances in test set
n^+ (n^-)	number of positive (negative) instances in test set
n_t^+ (n_t^-)	number of positive (negative) instances in test set, with score greater or equal to t
m	number of instances in each sample
m^+ (m^-)	number of positive (negative) instances in each sample
M_t^+ (M_t^-)	r.v. for the number of positive (negative) instances in a sample, with score greater or equal to t
TP_t^+ (FP_t^-)	r.v. for the true (false) positive rate at threshold t
relations:	$p_t^+ = n_t^+/n^+$, $TP_t^+ = M_t^+/m^+$, $p_t^- = n_t^-/n^-$, $FP_t^- = M_t^-/m^-$
Threshold averaging, paired design	
$n^+_{\overline{t_1},t_2}$ ($n^-_{\overline{t_1},t_2}$)	number of positive (negative) instances, in the test set, with score of model 1 below t_1 and score of model 2 greater or equal to t_2
$n^+_{t_1,\overline{t_2}}$ ($n^-_{t_1,\overline{t_2}}$)	number of positive (negative) instances, in the test set, with score of model 1 greater or equal to t_1 and score of model 2 below t_2
$\Delta TP^+_{t_1,t_2}$	r.v. for the difference in true positive rates when thresholds are set at t_1 and t_2 for models 1 and 2, resp
$\Delta FP^-_{t_1,t_2}$	same for false positive rates
relations:	$p^+_{t_1,\overline{t_2}} = n^+_{t_1,\overline{t_2}}/n^+$, $p^+_{\overline{t_1},t_2} = n^+_{\overline{t_1},t_2}/n^+$ $p^-_{t_1,\overline{t_2}} = n^-_{t_1,\overline{t_2}}/n^-$, $p^-_{\overline{t_1},t_2} = n^-_{\overline{t_1},t_2}/n^-$
Vertical averaging, single design	
T_r^-	r.v. for threshold at false positive rate r/n^-
s_k	kth largest negative instance score, in the test set
n_k^+	number of positive instances with score greater or equal to s_k, in test set
M_k^+	r.v. for the number of positive instances with score greater or equal to s_k, in a sample
TP_r^+	r.v. for the true positive rate at false positive rate r/n^-
relations:	$p_k^+ = n_k^+/n^+$
Vertical averaging, paired design	
$s_{1,k}$ ($s_{2,j}$)	kth (jth) largest score of negative instances in test set, according to model 1 (model 2)
$T^-_{1,r}$ ($T^-_{2,r}$)	r.v. for threshold at false positive rate r/n^- according to model 1 (model 2)
ΔTP_r^+	r.v. for difference in true positive rates, at false positive rate r/n^-

disjoint. Solutions obtained for single designs can therefore serve to obtain solutions for unpaired designs as well. On the other hand, with paired designs, we must account for the fact that out-of-sample performances of two models are usually positively correlated across test sets since model scores are correlated across instances: an obvious fraud will usually score high on both models. This paper considers single and paired designs.

2.2 Advantage of ROC curves

An advantage of ROC graph analysis lies in the fact that ROC curves are independent of the relative proportions of positive and negative instances in the population as well as the relative values of error costs (Fawcett 2006). Since both axes are scaled as proportions to the total number of positive (y-axis) and negative (x-axis) instances, a change in these numbers should not affect the ROC curve (although this argument has recently been discussed (Webb and Ting 2005; Fawcett and Flach 2005)). Since all computations of ROC curves are made independently of cost values, these have no effect on the curve. On the other hand, changes in costs or proportions will cause changes in the value of the optimal (expected total error cost minimizing) threshold, corresponding to a different point on the otherwise unaffected ROC curve. One drawback of ROC graph analysis, is its inability to handle instance-varying benefits (or costs) but an extension has recently been proposed for that purpose (Fawcett 2006). Model performance assessment in terms of expected total error cost can hardly be done using ROC curves and for this reason (and others (Drummond and Holte 2006)), *cost curves* (Drummond and Holte 2006) and expected performance curves (Bengio et al. 2005) have been introduced as alternatives (or complements) to ROC curves.

2.3 Bootstrap resampling of the test

In order to obtain confidence intervals for out-of-sample performance, resampling methods such as the bootstrap technique can be used. Since the objective here is to estimate a model's out-of-sample performance distribution, instances of the training set can not be used for that purpose and we are limited to resampling from the test set.

The bootstrap (Efron and Tibshirani 1993) is a simulation method that allows the estimation of the distribution of complex statistics such as ROC curves. Given an original set of instances, a certain number of samples are drawn, with replacement, from the original set. For each of the samples, an ROC curve is computed and the bootstrap estimate is obtained as the average, over all bootstrap samples, of the individual curves. In certain cases, analytic derivations can be obtained that give *exact bootstrap* distribution, allowing to characterize the statistic's bootstrap distribution without performing any resampling. An exact bootstrap distribution can be interpreted as the one to which the bootstrap distribution converges as the number of bootstrap samples tends to infinity.

Mathematical derivations of exact bootstrap results for ROC curves are made difficult by the presence of normalization terms that vary from one sample to another: at each threshold, the true (false) positive rate is obtained as the number of positive (negative) instances correctly (falsely) labelled as positive divided by the total number of positive (negative) instances in the sample. Since these total numbers vary from one bootstrap sample to another, one must account for their distribution. In particular, the bootstrap distribution assigns non zero probability to samples that are entirely composed of instances drawn from one of the two classes. In such cases, either the true or false positive rate is undefined at any threshold. In this paper, we circumvent this difficulty through the use of a procedure referred to as *stratified bootstrap* according to which proportions of positive and negative instances of each bootstrap sample are fixed as equal to those of the original test set. In other words, each sample is obtained from the combination of two independent bootstrap samples: one drawn from the set of positive instances and the other drawn from the set of negative instances. This procedure has previously been used in the context of ROC curves (Bandos 2005; Drummond and Holte 2006). The underlying assumption of fixed class proportions is discussed and revisited in Sect. 5.5.

2.4 Averaging techniques

As mentioned above, bootstrap estimates are obtained through the averaging of a series of ROC curves. This averaging can be performed according to various methods of which we consider two that have previously been used in the machine learning community (Fawcett 2004; Macskassy et al. 2005): threshold averaging and vertical averaging. The choice of the appropriate averaging method should fall directly from the intended use of the model to be implemented as each method can be associated to a specific variable that is considered independent. For this reason, we view threshold and vertical averaging as two different *problems* rather *solutions* when trying to establish the distribution of an ROC curve. According to threshold averaging, the independent variable is the score threshold that serves to determine whether an instance is to be labelled as a positive or a negative. The averaging process therefore involves looping through a series of different threshold values. For each threshold, true and false positive rates of the different ROC curves are averaged. Connecting these average points forms the average ROC curve. According to vertical averaging, a series of false positive rates are considered. For each of these false positive rates, true positive rates of the different ROC curves are averaged. Again, connecting these average points forms the average ROC curve. Vertical averaging is therefore associated to an independent false positive rate variable.

Vertical averaging has the advantage that, in ROC space, it leads to one-dimensional confidence intervals rather than two-dimensional confidence regions such as those obtained through threshold averaging. On the other hand, false positive rates can only be measured ex-post, after the model has been implemented and used, thus can not correspond to any ex-ante decision made by the user of the model. Obviously, threshold setting is a potential ex-ante user decision so that threshold averaging can be directly associated to practical applications, an advantage of this method over vertical averaging.

2.5 Approaches to obtain confidence intervals

Much research has been devoted to smoothing the empirical ROC curve and estimating its dispersion. Approaches fall into one of three broad categories: parametric, semi-parametric and nonparametric (or empirical). According to parametric approaches, scores of positive instances are assumed to follow a particular distribution and scores of negative instances are assumed to follow a distribution of the same functional form but with different parameters. The functional form is usually either normal or logistic. Model parameters are estimated using available data and the smoothed ROC curve follows as a combination of the two estimated functions.

Semi-parametric kernel-based methods have been proposed (Zou et al. 1997; Lloyds 1998) in order to smooth the empirical ROC curve and have been shown (Lloyds and Wong 1999) to perform better, in terms of root mean squared error, than the empirical ROC curve itself, if kernel bandwidths are chosen appropriately such as in Hall and Hyndman (2003). Using the kernels, a distribution for the ROC curve can be derived and used to obtain confidence intervals or regions, the performance of which is generally evaluated through an estimation of the coverage accuracy. Kernel bandwidth selection for optimal coverage accuracy has been investigated in Hall et al. (2004).

When obtained through the use of a model for scoring instances of a test set, the curve is often referred to as the *empirical* ROC. Strong convergence properties of the empirical ROC curve (Hsieh and Turnbull 1996) and a formula for its bias (Lloyds and Wong 1999) have been obtained. Although bootstrap resampling is often used to obtain empirical dispersion measures of ROC curves, exact (stratified) bootstrap results have only been reported for the area under the ROC curve (AUC) (Bandos 2005) and cost curves (Dugas and Gadoury 2008). In this paper, we derive pointwise exact stratified bootstrap distributions for ROC curves.

2.6 Confidence intervals for multinomial probabilities

As will be shown in Sect. 3, the true and false positive rates follow independent binomial distributions, when exact stratified bootstrap is used. We must therefore estimate confidence intervals for binomial probabilities, an issue that has received substantial attention in the statistical literature. The most common approach is to fit the estimated first two moments of the distribution to a Gaussian distribution. Accordingly, a $1-\alpha$ confidence interval has bounds

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the Gaussian distribution, n is the sample size from which $\hat{p}=k/n$ was obtained, and k is the number of observed successes. These intervals are sometimes referred to as *Wald* intervals since they are obtained from inverting the Wald large-sample normal test. Similarly, *score* confidence intervals are obtained by inverting a score normal test which uses the null hypothesis variance $p(1-p)/n$ rather than its approximated value. Score confidence interval bounds for a binomial probability are

$$\bigl(\hat{p} + z^2_{1-\alpha/2}/2n \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + (z_{1-\alpha/2}/2n)^2}\bigr)/(1+z^2_{1-\alpha/2}/n).$$

It is interesting to note (and can easily be shown) that score confidence intervals, as opposed to Wald intervals, will strictly provide admissible bounds i.e., limited to the [0,1] range. As a consequence, score intervals are asymmetric with respect to $\hat{p}$, the midpoint being $(\hat{p} + z^2_{1-\alpha/2}/2n)/(1+z^2_{1-\alpha/2}/n)$. Note that the estimated value $\hat{p}$ is always within the bounds. Another element of importance is the fact that score intervals never collapse to zero-length, i.e. the upper bound is always strictly larger than the lower bound. Zero-length Wald intervals are obtained whenever $\hat{p} \in \{0,1\}$. Thus the closer to 0 or 1 is the true underlying event probability p and the smaller the sample size is, the more frequent zero-length intervals will be obtained. This explains why Wald intervals exhibit such poor coverage for values of p close to 0 or 1 and small samples.

Exact confidence intervals for which coverage is always greater or equal to target $(1-\alpha)$ have been proposed (Clopper and Pearson 1934) but they generally lead to intervals that are too wide i.e., too conservative. To avoid confusion with respect to the use of the term “exact”, let us mention that, for the rest of the paper, “exact” shall refer to the fact that we derive exact bootstrap distributions. Exact confidence intervals are used nowhere here.

In Agresti and Coull (1998), the three alternatives described above are compared with respect to their coverage accuracy. As shown in Agresti and Coull (1998), coverage accuracy of Wald intervals drops drastically for small samples when the underlying true probability is close to 0 or 1. Score intervals coverage is shown to be closest to target. The authors then show that score intervals can be approximated by simply adding two positive instances as well as two negative instances to the sample before Wald-type intervals are derived. Bounds of the confidence intervals therefore become

$$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}} \tag{1}$$

where, $\tilde{p}=\tilde{k}/\tilde{n}$, $\tilde{k}=k+2$, and $\tilde{n}=n+4$. We refer to these as the Agresti-Coull intervals.

In Sects. 3 and 4, we show how Agresti-Coull confidence intervals and other similar intervals can be adapted to the four problems we are interested in. Through numerical simulations, Sect. 5 shows that these confidence intervals provide improvements over existing techniques, in terms of coverage accuracy.

3 Threshold averaging

3.1 Threshold averaging and single designs

Consider a test set consisting of n instances from which stratified bootstrap samples of size m are drawn (often, $m = n$). Let n^+ and n^- be the number of positive and negative instances, respectively, in the test set. Let $\mathcal{X}^+$ and $\mathcal{X}^-$ be the sets of all bootstrap samples that can be drawn from the n^+ positive and n^- negative instances of the test set, respectively. The set of all stratified bootstrap samples that can be drawn from the test set is $\{(x^+, x^-), x^+ \in \mathcal{X}^+, x^- \in \mathcal{X}^-\}$. Throughout the paper, values depending on actual sample $x^+(x^-)$ of positive (negative) instances are superscripted with symbol $+$ $(-)$.

According to the stratified bootstrap procedure, the number of sampled positive and negative instances, $m^+ = m \cdot n^+/n$ and $m^- = m \cdot n^-/n$ are fixed. In this paper, we simply assume $m^+, m^- \in \mathbb{N}$. Otherwise, in cases where $m \neq n$ and $m^+, m^- \notin \mathbb{N}$, an additional procedure should be devised so that stratified sampling precisely reflects the actual proportions of the test set.

Let n_t^+ denote the number of instances, among the n^+ positive instances of the test set, with score greater or equal to t. Let M_t^+ be the random variable for the number of positive instances with score greater or equal to t, within a random bootstrap sample of size m^+ drawn from the n^+ positive instances of the test set. Then, M_t^+ follows binomial distribution with trial number m^+ and estimated event probability $p_t^+ = n_t^+/n^+$ which we note as $M_t^+ \sim \mathrm{Bin}(m^+, p_t^+)$. Let $TP_t^+ = M_t^+/m^+$ be the random variable for the true positive rate, at threshold t, where m^+ is fixed across all samples.

Similarly for negative instances, the random variable for the false positive rate is denoted $FP_t^- = M_t^-/m^-$ where M_t^- is the random variable for the number of negative instances, within a sample of size m^- with score greater or equal to threshold t. Let $p_t^- = n_t^-/n^-$ where n_t^- is the number of instances with score greater or equal to t among the n^- negative instances of the test set. Thus, $M_t^- \sim \mathrm{Bin}(m^-, p_t^-)$.

According to stratified bootstrap resampling, samples $x^+ \in \mathcal{X}^+$ and $x^- \in \mathcal{X}^-$ are drawn independently so that TP_t^+ and FP_t^- are independent as well and their confidence intervals can be built independently. These confidence intervals can be obtained by fitting a Gaussian distribution to the first and second moments of the binomial distributions defined above. Since p_t^+ and p_t^- are sample proportions that are estimates of the true underlying probabilities, then careful attention must be paid in order to derive confidence intervals that closely match their target coverage. Our approach is to build two Agresti-Coull (1998) confidence intervals, one for the true positive rate and another for the false positive rate, and to combine these two intervals to obtain a two-dimensional confidence region. In Sect. 5.1, we perform simulations in order to assess the coverage accuracy of the suggested approach. This approach is compared to two others.

For a set of h thresholds, confidence intervals are obtained in time[2] $O(n \ln n)$ because of the necessary sorting preprocessing. If instances are already sorted, then the computational time is $O(n)$ (linear). Algorithm 1 describes how this can be done. Figure 1 gives an example of the use of Algorithm 1 where the confidence regions are illustrated for a set of five thresholds.

[2] Throughout the paper, computational time analyses assume n^+, n^-, m^+, m^-, and m are proportional to n.

Algorithm 1 Confidence regions for threshold averaging with single design. Time: $O(n \ln n)$

```
**Input:** Scores of positive and negative instances, set of $h$ thresholds, size $m$.
**Output:** Set of $h$ confidence regions for the model performance.
$n_t^+ \leftarrow 0, n_t^- \leftarrow 0$
$\beta \leftarrow 1 - \sqrt{1-\alpha}$
$z \leftarrow (1-\beta/2)$th quantile of the Gaussian distribution.
$a \leftarrow 2$
**for** $j = 1, 2, \ldots, h$ **do**
    $t \leftarrow j$th largest threshold.
    Compute $n_t^+$
    $\hat{p}_t^+ \leftarrow (n_t^+ + a)/(n^+ + 2a)$
    $\hat{\sigma} \leftarrow \sqrt{\hat{p}_t^+(1-\hat{p}_t^+)/(m^+ + 2a)}$
    $(L_j^+, U_j^+) \leftarrow \hat{p}_t^+ \pm z \cdot \hat{\sigma}$
    Similarly, obtain interval $(L_j^-, U_j^-)$ using scores of negative instances.
    Define $j$th confidence region as the rectangle with lower left corner $(L_j^-, L_j^+)$ and upper right corner $(U_j^-, U_j^+)$.
**end for**
```

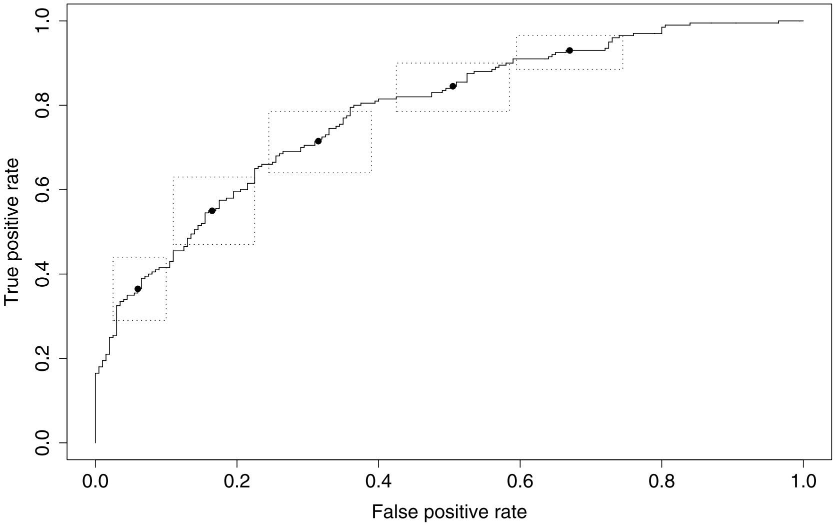

Fig. 1 Empirical ROC curve with pointwise exact bootstrap confidence intervals for a model performance (single design). Threshold averaging is used for five threshold values. All values were obtained using Algorithm 1

3.2 Threshold averaging and paired designs

When comparing the performance of two models, particular attention must be given to the choice of thresholds since the scores obtained using different models may bear different meanings: if a first model assigns scores between 0 and 100 while a second model assigns scores between 100 and 1000 then clearly, comparing their performance at fixed threshold values is meaningless. Even when score domains overlap, either partially or entirely, comparisons can be flawed if overall score distributions vary widely from one model to another.

In order to perform relevant comparisons between different model outputs, a first solution is to modify the scores through a process often referred to as score calibration (Platt 2000; Zadrozny and Elkan 2002; Fawcett and Niculescu-Mizil 2007) where the objective is to learn a monotonic mapping from the "raw" scores of a model to their corresponding calibrated scores representing class probability values, i.e., the conditional probability that an instance is positive.

Here, the purpose of calibration is merely to align scores of different models in order to perform relevant performance comparisons for thresholds that would, in practice, be comparable candidates for the two models at hand. The advantage of score calibration for this purpose is that it can be performed automatically, without the user's intervention. But this automation can also be viewed as a disadvantage since in some cases, the user might wish to compare models according to their respective performances at thresholds that are not necessarily aligned, even after calibration. The reason is as such: score calibration is usually performed with respect to score distribution, disregarding model performance, whereas threshold selection is based on model performance. This later approach is more flexible and likely more representative of practical implementations but requires additional user input.

In this paper, we handle both approaches by considering a set of h pairs of thresholds (t_1, t_2). If considering score calibration, then $t_1 = t_2$ for each of the h pairs, where t_1 and t_2 are the calibrated score thresholds of models 1 and 2, respectively. On the other hand, if thresholds are user-specified, then the h threshold pairs represent those for which the user wishes to compare the performance of the two models.

Given fixed threshold values t_1 and t_2, both true and false positive rates can vary from one model to another. It is therefore impossible to differentiate between two models when both true and false positive rates of one model are lower than the corresponding values for the other model, unless the costs of both types of errors (false positives and false negatives) and class probabilities (positive and negative) are known. We can still estimate $f_1(t_1, t_2)$, the probability that, at thresholds t_1 and t_2, model 1 *dominates* model 2, i.e. the probability that the false positive rate $FP^-_{t_1}$ of model 1 is less than or equal to that of model 2, $FP^-_{t_2}$, while model 1's true positive rate $TP^+_{t_1}$ is greater or equal to that of model 2, $TP^+_{t_2}$. The case where both true and false positive rates are equal must be excluded for strict dominance. Since samples x^+ and x^- are drawn independently, joint probabilities can be expressed as the product of marginals, thus

$$\begin{aligned} f_1(t_1, t_2) &= Pr\{\Delta TP^+_{t_1,t_2} \geq 0\} \cdot Pr\{\Delta FP^-_{t_1,t_2} \leq 0\} \\ &\quad - Pr\{\Delta TP^+_{t_1,t_2} = 0\} \cdot Pr\{\Delta FP^-_{t_1,t_2} = 0\}, \end{aligned} \tag{2}$$

where $\Delta TP^+_{t_1,t_2} = TP^+_{t_1} - TP^+_{t_2}$ and $\Delta FP^-_{t_1,t_2} = FP^-_{t_1} - FP^-_{t_2}$. Empirical estimates of $f_1(t_1, t_2)$ can be obtained with Algorithm 2 for which we provide here an intuitive description. Detailed mathematical derivations of the algorithm are left to Appendix A.

When paired designs are considered, one must account for the presence of dependencies between the scores of the two models, as explained in Sect. 1. Let us first consider positive instances. Differences in true positive rates are due to the presence of instances for which models disagree, i.e. given the thresholds considered, one model classifies an instance as positive while the other classifies the same instance as negative. Thus, estimates of the distribution of the difference in true positive rates are based on the number of test set instances in each of the following three categories: (a) instances for which score of model 1 is greater or equal to t_1 while score of model 2 is lower than t_2, (b) instances for which score of model 1 is lower than t_1 while score of model 2 is greater or equal to t_2, and (c) instances

Algorithm 2 Dominance probabilities for threshold averaging with paired design. Time: $O(n \cdot h)$

```
Input: Scores of positive and negative instances for both models, set of $h$ pairs of thresholds, size $m$.
Output: Set of $h$ probabilities of dominance.
Notation: Let $dtp1$, $dtp0$, $dfp1$, and $dfp0$ represent probabilities $Pr\{\Delta TP^+_{t_1,t_2} \geq 0\}$, $Pr\{\Delta TP^+_{t_1,t_2} = 0\}$, $Pr\{\Delta FP^-_{t_1,t_2} \geq 0\}$, and $Pr\{\Delta FP^-_{t_1,t_2} = 0\}$, respectively.

for $j = 1, 2, \dots, h$ do
    $(t_1, t_2) \leftarrow j$th pair of thresholds.
    Compute values $n^+_{t_1,\overline{t_2}}$, $n^+_{\overline{t_1},t_2}$, $n^-_{t_1,\overline{t_2}}$, and $n^-_{\overline{t_1},t_2}$.
    $u \leftarrow n^+_{t_1,\overline{t_2}}/n^+, \quad v \leftarrow n^+_{\overline{t_1},t_2}/n^+/(1-u)$
    $dtp0 \leftarrow 0, \quad dtp1 \leftarrow 1 - B(u, \lfloor m^+/2 \rfloor, m^+)$.
    for $i = 0, 1, \dots, \lfloor m^+/2 \rfloor$ do
        $dtp0 \leftarrow dtp0 + b(u, i, m^+) \cdot b(v, i, m^+ - i)$
        $dtp1 \leftarrow dtp1 + b(u, i, m^+) \cdot B(v, i, m^+ - i)$
    end for
    Values for $dfp1$ and $dfp0$ are obtain through similar computations for negative instances with $n^-$, $m^-$, $n^-_{\overline{t_1},t_2}$, and $n^-_{t_1,\overline{t_2}}$ in place of $n^+$, $m^+$, $n^+_{\overline{t_1},t_2}$, and $n^+_{t_1,\overline{t_2}}$, respectively.
    $f_1(t_1, t_2) \leftarrow dtp1 \cdot (1 - dfp1 + dfp0) - dtp0 \cdot dfp0$
    $f_2(t_1, t_2) \leftarrow (1 - dtp1 + dtp0) \cdot dfp1 - dtp0 \cdot dfp0$
end for
Return $f_1(t_1, t_2)$ and $f_2(t_1, t_2)$ for all pairs of thresholds
```

for which both models are in agreement, i.e. both scores are either greater or equal to or lower than their respective thresholds. Let $n^+_{\overline{t_1},t_2}$ and $n^+_{t_1,\overline{t_2}}$ represent the number of positive instances of categories (a) and (b), in the test set, respectively. For negative instances, we use a similar notation: $n^-_{\overline{t_1},t_2}$ and $n^-_{t_1,\overline{t_2}}$ represent the number of negative instances of categories (a) and (b), in the test set, respectively. When performing stratified bootstrap resampling, the joint distribution of the number of positive instances drawn from categories (a), (b) and (c) is trinomial (multinomial with three categories). The same is true of negative instances. This distributional property is used to derive Algorithm 2 to obtain dominance probability estimates $f_1(t_1, t_2)$ and $f_2(t_1, t_2)$.

Note that, unfortunately, evaluating the statistical significance of superior performance of one model over the other using (2) and Algorithm 2 leads to a test with very poor power. The main reason is that in many cases, probabilities $Pr\{\Delta TP^+_{t_1,t_2} < 0, \Delta FP^-_{t_1,t_2} < 0\}$ and $Pr\{\Delta TP^+_{t_1,t_2} > 0, \Delta FP^-_{t_1,t_2} > 0\}$ are sufficiently large so that the tests are inconclusive by usual standards, i.e. with p-values below 1, 5 or 10%. In other words, it is difficult for $f_1(t_1, t_2)$ to reach values above 0.99, 0.95 or even 0.90. This argument is illustrated in Fig. 2.

Let us now consider confidence intervals for $\Delta TP^+_{t_1,t_2}$ and $\Delta FP^-_{t_1,t_2}$, thus defining confidence regions in ROC space, for the difference in performance between two candidate models. The issue of building confidence intervals for the difference between two event probabilities, within the context of paired designs, has previously been addressed in the statistical literature. In particular, Agresti and Min (2005) suggest a procedure, similar to the one used in the previous subsection, whereby one half instance is added to each of cate-

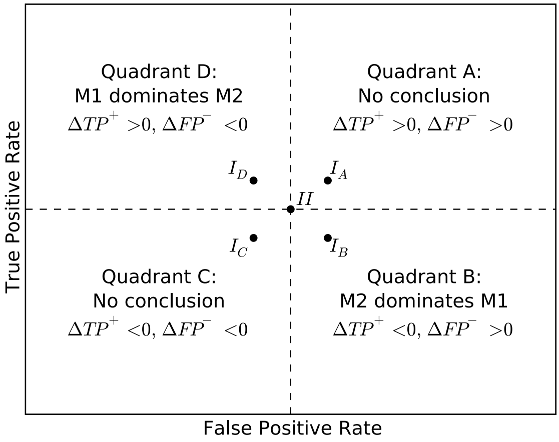

Fig. 2 Testing for dominance of one model over the other using (2). Point II represents, in ROC space, the performance of model 2 (M2) at a certain threshold t_2 and for a certain sample. The performance of model 1 (M1), at the corresponding threshold t_1 and for the same sample, can be located in any of *the four quadrants* (A, B, C, and D) delimited by the *two dashed lines* that cross each other at point II. If model 1's performance is located in *quadrant A* (e.g., point I_A) or in *quadrant C* (e.g., point I_C), then no conclusion can be reached: in *quadrant A*, a higher true positive rate is obtained at a cost of a higher false positive rate whereas in *quadrant C*, a lower false positive rate is obtained at a cost of a lower true positive rate. This leads to a dilemma that is not solved in the context of ROC analysis and no conclusion can be reached. If model 1's performance is located in one the other two quadrants, then we conclude that either model 1 (in *quadrant D*, e.g. point I_D) or model 2 (in *quadrant B*, e.g. point I_B) dominates

gories (a) and (b) and one full instance is added to category (c) (i.e. a total of two instances are added), before computing Wald-type confidence intervals.

Given threshold t_1 and t_2, computing values $n^{+}_{\overline{t_1},t_2}$ and $n^{+}_{t_1,\overline{t_2}}$ requires linear time. The same is true for the corresponding values for negative instances. For a set of h threshold pairs, confidence regions are thus obtained in time $O(n \cdot h)$. Algorithm 3 shows how these confidence regions are obtained. In Sect. 5.2, the proposed approach is compared with two others.

4 Vertical averaging

4.1 Vertical averaging and single designs

We wish to evaluate the exact bootstrap distribution of the true positive rate for each element of the set $\{r_i/m^-, 1 \leq r_1 \leq r_2 \leq \cdots \leq r_h \leq m^- - 1, i = 1, 2, \ldots, h\}$ of h false positive rates of interest. Let TP^+_r be the random variable for the true positive rate, at false positive rate r/m^-. Also, let M^+_r be the random variable for the number of positive instances correctly labelled as positives, at the same false positive rate r/m^- so that $TP^+_r = M^+_r/n^+$. Note that M^+_r depends on false positive rate r/m^- which in turn varies across samples $x^- \in \mathcal{X}^-$, so that M^+_r depends on both samples x^+ and x^- of positive and negative instances, respectively.

The following mathematical derivation can be summarized as such: given stratified sample (x^+, x^-) and fixed false positive rate r/m^-, r negative instances of the sample, those with the highest scores, are falsely labelled as positives. The lowest score among these false positives is the threshold T^-_r associated to false positive rate r/m^-. Once that threshold is

Algorithm 3 Confidence regions for threshold averaging with paired design. Time: $O(n \cdot h)$

```
**Input:** Scores of positive and negative instances for both models, set of $h$ pairs of thresholds, size $m$.
**Output:** Set of $h$ confidence regions for the difference in performance.
$\beta \leftarrow 1 - \sqrt{1-\alpha}$
$z \leftarrow (1-\beta/2)$th quantile of the Gaussian distribution.
$a \leftarrow 0.5$
**for** $j = 1, 2, \ldots, h$ **do**
    $(t_1, t_2) \leftarrow j$th pair of thresholds.
    Obtain values $n^+_{t_1,\overline{t_2}}$, $n^+_{\overline{t_1},t_2}$, $n^-_{t_1,\overline{t_2}}$, and $n^-_{\overline{t_1},t_2}$.
    $\hat{p}_{t_1,\overline{t_2}} \leftarrow (n^+_{t_1,\overline{t_2}} + a)/(n^+ + 4a)$
    $\hat{p}_{\overline{t_1},t_2} \leftarrow (n^+_{\overline{t_1},t_2} + a)/(n^+ + 4a)$
    $\hat{\sigma} \leftarrow \sqrt{\frac{\hat{p}_{t_1,\overline{t_2}} + \hat{p}_{\overline{t_1},t_2} - (\hat{p}_{t_1,\overline{t_2}} - \hat{p}_{\overline{t_1},t_2})^2}{m^+ + 4a}}$
    $(L^+_j, U^+_j) \leftarrow \hat{p}_{t_1,\overline{t_2}} - \hat{p}_{\overline{t_1},t_2} \pm z \cdot \hat{\sigma}$
    Similarly, obtain interval $(L^-_j, U^-_j)$ using scores of negative instances.
    Define $j$th confidence region as the rectangle with lower left corner $(L^-_j, L^+_j)$ and upper right corner $(U^-_j, U^+_j)$.
**end for**
```

determined, we can establish the distribution of the true positive rate over $\mathcal{X}^+$, *conditional* on the value of T^-_r. Integrating over the distribution of T^-_r, we obtain the *unconditional* distribution TP^+_r.

First, let $s_1 \geq s_2 \geq \cdots \geq s_{n^-}$ be the scores of the negative instances, sorted in decreasing order. Given a particular sample (x^+, x^-), if at least r of the m^- negative instances of sample x^- are chosen from the k negative instances with the largest scores, then the corresponding threshold T^-_r is greater or equal to s_k. This can be expressed as a sum of binomial probabilities:

$$\begin{aligned} Pr\{T^-_r \geq s_k\} &= \sum_{j=r}^{m^-} b(k/n^-, j, m^-) \\ &= 1 - B(k/n^-, r-1, m^-) \end{aligned} \tag{3}$$

where $b(p, k, n)$ is the binomial probability of obtaining k successes out of n independent trials, each with success probability p. The associated cumulative distribution is denoted $B(p, k, n)$. The exact bootstrap distribution of the threshold T^-_r follows as

$$\begin{aligned} Pr\{T^-_r = s_k\} &= Pr\{T^-_r \geq s_k\} - Pr\{T^-_r \geq s_{k-1}\} \\ &= B\left(\frac{k-1}{n^-}, r-1, m^-\right) - B(k/n^-, r-1, m^-). \end{aligned} \tag{4}$$

Each value of $P\{T^-_r = s_k\}$ is obtained in constant time.

Let us now consider the threshold conditional exact bootstrap distribution of the false positive rate. Let n^+_k be the number of positive instances with score greater or equal to s_k, out of the n^+ positive instances of the test set. Let M^+_k be the random variable for the number of instances with score above s_k, out of the m^+ positive instances of stratified bootstrap sample x^+. Then, M_k^+ follows binomial distribution $M_k^+ \sim \mathrm{Bin}(m^+, p_k^+)$ where $p_k^+ = n_k^+/n^+$. Finally, the unconditional true positive rate distribution is

$$\begin{aligned} Pr\{TP_r^+ = l/m^+\} &= \sum_{k=1}^{n^-} Pr\{T_r^- = s_k\} \cdot Pr\{M_k^+ = l\} \\ &= \sum_{k=1}^{n^-} Pr\{T_r^- = s_k\} \cdot b(l, m^+, p_k^+). \end{aligned} \tag{5}$$

Each probability $Pr\{TP_r^+ = l/m^+\}$ is computed in linear time and this is done for all values of l in $\{0, 1, \dots, m^+\}$ and all h false positive rates. Thus, using (4) and (5), we obtain true positive rate distributions for all h false positive rate values in time $O(n^2 \cdot h)$. Note that the true positive rate distribution is a mixture of binomial distributions, not a binomial itself, except for trivial cases.

Here again, exact bootstrap distributions can be summarized by their first two moments and fitted to Gaussian distributions in order to perform computations more efficiently and derive Wald-type confidence intervals. In order to avoid coverage accuracy breaks, we use the Agresti-Coull (1998) smoothed estimator for $p_k^+ : \hat{p}_k^+ = (n_k^+ + 2)/(n^+ + 4)$ so that, at false positive rate r/m^-, we have:

$$\begin{aligned} \hat{E}\{TP_r^+\} &= \sum_{k=1}^{n^-} Pr\{T_r^- = s_k\} \hat{E}\{M_k^+/m^+\} \\ &= \sum_{k=1}^{n^-} Pr\{T_r^- = s_k\} \cdot \hat{p}_k^+. \end{aligned} \tag{6}$$

Similarly for the second moment of TP_r^+:

$$\begin{aligned} \hat{E}\{(TP_r^+)^2\} &= \sum_{k=1}^{n^-} Pr\{T_r^- = s_k\} \hat{E}\{(M_k^+/m^+)^2\} \\ &= \sum_{k=1}^{n^-} Pr\{T_r^- = s_k\} [(\hat{p}_k^+)^2 + \hat{p}_k^+(1 - \hat{p}_k^+)/(m^+ + 4)]. \end{aligned} \tag{7}$$

Equations (6) and (7) are obtained in linear time and computed for each false positive rate so that pointwise confidence intervals are obtained in time $O(n \cdot h)$. Algorithm 4 shows how these computations are performed. Numerical results comparing the suggested approach with others are presented in Sect. 5.3.

4.2 Vertical averaging and paired designs

We conclude this section by considering the distribution of the difference between two ROC curves within the context of vertical averaging and paired designs. Here again, we provide an intuitive description and detailed mathematical derivations are left to Appendix B.

Algorithm 4 Confidence intervals for vertical averaging with single design. Time: $O(n \cdot h)$

```
Input: Scores of positive and negative instances, set of $h$ false positive rates, size $m$.
Output: Set of $h$ confidence intervals for the true positive rate.
Notation: Let $ei$, $ei2$, and $pik$ represent values $\hat{E}\{TP^+_{r_i}\}$, $\hat{E}\{(TP^+_{r_i})^2\}$, and $Pr\{T^-_{r_i} = s_k\}$, respectively.
$a \leftarrow 2$
for $k = 1, 2, \dots, n^-$ do
    Compute $n^+_k$
    $\hat{p}^+_k \leftarrow (n^+_k + a)/(n^+ + 2a)$
    $ek \leftarrow \hat{p}^+_k$
    $ek2 \leftarrow (\hat{p}^+_k)^2 + \hat{p}^+_k(1 - \hat{p}^+_k)/(m^+ + 2a)$
end for
$z \leftarrow (1 - \alpha/2)$th quantile of the standard Gaussian distribution.
for $i = 1, 2, \dots, h$ do
    $ei \leftarrow 0$
    $ei2 \leftarrow 0$
    for $k = 1, 2, \dots, n^-$ do
        $pik \leftarrow B(\frac{k-1}{n^-}, r_i - 1, m^-) - B(k/n^-, r_i - 1, m^-)$
        $ei \leftarrow ei + pik \cdot ek$
        $ei2 \leftarrow ei2 + pik \cdot ek2$
    end for
    $vi \leftarrow ei2 - (ei)^2$
    $(L^+_i, U^+_i) \leftarrow ei \pm z\sqrt{vi}$
end for
```

Let $T^-_{1,r}$ and $T^-_{2,r}$ be the random variables for the thresholds when false positive rate is r/m^-, for the first and second models, respectively. Let $s_{1,1} \geq s_{1,2} \geq \cdots \geq s_{1,n^-}$ and $s_{2,1} \geq s_{2,2} \geq \cdots \geq s_{2,q}$ be the scores of the negative instances, sorted in decreasing order, using the first and second models, respectively. Suppose $T^-_{1,r} = s_{1,k}$ and $T^-_{2,r} = s_{2,j}$. Let us first emphasize the fact that k is not necessarily equal to j and a simple numerical example will illustrate this situation: suppose there are $n^- = 4$ negative instances. According to the first model, their scores are $\{1, 2, 3, 4\}$ and according to the second model, scores are $\{5, 8, 6, 7\}$. We draw samples of size $m^- = 4$ and fix the false positive rate at $r/m^- = 25\%$, i.e. $r = 1$. First, consider the case where we draw each instance exactly once. According to the first model, the threshold is set equal to $T^-_{1,r} = s_{1,1} = 4$, the score of the fourth instance and the largest. The second model sets threshold at $T^-_{2,r} = s_{2,1} = 8$, the score of the second instance, also the largest. Thus, here we obtain thresholds such that $k = j = 1$. Now, let us consider a sample for which we draw the first instance twice, the second instance once and the third instance once as well. The first model sets the threshold equal to $T^-_{1,r} = s_{1,2} = 3$, the score of the third instance and second largest. The second model sets the threshold at $T^-_{2,r} = s_{2,1} = 8$, the score of the second instance and largest. Thus, in that case, $k = 2$ and $j = 1$ so that $k \neq j$. The joint distribution of $T^-_{1,r}$ and $T^-_{2,r}$ has $O(n^2)$ support.

The distribution of ΔTP^+_r is obtained in three steps: first, each probability density function value

$$f_r(k, j) = Pr\{T^-_{1,r} = s_{1,k}, T^-_{2,r} = s_{2,j}\}$$

of the joint thresholds distribution is obtained in linear time. Since $f_r(k, j)$ is computed for all values of r, k and j, total computational time is $O(n^4)$. Second, given thresholds $T^-_{1,r}$ and

Algorithm 5 Confidence intervals for vertical averaging with paired design. Time: $O(n^4)$

```
Input: Scores of positive and negative instances for both models, set of $h$ false positive rates, size $m$.
Output: Set of $h$ confidence intervals for the difference between true positive rates.
for $r = m^-, m^- - 1, \ldots, 1$ do
  for $k = 1, 2, \ldots, n^-$ do
    for $j = 1, 2, \ldots, n^-$ do
      Compute joint threshold probability density function $f_r(k, j)$ using (15), (16), (17), (18), and (19)
    end for
  end for
end for
for $k = 1, 2, \ldots, n^-$ do
  for $j = 1, 2, \ldots, n^-$ do
    for $d = -1, (m^+ - 1)/m^+, \ldots, -1/m^+, 0, 1/m^+, \ldots, (m^+ - 1)/m^+, 1$ do
      Compute thresholds conditional probability density function $g_{k,j}(d)$ using (22)
    end for
  end for
end for
for $r = 1, 2, \ldots, m^-$ do
  Using (23), compute probability density function $h_r(d)$ and obtain confidence interval for $d$, at false positive rate $r/m^-$.
end for
```

$T^-_{2,r}$, each conditional probability distribution function value for ΔTP^+_r

$$g_{k,j}(d) = Pr\{\Delta TP^+_r = d/m^+ | T^-_{1,r} = s_k, T^-_{2,r} = s_j\}$$

is obtained in linear time. Since $g_{k,j}(d)$ is computed for values of k, j and d, total computational time is $O(n^4)$. Finally, each value of the unconditional distribution of ΔTP^+_r

$$h_r(d) = \sum_{k,j} f_r(k, j) \cdot g_{k,j}(d)$$

is obtained in quadratic time. Since $h_r(d)$ is computed for all values of r and d, total computational time is, again, $O(n^4)$. Algorithm 5 summarizes these steps and relevant equations of Appendix B are identified.

As detailed in Appendix B, the conditional distributions $g_{k,j}(d)$ can be approximated with conditional moments. Unconditional moments are then obtained using the joint threshold distribution and the usual Gaussian approximation can be used to obtain confidence intervals. This procedure, although faster, still requires $O(n^4)$ computational time since the threshold distribution still needs to be obtained.

Of course, an $O(n^4)$ algorithm has limited practical applicability. In order to improve on the algorithmic efficiency, one would be required to approximate the joint threshold distribution. Also, we have observed that most entries of the matrix of the joint probability mass function have values very close to 0. Sparsifying this matrix would certainly help accelerate computations. We leave these avenues for future work.

In Sect. 5.4, we compare the confidence intervals obtained with this approach to those obtained by effectively drawing a series of bootstrap samples from the test set.

5 Numerical results

In this section, we conduct a series of experiments in order to assess the performance of the equations and algorithms described in Sects. 3 and 4. For each of the four potential combinations of averaging (threshold or vertical) and design (single or paired), we compare the proposed solutions of Table 1 with other popular methods. Performance is measured in terms of coverage accuracy of the confidence intervals derived using these solutions.

Simulation experiments are inspired from Macskassy et al. (2005) which reported results in terms of coverage accuracy as a function of false positive rate and can thus be compared to the results we present in Sect. 5.3. An exception are their results on threshold averaging, reported in terms of coverage accuracy as a function of threshold which are similar to those of Sect. 5.1. In addition, experiments with six real-world data sets, available from the UCI repository, are also presented. More details on these data sets appear in Appendix C.

Note that, for threshold averaging results of Sects. 5.1 and 5.2, we choose to report coverage accuracy as a function of *total positive rate* which is the number of test instances labelled as positive divided by the total number of instances in the test set. Using the total positive rate, rather than threshold, allows us to perform comparisons that are independent of score scales.

5.1 Threshold averaging and single designs

In this subsection, we compare the coverage accuracies of three methods used to derive confidence regions of threshold averaged ROC curves. According to the first method, the pointwise exact bootstrap distributions of both the observed true and false positive rates are smoothed using Agresti-Coull estimators. This method was described in detail in Sect. 3.1 by Algorithm 1. We refer to this method as the *Agresti* method. The second method consists of deriving Wald-type confidence intervals from the exact bootstrap distributions of the true and false positive rates. This second method is referred to as the *Wald* method. Finally, the third method consists of actually performing a certain number (here, 100) of bootstrap resamples of the test set. For each resample, an ROC curve is obtained. Quantiles of the empirical distribution serve to define confidence intervals and regions. This last method is referred to as the *Empirical* method.

As a first experiment, we reproduce the one that appears in Macskassy et al. (2005) in which positive and negative instances scores both follow normal distributions but with different parameters. Such a pair of normal distributions is often referred to as a binormal distribution. In Macskassy et al. (2005), the scale parameter is set to 3.75 for positive instances and 3.00 for negative instances and confidence intervals are obtained for a significance level of 10%. The location parameter θ for positive instances varies within the set $\{0.75, 1.5, 3.0, 5.0\}$ and the location parameter for negative instances is set equal to $-\theta$. Sample size is set to 10,000, i.e. a set of 10,000 instances is drawn from the positive instances distribution and another set of 10,000 negative instances is drawn from the negative instances distribution. The sampling procedure is repeated 1,000 times, i.e. 1,000 simulations are performed for each value of θ. We refer to this experiment as the *spread* experiment.

Figure 3 provides simulation results for the spread experiment. First, note that asymmetry of the plots is due to the difference in scale parameters used for positive and negative instances distributions. With a spread parameter of $\theta = 0.75$, the three methods are barely distinguishable. With a parameter of either $\theta = 1.50$ or $\theta = 3.00$, coverages of the Wald and Empirical methods drop sharply for low values of the total positive rate. The Agresti method

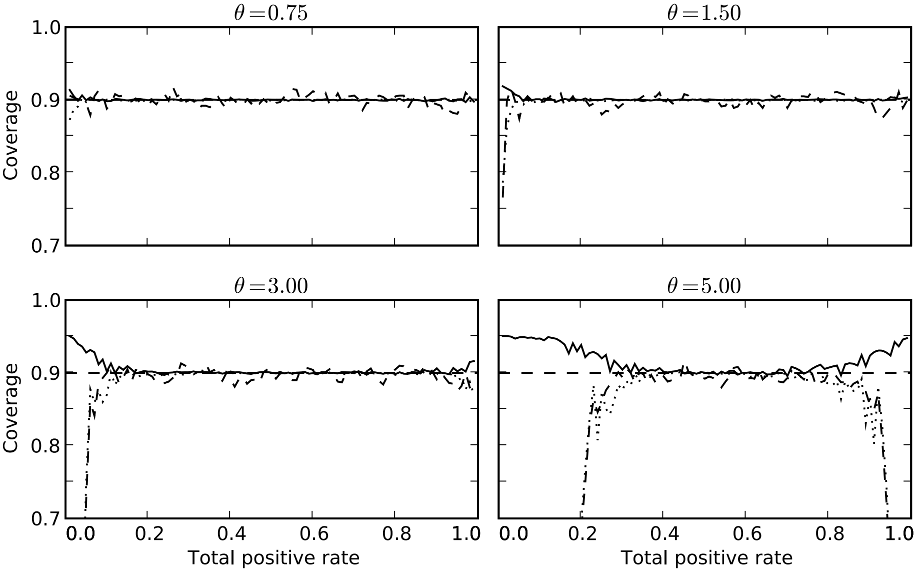

Fig. 3 Threshold averaging and single design. Effect of spread between distributions on coverage. Four different values for the location parameter θ are considered. The coverage accuracies of three methods: Agresti (*solid*), Wald (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 10,000 and target coverage is 90% (*dashed*)

is somewhat too conservative. With a parameter of $\theta = 5.00$, the pattern is emphasized on the left-hand side of the plot and appears on the right-hand side as well. Thus, better accuracy is obtained when score distributions of positive and negative instances have strong overlap, i.e. for low values of θ. This suggests that coverage accuracy is better for "difficult" problems, an observation that may seem counterintuitive. Digging into a simple numerical example will help clarify this point.

Consider the bottom right plot of Fig. 3, i.e. with spread parameter $\theta = 5.0$, and a total positive rate of 0.2. The target theoretical values for the true and false positive rates are 0.3999 and 1.3090e–04, respectively. Threshold is set at 5.9513. In other words, the probability that a negative instance scores above 5.9513, and thus becomes a false positive, is 1.3090e–04. With sample size of 10,000, the probability of obtaining no such false positive is $(1 - 1.3090\text{e–}04)^{10{,}000} = 0.2701$. In such cases, the Wald confidence interval for the false positive rate, is [0,0]. Then, the theoretical false positive rate is excluded from the confidence interval and there can be no coverage, i.e. the false positive rate can only be covered at most 73% of the time. Assuming coverage of the true positive rate is close to its target of $1 - \sqrt{1 - 0.1} = 0.9487$, then simultaneous coverage of both true and false positive rates should be close to the product of these two figures, i.e. 0.692. Indeed, we observe a coverage of 0.692 for the Wald method.

With $\theta = 0.75$, as in the top left plot of Fig. 3, and total positive rate of 0.2, target theoretical values for the true and false positive rates are 0.2861 and 0.1139, respectively. Threshold is set at 2.8681. Here, the probability that no false positive is observed is zero and this explains why coverage is much better for the Wald method.

The Empirical method exhibits similar patterns. On the other hand, the Agresti method provides confidence intervals that are too conservative. When no false positives are ob-

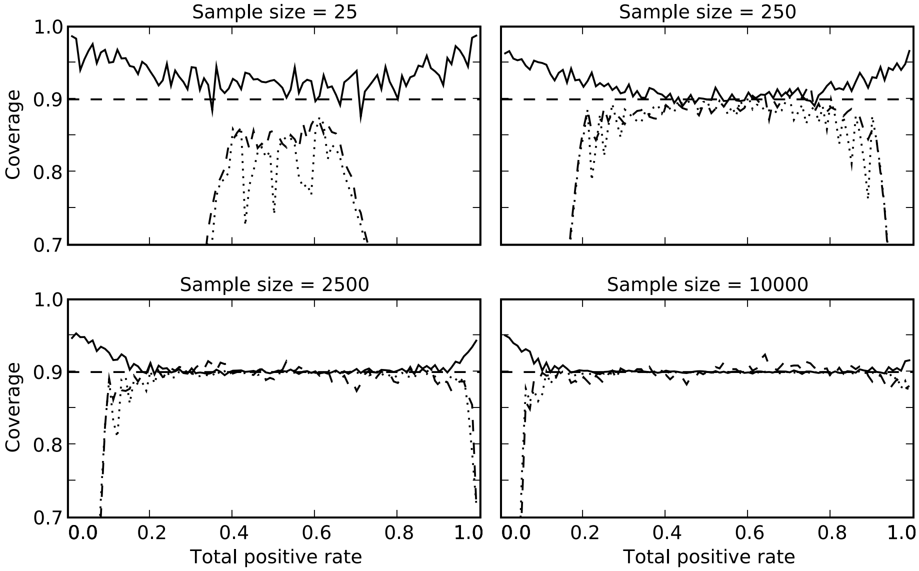

Fig. 4 Threshold averaging and single design. Effect of sample size on coverage. Four sizes are considered. The coverage accuracies of three methods: Agresti (*solid*), Wald (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Location parameter is $\theta = 3.0$ and target coverage is 90% (*dashed*)

served, the confidence interval becomes [0, 4.7539e–04], thereby including the target theoretical value of 1.3090e–04 (with $\theta = 5.0$ and total positive rate of 0.2). Thus, whereas coverage of the Wald and Empirical methods tend to break, the Agresti method becomes somewhat too conservative.

Note that our interpretation that the results are better for "difficult" problems is related to how we compare the different plots. Here, we have chosen to compare performances on the basis of equal total positive rates. This forces us to look at different threshold ranges. With $\theta = 0.75$, thresholds span the $[-7.8430, 8.5299]$ range whereas with $\theta = 5$, this range widens to $[-11.1618, 12.7016]$. The conclusion that results are better for difficult problems would certainly differ, had we compared plots on the basis of equal threshold values.

In a second experiment, we consider the effect of sample size on coverage accuracy. This experiment, as the previous one, appears in Macskassy et al. (2005) and is everywhere similar to the previous (spread) experiment except for two modifications: (1) the location parameter no longer varies: it is set to $\theta = 3.0$ and (2) the sample size takes values in $\{25, 250, 2\,500, 10{,}000\}$ instead of being fixed at 10,000. This second experiment will be referred to as the *size* experiment. Simulation results appear in Fig. 4. Clearly, results improve as the sample size increases, for all three methods considered. Once again in this experiment, severe coverage breaks affect the Wald and Empirical methods for low and high values of the total positive rate. In these regions, the Agresti method leads to coverage that is too conservative.

Finally, as a third experiment, we consider "real" data sets, available through University of California at Irvine's repository (Asuncion and Newman 2007). Six data sets are used and their detailed descriptions appear in Appendix C. We refer to them as Abalone, Adult, Chess, Covertype, Credit, and Telescope. Using training sets of lengths 1000 (Abalone),

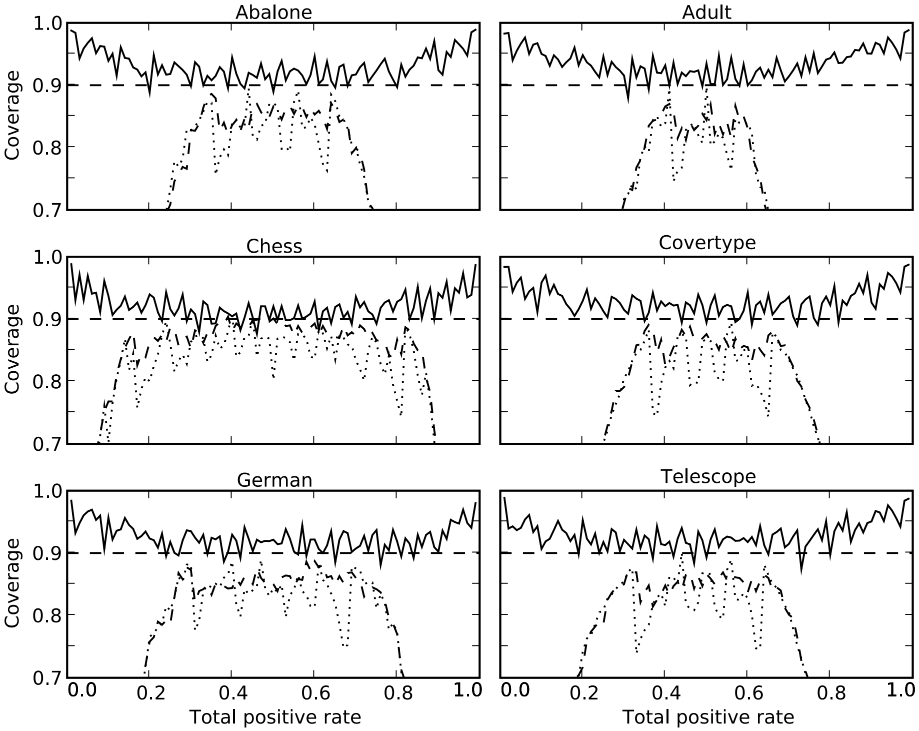

Fig. 5 Threshold averaging and single design. Coverage accuracy for six real data sets. The coverage accuracies of three methods: Agresti (*solid*), Wald (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 25 and target coverage is 90% (*dashed*)

32,561 (Adult), 5000 (Chess), 500 (Credit), and 1000 (Telescope), we trained a logistic regression model for each of the six data sets. From the set of remaining observations, we derived the target ROC curve. Then, for each simulation, a test set was obtained by sampling with replacement from the same set of remaining observations. Note that, when using real data, score distributions spread and shapes are determined by the data itself, not under our control. Thus, the only parameter that can fluctuate is size. In Figs. 5 and 6, we consider sample sizes of 25 and 250, respectively. Results are similar to those obtained previously. We refer to this as the *UCI* experiment.

In conclusion, experiments of this subsection show that, within the context of threshold averaging and single design, the Wald and Empirical methods may lead to severe coverage breaks. These breaks tend to occur in the ends of the ROC curve, i.e. where the total positive rate is either low or high but as the sample size decreases or as the score distributions of positive and negative instances are further apart (a larger spread), coverage breaks affect a larger portion of the ROC curve. In these situations, the Agresti method avoids coverage breaks that poise the other two methods but its coverage is above target.

5.2 Threshold averaging and paired designs

In this subsection, we report coverage accuracy results for confidence intervals for the difference of two ROC curves, at fixed threshold values. Given desired significance level α and

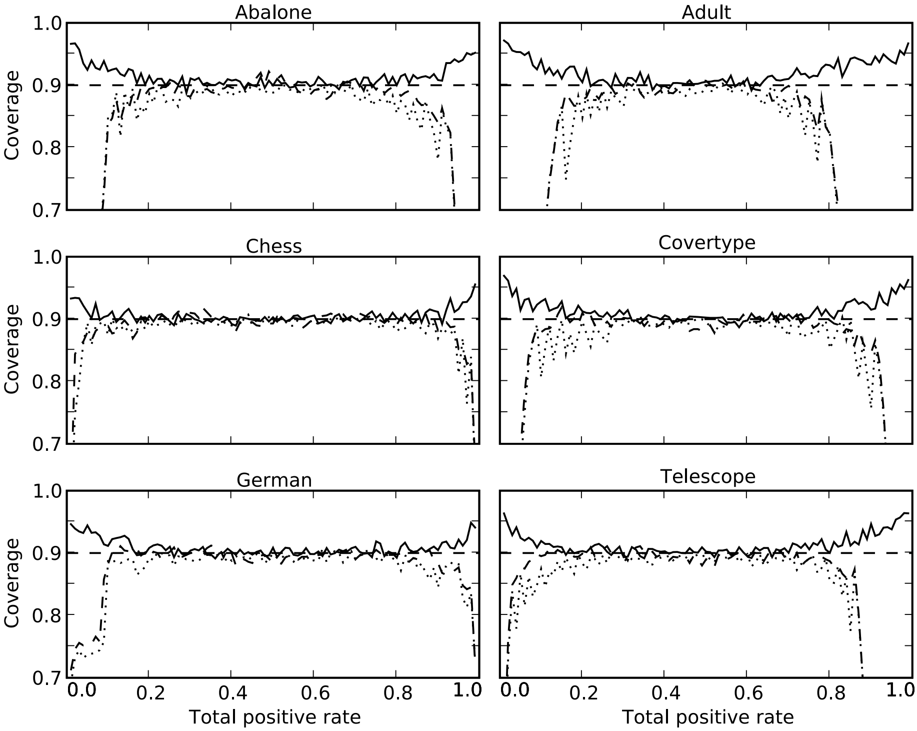

Fig. 6 Threshold averaging and single design. Coverage accuracy for six real data sets. The coverage accuracies of three methods: Agresti (*solid*), Wald (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 250 and target coverage is 90% (*dashed*)

since $\Delta TP^{+}_{t_1,t_2}$ and $\Delta FP^{-}_{t_1,t_2}$ are independent, a confidence interval of size $\sqrt{1-\alpha}$ is defined for each variable. The intersection of these two confidence intervals is a two-dimensional rectangular confidence region of desired size $1-\alpha$.

We compare three methods derived according to the same three approaches considered in the previous subsection. The *Agresti* method, as applied to the case of threshold averaging and paired design, is described in Algorithm 3. *Wald* intervals are easily obtained using the same algorithm but by setting $a \leftarrow 0$. Third, according to the *Empirical* method, 100 bootstrap samples are drawn, according to distributions we now define.

The experiment design is similar to the one used for the spread and size experiments of the previous subsection. Scores are distributed according to a binormal distribution with scale parameter set to 3.75 for positive instances and 3.00 for negative instances. Confidence intervals are obtained for a significance level of $\alpha = 10\%$. The location parameters are set as follows: for positive instances of the first model, we consider two values: $\theta \in \{1.0, 3.0\}$. For negative instances of both models the location parameter is set equal to $-\theta$. For positive instances of the second model it is set to $\theta + 2.0$. In order to include some form of dependency between the scores of the two models, three values of a correlation factor are considered: $\rho \in \{0.3, 0.6, 0.9\}$.

Simulation results appear in Fig. 7. As was the case with single design, coverage accuracy declines as the spread (θ) parameter increases. The same is true of the correlation

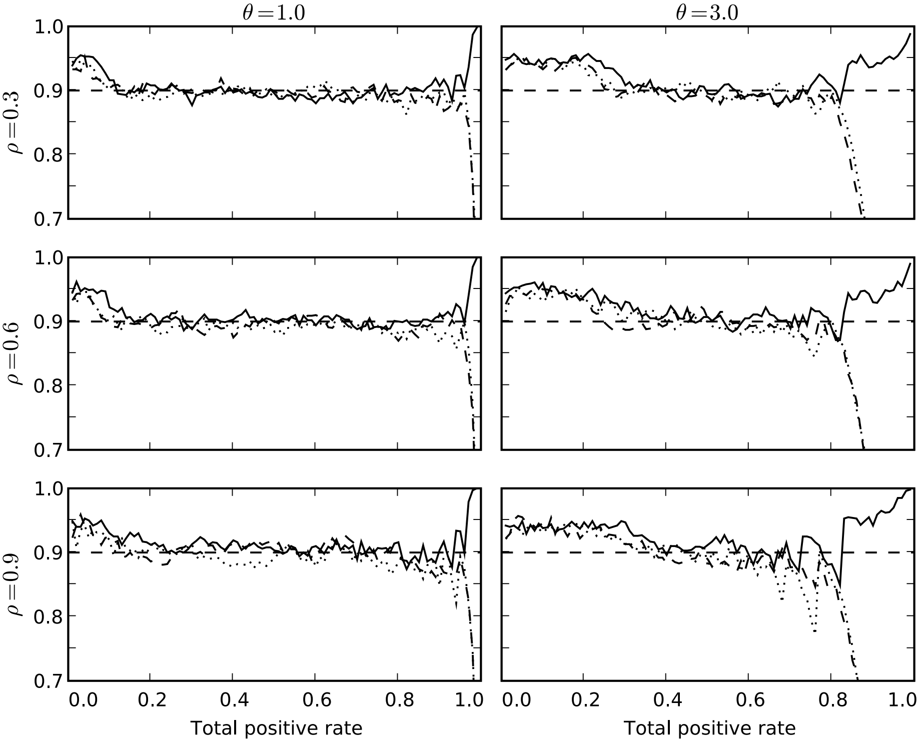

Fig. 7 Threshold averaging and paired design. Location parameter for positive instances of first model is set to $\theta = 1.0$ (*left*) and $\theta = 3.0$ (*right*). Correlation factor is equal to 0.3 (*top*), 0.6 (*middle*) and 0.9 (*bottom*). The coverage accuracies of three methods: Agresti (*solid*), Wald (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 100 and target coverage is 90% (*dashed*)

parameter: as ρ increases, zigzag patterns become more pronounced. Once again, in cases where coverages of the Wald and Empirical drop, coverage of the Agresti method increases above target.

We conclude, as we did in the previous subsection, that in situations where the coverages of the Wald and Empirical methods drop, the Agresti method leads to conservative confidence intervals with coverage above target.

5.3 Vertical averaging and single designs

Out of the four problems we consider in this paper, the case of vertical averaging with single design clearly stands out as the one that has received the most attention in the literature. In particular, semi-parametric approaches involving kernel-based methods have drawn much attention in the statistical literature (e.g. Hall and Hyndman 2003; Hall et al. 2004; Lloyds and Wong 1999) where the main issue is the selection of the appropriate value for the bandwidth parameter.

Again in this subsection, we compare three methods. The first method has been detailed in Sect. 4.1 by Algorithm 4. As this method uses an Agresti-type adjustment of pointwise

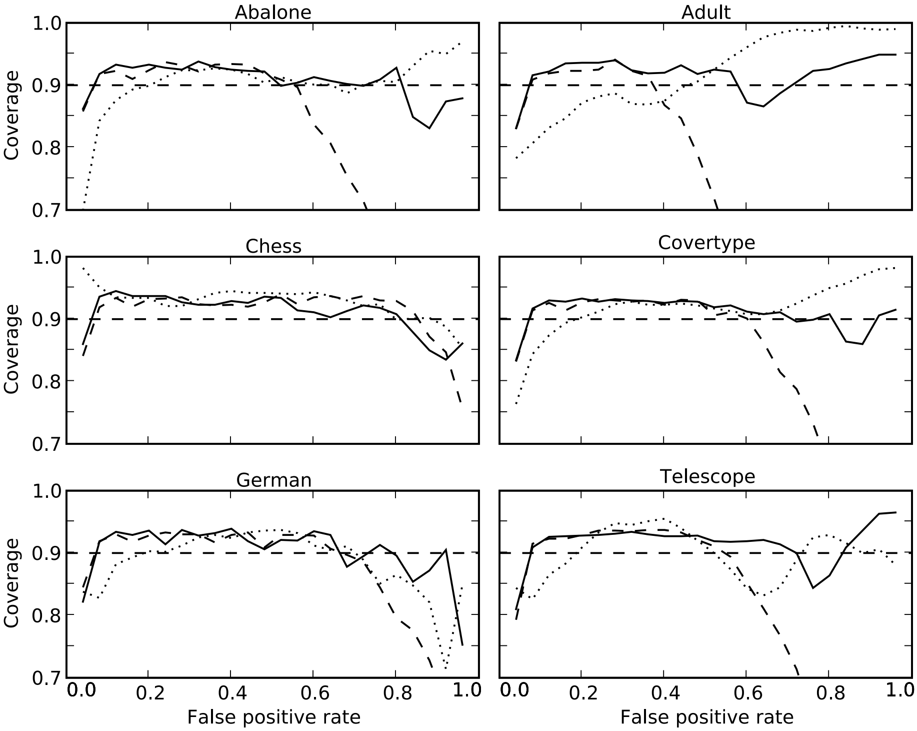

Fig. 8 Vertical averaging and single design. Coverage accuracy for six real data sets. The coverage accuracies of three methods: Agresti (*solid*), Kernel (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 25 and target coverage is 90% (*dashed*)

exact bootstrap distributions, we refer to it as the *Agresti* method. Second, we consider a kernel-based method. Kernels are Gaussian-shaped with bandwidth parameter chosen as in Hall et al. (2004). We refer to this as the *Kernel* method. Third, we report results for the *Empirical* method whereby 100 bootstrap samples are drawn from the test set. Confidence intervals are obtained from the quantiles of the distribution of the true positive rate for a series of fixed values of the false positive rate.

Figures 8 and 9 provide comparative results for the three methods considered here, using our 6 UCI data sets with test set sizes of 25 and 250, respectively. With a sample size of 25 (Fig. 8) and may be with the exception of the Chess data set, the Empirical (dashed) method exhibits severe coverage breaks at high false positive rates. The Kernel method performs better with a single severe break at low false positive rates for the Abalone data set. But it also leads to substantially overconservative intervals on the Adult, Chess and Covertype data sets. Overall, the Agresti method performs best with the most severe break of 0.753 occurring at high false positive rates on the German data set and the highest coverage of 0.965 at high false positive ratios on the Telescope data set.

With sample sizes of 250 (Fig. 9), coverage is generally better. Here again, the Empirical method exhibits severe coverage breaks at high false positive rates, with the exception of the Chess and German data sets. Severe breaks also affect the Kernel method on the Abalone,

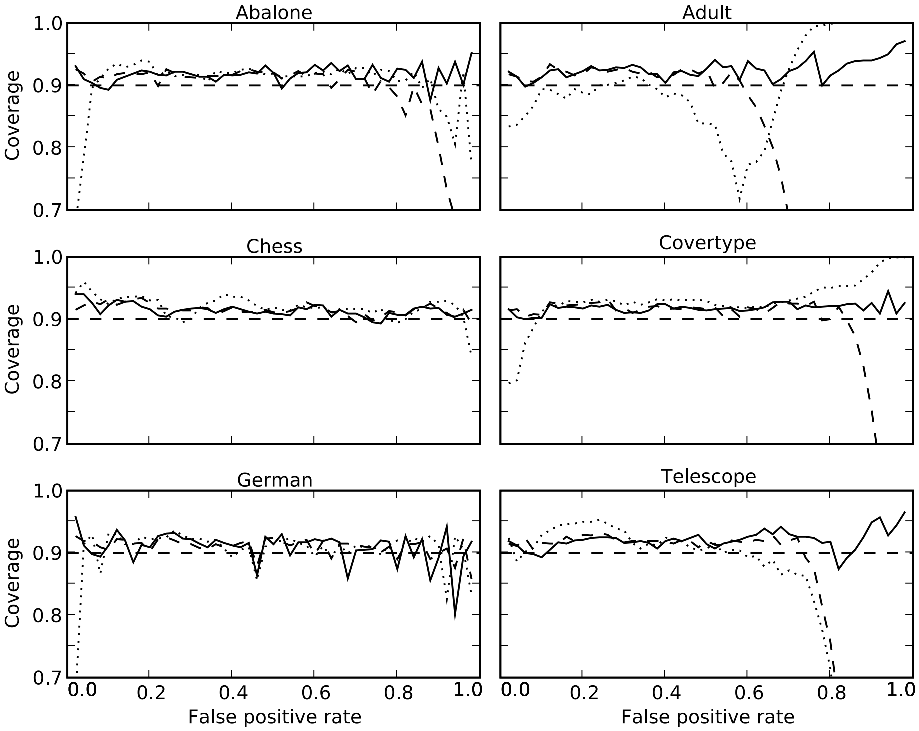

Fig. 9 Vertical averaging and single design. Coverage accuracy for six real data sets. The coverage accuracies of three methods: Agresti (*solid*), Kernel (*dotted*), and Empirical (*dashed*) are plotted against the theoretical total positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 250 and target coverage is 90% (*dashed*)

German, and Telescope data sets. Also, on the Adult and Covertype data sets, coverage reaches 1. The Kernel method behaves somewhat chaotically on the Adult data set. Here again, the Agresti method fares best with lowest coverage of 0.802 on the German data set and highest coverage of 0.971 on the Adult data set.

In conclusion, Figs. 8 and 9 show that the Agresti method succeeds in avoiding the most severe coverage breaks that affect both other methods, particularly the Empirical method. The Agresti method also avoids producing too high coverage.

5.4 Vertical averaging and paired designs

This subsection presents coverage accuracy results for the case of vertical averaging with paired designs. The *Agresti* method as described in Algorithm 5 is compared to its *Wald* counterpart and the *Empirical* method. The experiment of Sect. 5.2 is used for that purpose. Results appear in Fig. 10. Coverage accuracy is generally too conservative for all three methods. For very low false positive rates, coverage drops for all methods. For high false positive rates, coverages for the Wald and Empirical methods drop severely. On the other hand, coverage for the Agresti method is above target, for high false positive rates.

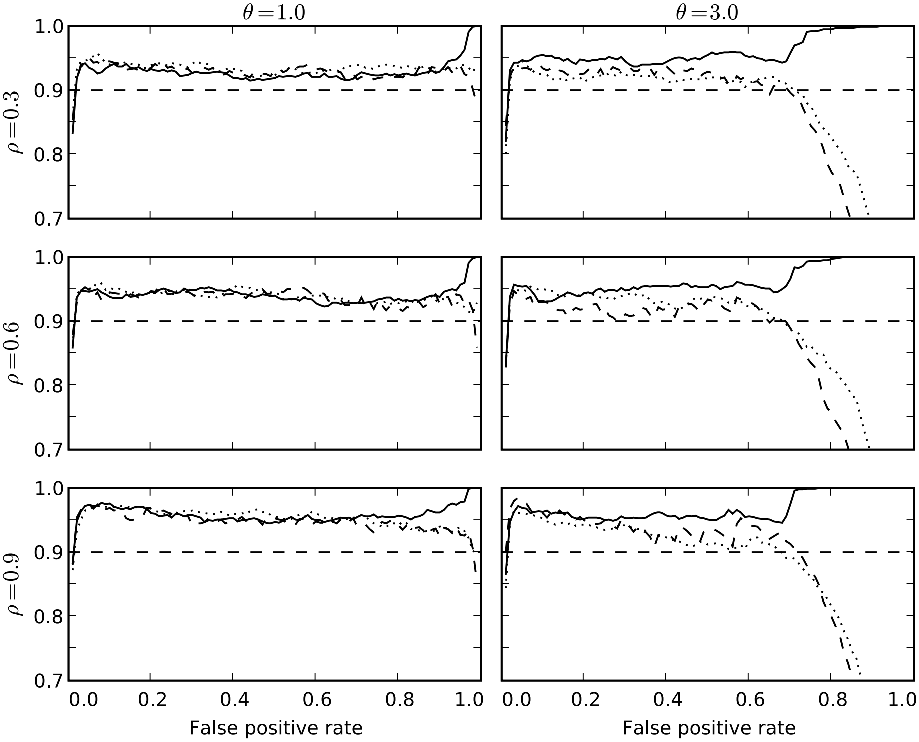

Fig. 10 Vertical averaging and paired design. Location parameter for positive instances of first model is set to $\theta = 1.0$ (*left*) and $\theta = 3.0$ (*right*). Correlation factor is equal to 0.3 (*top*), 0.6 (*middle*) and 0.9 (*bottom*). The coverage accuracies of three methods: Agresti (*solid*), Wald (*dotted*), and Empirical (*dashed*) are plotted against the false positive rate. Confidence intervals are built with significance level $\alpha = 10\%$. Coverage proportion is obtained from 1,000 simulations. Sample size is 100 and target coverage is 90% (*dashed*)

The conclusion of this subsection is similar to those of Sects. 5.1 and 5.2: at the expense of being somewhat too conservative, the Agresti method avoids the important coverage accuracy breaks that poise the Wald and Empirical methods.

5.5 The impact of stratified sampling

Formulas of Sects. 3 and 4 have been obtained assuming stratified sampling. As described in Sect. 1, this corresponds to assuming that the true underlying proportions of positive and negative instances are equal to those of the test set. In that case, each sample is in fact obtained as the union of two independent samples that have been drawn from two disjoint pools of instances: positive and negative instances of the test set. One may wonder whether the confidence intervals we have developed so far are robust to changes in the proportions of true and false positives, i.e. whether confidence intervals obtained assuming stratified sampling perform as well when the actual sampling is full, i.e. samples are drawn from the entire test set, a single pool of instances including positive as well as negative instances.

In Fig. 11, we reconsider the size experiment of Sect. 5.1 where threshold averaging and single design are used. As is apparent, both sampling approaches yield very similar coverage.

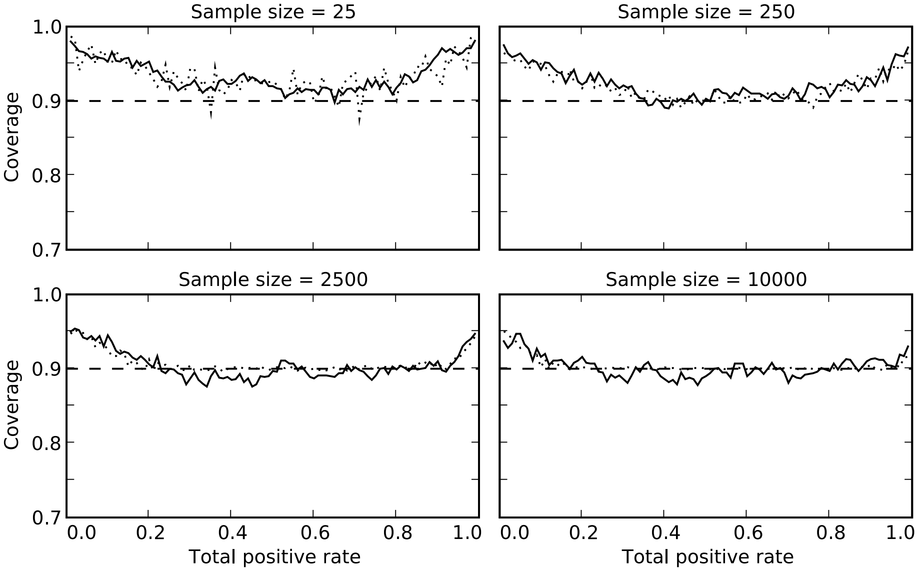

Fig. 11 Comparison of full and stratified sampling. The size experiment with threshold averaging and paired design (Sect. 5.1) is used to compare the two sampling approaches. Samples are drawn using stratified (*dotted*) and full (*solid*) sampling. Target coverage of 90% is also plotted (*dashed*) against total positive rate

6 Conclusion

In this paper, we derived pointwise exact bootstrap distributions of ROC curves (single design) and for the difference between two ROC curves (paired design). Paired design is of particular importance to the machine learning community as it directly pertains to the issue of model selection. Paired design has received very little attention and by addressing it in great details in this paper, we provide new analysis tools for machine learning research. Combining designs with the two averaging techniques considered (threshold and vertical), leads to the four problems that were addressed. Using these pointwise exact bootstrap distributions and building upon previous work, mainly Agresti and Coull (1998) and Agresti and Min (2005), confidence intervals and regions were obtained.

For the problem involving vertical averaging and single design, by far the one that has received most attention in the literature, *Wald*-type confidence intervals and a *Kernel*-based semi-parametric approach (Hall et al. 2004) were considered. On this specific problem, the suggested (Agresti) approach was shown to outperform both the Wald and Kernel approaches as the former had the tightest range of observed coverage values around target coverage. For the three other problems, the proposed Agresti approach was compared to *Wald* confidence intervals derived from pointwise exact bootstrap distributions and an approach that consists of actually taking a certain number of bootstrap samples of the test set. We referred to this later approach as *Empirical*. On all three problems, both Wald and Empirical approaches were shown to be poised with severe drops in coverage accuracy. These coverage breaks can be avoided by using the proposed approach. Unfortunately, in these cases, the Agresti approach obtains coverage that is somewhat above target. For threshold averaging, coverage accuracy was plotted against *total* positive rate, thereby allowing us to perform comparisons that are independent of score scales.

In essentially all of the literature on ROC curves, confidence intervals are derived assuming fixed proportions of positive and negative instances, what we referred to as *stratified* sampling. This element is never questioned, although is does represent a departure from real-world applications. In reality, the observed test set proportions may deviate from the true underlying population proportions. We relaxed this hypothesis of fixed proportions by considering *full* sampling whereby bootstrap samples of varying proportions are drawn from the test set in order to derive confidence intervals. In our experiments, the use of either sampling technique had marginal impact on coverage, a result that we consider as reassuring.

In the future, we wish to investigate algorithmic accelerations through approximations of the Agresti approach, when applied to the case of vertical averaging and paired design. Also, for some applications, the number of instances that may be labelled and *treated* as positive may be limited, due to budget constraints. This suggests that a different averaging technique may be appropriate, one that derives confidence intervals given fixed values for the total positive rate, rather than fixed thresholds or fixed false positive rates.

Appendix A: Probability of dominance for threshold averaging

In this section, we derive exact stratified bootstrap probabilities that scoring model 1 dominates model 2, given that model 1 and model 2 assign a positive label to all instances with scores greater or equal to t_1 and t_2, respectively.

Let $n^+_{t_1,t_2}$ be the number of positive instances in the test set with first score greater or equal to t_1 and second score greater or equal to t_2. Also, let $n^+_{t_1,\overline{t_2}}$ ($n^+_{\overline{t_1},t_2}$) be the number of positive instances with first score greater or equal to (below) t_1 and the second score below (greater or equal to) t_2. Finally, $n^+_{\overline{t_1},\overline{t_2}}$ represents the number of positive instances with both scores below their respective thresholds. Thus, the following relationship holds:

$$n^+ = n^+_{t_1,t_2} + n^+_{t_1,\overline{t_2}} + n^+_{\overline{t_1},t_2} + n^+_{\overline{t_1},\overline{t_2}}.$$

Let $p^+_{t_1,t_2} = n^+_{t_1,t_2}/n^+$ be the probability that a positive instance drawn at random from the test set has both scores above their respective thresholds. Similarly, $p^+_{t_1,\overline{t_2}} = n^+_{t_1,\overline{t_2}}/n^+$, $p^+_{\overline{t_1},t_2} = n^+_{\overline{t_1},t_2}/n^+$, and $p^+_{\overline{t_1},\overline{t_2}} = n^+_{\overline{t_1},\overline{t_2}}/n^+$.

Consider a random bootstrap sample of size m^+ drawn from the set of n^+ positive instances of the test set. Let $M^+_{t_1,t_2}$, $M^+_{t_1,\overline{t_2}}$, $M^+_{\overline{t_1},t_2}$ and $M^+_{\overline{t_1},\overline{t_2}}$ be the random variables for the numbers of instances, defined similarly as above but for the set of instances of the bootstrap sample. We have,

$$m^+ = M^+_{t_1,t_2} + M^+_{t_1,\overline{t_2}} + M^+_{\overline{t_1},t_2} + M^+_{\overline{t_1},\overline{t_2}}.$$

Then, the difference in true positive rates is

$$\begin{aligned}\Delta TP^+_{t_1,t_2} &= TP^+_{t_1} - TP^+_{t_2}\\ &= \frac{M^+_{t_1,t_2} + M^+_{t_1,\overline{t_2}}}{m^+} - \frac{M^+_{t_1,t_2} + M^+_{\overline{t_1},t_2}}{m^+}\\ &= \frac{M^+_{t_1,\overline{t_2}} - M^+_{\overline{t_1},t_2}}{m^+}. \end{aligned} \tag{8}$$

According to the stratified bootstrap sampling approach, m^+ is fixed and the difference in true positive rates depends on the two values $M^+_{t_1,\overline{t_2}}$ and $M^+_{\overline{t_1},t_2}$ which have trinomial joint distribution:

$$Pr\{M^{+}_{t_1,\overline{t_2}} = i, M^{+}_{\overline{t_1},t_2} = l\} = \binom{m^{+}}{i,l} (p^{+}_{t_1,\overline{t_2}})^{i} (p^{+}_{\overline{t_1},t_2})^{l} (p^{+}_{t_1,t_2} + p^{+}_{\overline{t_1,t_2}})^{m^{+}-i-l}$$
$$= b(u,i,m^{+}) \cdot b(v,l,m^{+}-i) \tag{9}$$

where $\binom{n}{a,b} = \frac{n!}{a!\cdot b!\cdot (n-a-b)!}$, $u = p^{+}_{t_1,\overline{t_2}}$, and $v = \frac{p^{+}_{\overline{t_1},t_2}}{1-p^{+}_{t_1,\overline{t_2}}}$ so that,

$$Pr\{\Delta TP^{+}_{t_1,t_2} = 0\} = \sum_{i=0}^{\lfloor m^{+}/2\rfloor} b(u,i,m^{+}) \cdot b(v,i,m^{+}-i), \tag{10}$$

$$Pr\{\Delta TP^{+}_{t_1,t_2} \geq 0\} = \sum_{i=0}^{\lfloor m^{+}/2\rfloor} b(u,i,m^{+}) \cdot B(v,i,m^{+}-i)$$
$$+ 1 - B(u, \lfloor m^{+}/2\rfloor, m^{+}). \tag{11}$$

Similar results are obtained for negative instances and $f_{1,t}$ is obtained using (2). Using (10) and (11), we obtain probabilities $Pr\{\Delta TP^{+}_{t_1,t_2} \geq 0\}$, $Pr\{\Delta FP^{-}_{t_1,t_2} \leq 0\}$, $Pr\{\Delta TP^{+}_{t_1,t_2} = 0\}$, $Pr\{\Delta FP^{-}_{t_1,t_2} = 0\}$ in linear time. Assuming we wish to evaluate the probability of dominance for h different thresholds, this leads to an $O(n \cdot h)$ algorithm.

Finally, from (9), first and second moments for the distribution of $\Delta TP^{+}_{t_1,t_2}$ are easily obtained:

$$E\{\Delta TP^{+}_{t_1,t_2}\} = p^{+}_{t_1,\overline{t_2}} - p^{+}_{\overline{t_1},t_2},$$

$$\text{Var}\{\Delta TP^{+}_{t_1,t_2}\} = \frac{p^{+}_{t_1,\overline{t_2}} + p^{+}_{\overline{t_1},t_2} - (p^{+}_{t_1,\overline{t_2}} - p^{+}_{\overline{t_1},t_2})^2}{m^{+}}.$$

Values $p^{+}_{t_1,t_2}$, $p^{+}_{\overline{t_1},t_2}$, $p^{+}_{t_1,\overline{t_2}}$, $p^{+}_{\overline{t_1,t_2}}$ are computed for each pair of thresholds and this is performed in $O(n \cdot h)$ time.

Note that a naive approach that considers scores of different models as independent random variables leads to the same expected value for $\Delta TP^{+}_{t_1,t_2}$ but a larger variance $\text{Var}^{N}\{\Delta TP^{+}_{t_1,t_2}\}$:

$$\text{Var}^{N}\{\Delta TP^{+}_{t_1,t_2}\} = \text{Var}\{TP^{+}_{t_1}\} + \text{Var}\{TP^{+}_{t_2}\}$$
$$= \frac{(p^{+}_{t_1,t_2} + p^{+}_{t_1,\overline{t_2}})(p^{+}_{\overline{t_1},t_2} + p^{+}_{\overline{t_1,t_2}})}{m^{+}}$$
$$+ \frac{(p^{+}_{t_1,t_2} + p^{+}_{\overline{t_1},t_2})(p^{+}_{t_1,\overline{t_2}} + p^{+}_{\overline{t_1,t_2}})}{m^{+}}$$
$$= \text{Var}\{\Delta TP^{+}_{t_1,t_2}\}$$
$$+ 2 \cdot \frac{p^{+}_{t_1,t_2} \cdot p^{+}_{\overline{t_1,t_2}} - p^{+}_{t_1,\overline{t_2}} \cdot p^{+}_{\overline{t_1},t_2}}{m^{+}}. \tag{12}$$

This last result has an intuitive interpretation: larger positive correlations between scores of two models lead them to disagree less often so that values for $p^{+}_{\overline{t_1},t_2}$ and $p^{+}_{t_1,\overline{t_2}}$ should be smaller, relative to $p^{+}_{t_1,t_2}$ and $p^{+}_{\overline{t_1,t_2}}$, and the variance overestimation should be greater.

Appendix B: Distribution of ΔTP_r^+ for vertical averaging

In this section, we derive the exact stratified bootstrap distribution of the difference in the true positive rates between two scoring models, given fixed false positive rate r. From this distribution, confidence intervals can be built. In the first part of the section, we obtain the joint (exact stratified bootstrap) distribution of the two model thresholds, given false positive rate r. Then, conditional on these threshold values, the distribution of the difference between the true positive rates follows.

Let $\mathcal{N}^-_{k,j}$ be the set of negative instances with first score greater or equal to $s_{1,k}$ and second score greater or equal to $s_{2,j}$. Similarly, $\mathcal{N}^-_{\bar{k},j}$ ($\mathcal{N}^-_{k,\bar{j}}$) denotes the set of negative instances with first score below (greater or equal to) $s_{1,k}$ and second score greater or equal to (below) $s_{2,j}$ and $\mathcal{N}^-_{\overline{k,j}}$ is the set of negative instances with first score below $s_{1,k}$ and second score below $s_{2,j}$. Let $n^-_{k,j}$, $n^-_{\bar{k},j}$, $n^-_{k,\bar{j}}$, $n^-_{\overline{k,j}}$ be the cardinalities of the four sets defined above. Finally, let $p^-_{k,j} = n^-_{k,j}/n^-$ be the probability that a negative instance drawn at random from the test set has first and second scores greater or equal to $s_{1,k}$ and $s_{2,j}$, respectively. Similarly, $p^-_{k,\bar{j}} = n^-_{k,\bar{j}}/n^-$, $p^-_{\bar{k},j} = n^-_{\bar{k},j}/n^-$, and $p^-_{\overline{k,j}} = n^-_{\overline{k,j}}/n^-$.

Given fixed false positive rate r, let $T^-_{1,r}$ and $T^-_{2,r}$ be the random variables for the thresholds obtained, according to the first and second models, for random stratified bootstrap sample $x^- \in \mathcal{X}^-$, of size m^-, drawn from the n^- negative instances of the test set. We are interested in evaluating the joint probability that $T^-_{1,r} \geq s_{1,k}$ and $T^-_{2,r} \geq s_{2,j}$. Surely, these two conditions are simultaneously satisfied if at least r negative instances, have been chosen from set $\mathcal{N}^-_{k,j}$. There is also the possibility that less than r instances have been drawn from set $\mathcal{N}^-_{k,j}$ but that sufficiently many have been drawn from sets $\mathcal{N}^-_{\bar{k},j}$ and $\mathcal{N}^-_{k,\bar{j}}$ so that both conditions are still respected. Thus,

$$\begin{aligned} F_{r,k,j} &= Pr\{T^-_{1,r} \geq s_{1,k}, T^-_{2,r} \geq s_{2,j}\} \\ &= \sum_{l=r}^{m^-} b(p^-_{k,j}, l, m^-) + \sum_{l=(2r-m^-)_+}^{r-1} b(p^-_{k,j}, l, m^-) \cdot C^-_{k,j}(l, r) \end{aligned} \tag{13}$$

where $(x)_+ = \max(0, x)$ and $C^-_{k,j}(l, r)$ is a sum of trinomial coefficients:

$$C^-_{k,j}(l, r) = \sum_{a=r-l}^{m^- - r} \sum_{b=r-l}^{m^- - l - a} \binom{m^- - l}{a, b} \frac{(p^-_{\bar{k},j})^a (p^-_{k,\bar{j}})^b (p^-_{\overline{k,j}})^{m^- - l - a - b}}{(p^- - p^-_{k,j})^{m^- - l}} \tag{14}$$

with $C^-_{k,j}(r, r) = 1$. We define $F_{0,k,j} = 1$ and $F_{m^-+1,k,j} = 0$ for all values of k and j. From (13), we obtain the following recursive relationship for $r = 1, 2, \ldots, m^-$:

$$\begin{aligned} F_{r+1,k,j} = F_{r,k,j} &- \sum_{l=(2r+2-m^-)_+}^{r} b(p^-_{k,j}, l, m^-)[C^-_{k,j}(l, r) - C^-_{k,j}(l, r+1)] \\ &- b(p^-_{k,j}, 2r - m^- + 1, m^-) \cdot C^-_{k,j}(2r - m^- + 1, r) \cdot I_{\{2r+1-m^- \geq 0\}} \\ &- b(p^-_{k,j}, 2r - m^-, m^-) \cdot C^-_{k,j}(2r - m^-, r) \cdot I_{\{2r-m^- \geq 0\}} \end{aligned} \tag{15}$$

where $I_{\{x \geq 0\}}$ is 1 if $x \geq 0$ and zero otherwise. The difference between trinomial coefficients simplifies to the following:

$$C^-_{k,j}(l, r) - C^-_{k,j}(l, r+1)$$

$$
\begin{aligned}
&= b\left(\frac{p^-_{\overline{k},j}}{p^- - p^-_{k,j}}, r-l, m^- - l\right) \sum_{b=r-l}^{m^- - r} b\left(\frac{p^-_{k,\overline{j}}}{p^-_{k,\overline{j}} + p^-_{\overline{k,j}}}, b, m^- - r\right) \\
&\quad + b\left(\frac{p^-_{k,\overline{j}}}{p^- - p^-_{k,j}}, r-l, m^- - l\right) \sum_{a=r-l}^{m^- - r} b\left(\frac{p^-_{\overline{k},j}}{p^-_{\overline{k},j} + p^-_{\overline{k,j}}}, a, m^- - r\right) \\
&\quad - b\left(\frac{p^-_{k,\overline{j}}}{p^- - p^-_{k,j}}, r-l, m^- - l\right) \cdot b\left(\frac{p^-_{\overline{k},j}}{p^-_{\overline{k},j} + p^-_{\overline{k,j}}}, r-l, m^- - r\right).
\end{aligned}
\tag{16}
$$

Binomial coefficients that appear in the summations of (16) are independent of l but the summation itself includes additional terms as l increases. Thus, the sum of terms can be accumulated so that $F_{r,k,j} - F_{r+1,k,j}$ is computed in linear time. The last two terms of (15) can be computed in constant time. If $2r + 1 - m^- \geq 0$, the next to last term is given by

$$
\begin{aligned}
&\binom{m^-}{m^- - r, m^- - r - 1} (p^-_{k,j})^{2r - m^- + 1} (p^-_{\overline{k},j} \cdot p^-_{k,\overline{j}})^{m^- - r - 1} \\
&\times [(m^- - r) n^-_{\overline{k,j}} + n^-_{k,\overline{j}} + n^-_{\overline{k},j}]
\end{aligned}
\tag{17}
$$

and if $2r - m^- \geq 0$, the last terms simplifies to

$$
\binom{m^-}{m^- - r, m^- - r} (p^-_{k,j})^{2r - m^-} (p^-_{\overline{k},j} \cdot p^-_{k,\overline{j}})^{m^- - r}. \tag{18}
$$

Using (15), (16), (17), and (18), values of $F_{r,k,j}$ are obtained for all values of r, k and j. Each probability $F_{r,k,j}$ is computed in linear time so that computing all of them takes time $O(n^4)$.

Then, the probability density function $f_r(k, j)$ is easily obtained as

$$
\begin{aligned}
f_r(k, j) &= Pr\{T^-_{1,r} = s_{1,k}, T^-_{2,r} = s_{2,j}\} \\
&= F_{r,k,j} - F_{r,k,j-1} - F_{r,k-1,j} + F_{r,k-1,j-1}
\end{aligned}
\tag{19}
$$

and this concludes the first part of the section.

We now turn to the distribution of the difference in true positive rates between two scoring models, conditional on threshold values $T^-_{1,r} = s_{1,k}$ and $T^-_{2,r} = s_{2,j}$. Let $n^+_{k,j}, n^+_{\overline{k},j}, n^+_{k,\overline{j}}, n^+_{\overline{k,j}}$ be defined as counts of the positive instances as $n^-_{k,j}, n^-_{\overline{k},j}, n^-_{k,\overline{j}}, n^-_{\overline{k,j}}$ were defined for the negative instances. Dividing these counts of positive instances by n^+, we obtain the associated probabilities $p^+_{k,j}, p^+_{\overline{k},j}, p^+_{k,\overline{j}}$, and $p^+_{\overline{k,j}}$. Finally, let $M^+_{k,j}, M^+_{\overline{k},j}, M^+_{k,\overline{j}}, M^+_{\overline{k,j}}$ be the corresponding counts for sample $x^+ \in \mathcal{X}^+$. Let $TP^+_{1,r}$ and $TP^+_{2,r}$ be the true positive rates for the two scoring models. Their difference is then

$$
\begin{aligned}
\Delta TP^+_r &= TP^+_{1,r} - TP^+_{2,r} \\
&= \frac{M^+_{k,j} + M^+_{k,\overline{j}}}{m^+} - \frac{M^+_{k,j} + M^+_{\overline{k},j}}{m^+} \\
&= \frac{M^+_{k,\overline{j}} - M^+_{\overline{k},j}}{m^+}.
\end{aligned}
\tag{20}
$$

According to the stratified sampling procedure, m^+ is fixed and the difference in true positive rates depends on the two values $M^+_{k,\bar{j}}$ and $M^+_{\bar{k},j}$ with trinomial joint distribution:

$$Pr\{M^+_{k,\bar{j}} = i,\ M^+_{\bar{k},j} = l\} = \binom{m^+}{i, l} (p^+_{k,\bar{j}})^i (p^+_{\bar{k},j})^l (p^+_{k,j} + p^+_{\bar{k},\bar{j}})^{m^+-i-l}. \tag{21}$$

The thresholds conditional distribution of ΔTP^+_r is a sum of some of these trinomial probabilities. For $d \geq 0$, we have:

$$\begin{aligned} g_{k,j}(d) &= Pr\{\Delta TP^+_r = d/m^+ | T^-_{1,r} = s_k, T^-_{2,j} = s_j\} \\ &= Pr\{M^+_{k,\bar{j}} = i,\ M^+_{\bar{k},j} = i - d\} \\ &= \sum_{i=d}^{m^+} \binom{m^+}{i, i-d} (p^+_{k,\bar{j}})^i (p^+_{\bar{k},j})^{i-d} (p^+_{k,j} + p^+_{\bar{k},\bar{j}})^{m^+-2i+d} \end{aligned} \tag{22}$$

and for $d < 0$, the summation index i ranges from 0 to $m^+ - d$, inclusively. There are $2m^+ + 1$ different possible values for the difference d/m^+: $-1, (-m^+ + 1)/m^+, \dots, -1/m^+, 0, 1/m^+, \dots, (m^+ - 1)/m^+, 1$. Obtaining the conditional distribution $g_{k,j}(d)$ for all values of d takes quadratic time. Repeating the procedure for all values of k and j takes time $O(n^4)$. With these values, the unconditional distribution of ΔTP^+_r is obtained as

$$h_r(d) = \sum_{k,j} f_r(k,j) \cdot g_{k,j}(d). \tag{23}$$

Computing this distribution for all values of r and d takes overall computational time $O(n^4)$. Algorithm 5 summarizes the steps described in this section in order to obtain the distribution of ΔTP^+_r.

An alternative is to use (21) in order to derive the first two conditional moments of ΔTP^+_r:

$$\begin{aligned} \mu'_{k,j} &= E\{\Delta TP^+_r | T^-_{1,r} = s_k, T^-_{2,r} = s_j\} \\ &= p^+_{k,\bar{j}} - p^+_{\bar{k},j}, \end{aligned} \tag{24}$$

$$\begin{aligned} \mu''_{k,j} &= E\{(\Delta TP^+_r)^2 | T^-_{1,r} = s_k, T^-_{2,r} = s_j\} \\ &= (1 - 1/m^+)(p^+_{k,\bar{j}} - p^+_{\bar{k},j})^2 + \frac{p^+_{k,\bar{j}} + p^+_{\bar{k},j}}{m^+}. \end{aligned} \tag{25}$$

Computing (24) and (25) for all values of $k, j \in \{1, 2, \dots, q\}$ takes quadratic time. Then, combining these threshold conditional expectations with threshold distributions (15)–(19) in order to obtain unconditional moments and this is done, for each false positive rate, in quadratic time:

$$\mu'_r = E\{\Delta TP^+_r\} = \sum_{k,j} f_r(k,j) \cdot \mu'_{k,j}, \tag{26}$$

$$\mu''_r = E\{(\Delta TP^+_r)^2\} = \sum_{k,j} f_r(k,j) \cdot \mu''_{k,j}. \tag{27}$$

Repeating this procedure for all false positive rate values therefore takes cubic time. Using these unconditional moments, Gaussian distributions can be fitted in order to obtain confidence intervals for ΔTP_r^+. Here again, although faster, the algorithm is dominated by the time spent computing the joint threshold distribution so that the overall computational time remains $O(n^4)$.

Appendix C: UCI data sets

In this section, we describe how the six UCI repository data sets were preprocessed and split before obtaining the experimental results described above, in Sect. 5. Before we do so, let us first describe the technique of *one-hot encoding*. One-hot encoding is used to obtain a numerical representation for a categorical variable. A variable with n categories is represented by (mapped onto) a vector of $n - 1$ binary variables. One of the categories, the "base" category, is associated to the null vector (a vector of $n - 1$ zeros). Each of the other $n - 1$ categories is associated to a vector for which only one of the binary variables is equal to one while all other binary variables are equal to zero, thus the term *one-hot*. This technique has been used to preprocess some of the variables of the data sets which we now describe.

- *Abalone*: the purpose of this data set is to predict the age of an abalone from a set of 8 attributes. The first attribute, sex, was one-hot encoded. Other attributes are all numerical and were left unchanged. For the purpose of binary classification, we split the data set between young ($<$10 years-old) and old ($\geq$10 years-old) abalones.
- *Adult*: based on the 1994 census data, predict whether an individual's income is above $50,000. Categorical data was one-hot encoded. The original data set is already split between a training set (32,561 observations) and a test set (16,281 observations). We used the same split for our experiments.
- *Chess*: a database of chess endgames involving the white king and white rook against the black king. The goal is to identify whether a particular configuration of the three pieces will lead to a white win or a draw between the two players, assuming optimal moves on both sides. In the original data set, wins are categorized according to the number of moves necessary before the end of the game but, for the purpose of binary classification, we considered wins as a single class. Rows and columns were one-hot encoded and since the observations are ordered, we randomly shuffled them. The first 5,000 observations were used to train the models and the remaining 23,056 were used as a test set.
- *Covertype*: the objective of this data set is to predict forest cover type of undisturbed forests, given a set of 54 attributes. The original data set includes seven different cover types but, since Spruce-Fir (211,840) and Lodgepole Pine (283,301) make up for more than 80% of the observations, only these two types were retained for the purpose of binary classification. All 54 attributes of the original data set are numerical. The first 5,000 observations were used to train the models and the remaining 405,141 were used for testing.
- *Credit*: German credit approval data set. Given a set of 24 numeric attributes, the task is to discriminate individuals considered to have good credit from those with a bad credit record. The first 500 observations were used to train models and the remaining 500 were used as a test set.
- *Telescope*: this is the Major Atmospheric Gamma Imaging Cherenkov Telescope project (MAGIC) data set. The purpose is to discriminate between electromagnetic shower images initiated by primary gammas (signal) from hadronic shower images (background noise) caused by cosmic rays in the upper atmosphere. All 10 attributes are numeric.

Table 3 UCI's data sets. Starred (*) data sets required shuffling since their on-line version is sorted

Data Set	Training set		Test set		Total
	Positives	Negatives	Positives	Negatives	
Abalone	410	590	1,686	1,491	4,177
Adult	7,841	24,720	3,846	12,435	48,842
Chess*	4,505	495	20,755	2,301	28,056
Covertype	2,739	2,261	280,562	209,579	495,141
Credit	364	136	336	164	1,000
Telescope*	328	672	11,660	6,360	19,020

Rows were shuffled since the data set is sorted. The first 1,000 observations were used for model training.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician*, *52*(2), 119–226.

Agresti, A. & Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*, *24*, 729–740.

Asuncion, A., & Newman, D. J. (2007). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml.

Bandos, A. (2005). *Nonparametric methods in comparing two correlated ROC curves*. PhD thesis, Graduate School of Public Health, University of Pittsburgh.

Bengio, S., Mariéthoz, J., & Keller, M. (2005). The expected performance curve. In *Proceedings of the ICML 2005 workshop on ROC analysis in machine learning*, Bonn, Germany.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413.

Drummond, C., & Holte, R. C. (2006). Cost curves: an improved method for visualizing classifier performance. *Machine Learning*, *65*(1), 95–130.

Dugas, C., & Gadoury, D. (2008). Pointwise exact bootstrap distribution of cost curves. In A. McCallum & S. Roweis, (Eds.), *Proceedings of the twenty fifth international conference on machine learning*, Helsinki, Finland (pp. 280–287).

Efron, B., & Tibshirani, R. J. (1993). *Monographs on statistics and probability*. *Vol. 57*: *An introduction to the bootstrap*. London: Chapman & Hall.

Fawcett, T. (2004). *ROC graphs: Notes and practical considerations for researchers*. Technical report, HP Laboratories.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Fawcett, T. (2006). ROC graphs with instance varying costs. *Pattern Recognition Letters*, *27*(8), 882–891.

Fawcett, T., & Flach, A. (2005). A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*, *58*(1), 33–38.

Fawcett, T., & Niculescu-Mizil, A. (2007). PAV and the ROC convex hull. *Machine Learning*, *68*(1), 97–106.

Hall, P., Hyndman, R. J., & Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, *91*(3), 743–750.

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

Hsieh, F., & Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, *24*(1), 25–40.

Kerekes, J. (2008). Receiver operating characteristic curve confidence intervals and regions. *IEEE Geoscience and Remote Sensing Letters*, *5*(2), 251–255.

Lloyds, C. J. (1998). The use of smoothed ROC curves to summarise and compare diagnostic systems. *Journal of the American Statistical Association*, *93*, 1356–1364.

Lloyds, C. J., & Wong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters*, *44*, 221–228.

Macskassy, S. A., Provost, F., & Rosset, S. (2005). Pointwise ROC confidence bounds: an empirical evaluation. In *Proceedings of the ICML 2005 workshop on ROC analysis in machine learning*, Bonn, Germany.

Platt, J. (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Barlett, B. Schölkopf & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). Cambridge: MIT Press.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American*, *283*(4), 82–87.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*. San Diego: Academic Press.

Webb, G. I., & Ting, K. M. (2005). On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*, *58*(1), 25–32.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *KDD'02: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 694–699). New York: ACM.

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