Published in final edited form as:

Commun Stat Simul Comput. 2016 ; 45(5): 1689–1703. doi:10.1080/03610918.2015.1065327.

Validation of Nonparametric Two-Sample Bootstrap in ROC Analysis on Large Datasets

Jin Chu Wu, **Alvin F. Martin**, and **Raghu N. Kacker**

National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

The nonparametric two-sample bootstrap is applied to computing uncertainties of measures in ROC analysis on large datasets in areas such as biometrics, speaker recognition, etc., when the analytical method cannot be used. Its validation was studied by computing the SE of the area under ROC curve using the well-established analytical Mann-Whitney-statistic method and also using the bootstrap. The analytical result is unique. The bootstrap results are expressed as a probability distribution due to its stochastic nature. The comparisons were carried out using relative errors and hypothesis testing. They match very well. This validation provides a sound foundation for such computations.

Keywords

bootstrap; ROC analysis; validation; large datasets; uncertainty; biometrics; speaker recognition

1 Introduction

Receiver operating characteristic (ROC) analysis provides important statistical techniques in a wide variety of disciplines related to decision making. The uncertainties of different measures used in ROC analysis on large datasets such as the true accept rate (TAR) and the equal error rate (EER) in biometrics, a detection cost function defined as a weighted sum of probabilities of type I error and type II error in speaker recognition evaluation, etc., need to be determined [1, 2].

The analytical method cannot be used in those cases [1–2]. For instance, in speaker recognition evaluation, it is hard to calculate analytically the covariance term of the correlated probabilities of type I error and type II error for the detection cost function. In addition, the analytical method does not take account of the characteristics of the distributions of scores and thus may underestimate the uncertainties if it were to be used [3]. As a result, the nonparametric two-sample bootstrap method is employed to compute those uncertainties in terms of standard errors (SE) and confidence intervals (CI) [1–2, 4–8].

A challenging question then arises: Can the bootstrap method provide reliable estimates of the SEs of the measures in ROC analysis on large datasets? This validation was studied by calculating the SE of the area under the ROC curve (AUC) using the well-established

analytical method as well as using the nonparametric two-sample bootstrap method, and then comparing results from these two methods. Such a validation can provide a sound foundation for applying the bootstrap method to computing the uncertainties of measures in those areas where the analytical method cannot be used.

The AUC evaluates a classifier using a metric which depends on the classifier itself [9, 10]. Such arguments are out of the scope of this article. Nonetheless, the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [11, and references therein]. If the trapezoidal rule is employed, the AUC is equivalent to the Mann-Whitney statistic formed by independent and identically distributed (i.i.d.) genuine scores and impostor scores [12–17]. Hence, the variance of the Mann-Whitney statistic can be utilized as the variance of the AUC. In other words, the SE of the AUC can be computed analytically. This analytical approach is a deterministic process. The result may be treated as the ground truth.

On the other hand, the SE of the AUC can also be estimated using the nonparametric two-sample bootstrap method [4]. Unlike the analytical approach, the bootstrap method is a stochastic process. Hence, the SEs derived using the bootstrap method constitute a probability distribution. Some outcomes may be more probable and others less. Comparisons of such a probability distribution of the bootstrap estimated SÊs of AUC with the single analytically estimated SÊ are carried out using relative errors as well as hypothesis testing. If the differences are small, then the nonparametric two-sample bootstrap method is validated for estimating the uncertainties of measures in ROC analysis on large datasets.

Without loss of generality, in this article, biometric applications on large datasets will be taken as examples. Genuine scores are created by comparing two different images of the same subject, and impostor scores are generated by matching two images of two different subjects. The two distributions of continuous scores are schematically depicted in Figure 1(A). The cumulative probabilities of genuine and impostor scores ranging from a specified score (i.e., threshold) to the highest score are defined as the TAR and the false accept rate (FAR), respectively. As the threshold moves from the highest score down to the lowest score, an ROC curve is constructed in the far-tar coordinate system as drawn in Figure 1(B).

As extensively investigated, the above two distributions usually do not have well defined parametric forms [1, 12]. The bootstrap method assumes that the data are randomly selected and i.i.d.. Our large government data bases used for developing scores in biometric technology were randomly collected from real practice, and thus the data dependency is not involved. As a result, the nonparametric two-sample bootstrap method is pertinent to estimating the SE of the AUC. The empirical distribution is assumed for each of the observed scores.

The number of bootstrap replications is a very important parameter in bootstrap method [6–8]. In our applications the data samples are large, the statistics of interest are probabilities, and no normality assumption can be made for score distributions. In order to reduce the bootstrap variance and ensure the accuracy of computation in such applications, the

bootstrap variability was empirically studied, and the appropriate number of bootstrap replications was determined to be 2000 [4].

In this article, 14 datasets generated by 14 image matching algorithms[1] are employed as examples. These 14 datasets are completely different with respect to scoring methods in terms of using integers or real numbers in different ranges, the shapes of the score distributions, the overlap between the genuine-score distribution and the impostor-score distribution, etc. However, in each dataset, the total number of genuine scores is a little over 60,000 and the total number of impostor scores is a little over 120,000. As demonstrated in our previous studies of sample size carried out by applying Chebyshev's inequality in biometric applications, if the numbers of scores get larger than these, the measurement accuracy will improve little [18].

To support the main objective in the paper [19], a very preliminary comparison was carried out, in which the stochastic characteristic of the bootstrap method was not taken into account and thus only a single SE estimated by a random execution of the bootstrap rather than a distribution of the bootstrap results was compared with the unique analytically estimated SE. So, the hypothesis testing was not used. And the datasets used in that paper had data dependency but were assumed to be i.i.d..

The exact bootstrap variance of the single AUC was proposed in Ref. [20], which contains a double summation over the total number of the genuine scores and the total number of the impostor scores, and thus is computationally impractical in the applications where tens to hundreds of thousands of scores are involved. Further, it is impossible to figure out the exact bootstrap variance for most statistics of interest. The nonparametric two-sample bootstrap algorithms can be used to compute variance of any measures in ROC analysis. In addition, our bootstrap variability studies determine the appropriate number of bootstrap replications and thus provide a theoretical basis to reduce the bootstrap variance and ensure the accuracy of the computation for the bootstrap applications in ROC analysis on large datasets [4].

The formulas of AUC are presented in Section 2. The analytical method using the Mann-Whitney statistic to estimate the SE of the AUC is shown in Section 3. An algorithm of the nonparametric two-sample bootstrap method for estimating the SE of the AUC is provided in Section 4. The probability distribution of the bootstrap estimated SE of the AUC is explored in Section 5. The 14 different datasets are discussed in Section 6. The results of the analytical method and the bootstrap method, and the comparisons of these two types of results are offered in Section 7. Finally, conclusions and discussion can be found in Section 8.

[1]Specific hardware and software products identified in this paper were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

2 The formulation of AUC

Let $f_G(s)$ and $f_I(s)$ denote the continuous probability density functions (pdf) of the genuine scores and the impostor scores, respectively. Then, the TAR and FAR at a score s are expressed by

$$\mathrm{TAR}(s)=\int_s^{+\infty} f_G(t)dt, \quad (1)$$

and

$$\mathrm{FAR}(s)=\int_s^{+\infty} f_I(t)dt, \quad (2)$$

where $s \in (-\infty, +\infty)$. Defining $\upsilon \equiv \mathrm{FAR}(s)$ where v increases from 0 to 1 as s decreases from $+\infty$ down to $-\infty$, it is clear that $dv = -f_I(s)\,ds$. Then, form Figure 1(B) the AUC can be expressed as

$$\mathrm{AUC}=\int_0^1 \mathrm{TAR}(s)dv=\int_{-\infty}^{+\infty}\left[\int_s^{+\infty} f_G(t)dt\right]\times f_I(s)ds. \quad (3)$$

In other words, the AUC is the integral over score s of the cumulative probability of the genuine pdf from a score s up to positive infinity multiplied by the impostor pdf at the score s.

This formulation of AUC can also be interpreted in terms of discrete distribution functions of scores. All scores were converted into integers if they were not for implementation purposes. Without loss of generality, the scores are expressed inclusively using the integer score set $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$. The genuine score set and the impostor score set in the sense of multiset (i.e., the same score can appear multiple times) are denoted as

$$\mathbf{G}=\{m_i | m_i \in \{s\} \text{ and } i=1,\ldots,N_G\}, \quad (4)$$

and

$$\mathbf{I}=\{n_i | n_i \in \{s\} \text{ and } i=1,\ldots,N_I\}, \quad (5)$$

where N_G and N_I are the total numbers of genuine scores and impostor scores.

Let $P_i(s)$, where $s \in \{s\}$ and $i \in \{G, I\}$, denote the discrete empirical probabilities of genuine scores and impostor scores occurring at a score s, respectively. Thus, the two discrete probability distribution functions can be expressed as

$$\boldsymbol{P_i}=\{\mathrm{P_i(s)}|\forall \mathrm{s}\in\{\mathrm{s}\}\ \text{and}\ \sum_{\tau=s\,\min}^{s\,\max}\mathrm{P_i}(\tau)=1\},\mathrm{i}\in\{\mathrm{G,I}\}. \tag{6}$$

Suppose that the discrete cumulative probabilities of genuine scores and impostor scores at a score s, C_i (s), i ∈ {G, I}, are defined in this article to be the probabilities cumulated from the integer score s up to the highest score s_{max}. Then, the two discrete cumulative probability distribution functions can be expressed as

$$\boldsymbol{C_i}=\{\mathrm{C_i(s)}=\sum_{\tau=s}^{s\,\max}\mathrm{P_i}(\tau)|\forall \mathrm{s}\in\{\mathrm{s}\}\},\mathrm{i}\in\{\mathrm{G,I}\}. \tag{7}$$

The ROC curve in the discrete formation is no longer a smooth curve. While cumulating probabilities of genuine scores and impostor scores starting from the highest score, an ROC curve may move horizontally, vertically, toward the upper right, or stay where it is for each score decrement, depending on whether $P_I(s)$ and/or $P_G(s)$ are nonzero or not. Thus, the AUC consists of a set of trapezoids, each of which is built by a rectangle and a triangle in general. The trapezoid may be reduced to a rectangle, a vertical line, or a point.

At a score s ∈ {s}, in the far-tar coordinate system, a trapezoid depicted schematically in Figure 2 is constructed by four points: A ($C_I(s + 1)$, 0), B ($C_I(s + 1)$, $C_G(s + 1)$), C ($C_I(s)$, $C_G(s)$), and D ($C_I(s)$, 0). It is assumed that $C_I(s_{max} + 1) = C_G(s_{max} + 1) = 0$. This boundary condition indicates that Points A and B in the very first trapezoid corresponding to $s = s_{max}$ are at the origin of the far-tar coordinate system. It also implies $\sum_{\tau=s\,\max+1}^{s\,\max}=0$ due to Eq. (7). Thus, the lengths ($C_I(s) - C_I(s + 1)$) (i.e., $P_I(s)$) and ($C_G(s) - C_G(s + 1)$) (i.e., $P_G(s)$) form a triangle, and the lengths ($C_I(s) - C_I(s + 1)$) (i.e., $P_I(s)$) and $C_G(s + 1)$ (i.e., $\sum_{\tau=s+1}^{s\,\max}\mathrm{P_G}(\tau)$) create a rectangle.

If the trapezoidal rule is employed, the AUC expressed in Eq. (3) can be estimated as

$$\begin{aligned}\hat{\mathrm{A}}&=\sum_{s=s\,\max}^{s\,\min}\mathrm{trapezoid(s)}\\&=\sum_{s=s\,\max}^{s\,\min}\mathrm{triangle(s)}+\sum_{s=s\,\max}^{s\,\min}\mathrm{rectangle(s)}\\&=\sum_{s=s\,\max}^{s\,\min}\mathrm{P_I(s)}\times[\tfrac{1}{2}\times\mathrm{P_G(s)}+\sum_{\tau=s+1}^{s\,\max}\mathrm{P_G}(\tau)].\end{aligned} \tag{8}$$

Note that the summation runs consecutively in the descending order from s_{max} to s_{min}.

3 The analytical method to compute the estimated $\hat{SE}_A(A)$ of AUC

The AUC is estimated using the trapezoidal rule. Thus, the AUC is equivalent to the Mann-Whitney statistic formed by the discrete genuine and impostor scores. Thereafter, the variance of the Mann-Whitney statistic that can be computed analytically is utilized as the variance of AUC [12–17].

In order to relate AUC to the Mann-Whitney statistic, all the N_I scores in the impostor score set **I** in Eq. (5) are compared with all the N_G scores in the genuine score set **G** in Eq. (4). The order relations among scores can be expressed as

$$\mathbf{R}(S_G, S_I) = \begin{cases} 1 & \text{if } S_I < S_G \\ \frac{1}{2} & \text{if } S_I = S_G \\ 0 & \text{if } S_I > S_G \end{cases} \quad (9)$$

By converting probabilities of genuine and impostor scores in Eq. (8) back to frequencies and including zero-frequency scores, the first term in Eq. (8) shows the total number of score pairs, in which both the impostor score and the genuine score are equal to s, divided by $N_G N_I$ and weighted by ½. And the second term in Eq. (8) represents the total number of score pairs, in which the impostor score is less than the genuine score, divided by $N_G N_I$ and weighted by 1 [21]. Hence, the estimated $\hat{AUC}$ can be re-written as

$$\hat{A} = \frac{1}{N_G N_I} \times \sum_{s_G=1}^{N_G} \sum_{s_I=1}^{N_I} \mathbf{R}(s_G, s_I). \quad (10)$$

Except for the coefficient, this is exactly the Mann-Whitney statistic formed by the genuine and impostor scores.

To compute the variance of the Mann-Whitney statistic, two more cumulative probability distribution functions are required [11–15]. One is

$$\boldsymbol{Q}_{\mathbf{G}} = \{Q_G(s) = \sum_{\tau=s+1}^{s\max} P_G(\tau) | \forall s \in \{s\}\}. \quad (11)$$

The other one is

$$\boldsymbol{Q}_{\mathbf{I}} = \{Q_I(s) = \sum_{\tau=s\min}^{s-1} P_I(\tau) | \forall s \in \{s\}\}. \quad (12)$$

Where $\sum_{\tau=s\min}^{s\min-1} = 0$ is assumed.

The probability B_{GGI}, that two randomly chosen genuine matches will obtain higher scores than one randomly chosen impostor match, can be written as

$$B_{GGI} = \sum_{s=s\,\min}^{s\,\max} P_I(s) \times [Q_G{}^2(s) + Q_G(s) \times P_G(s) + \frac{1}{3} \times P_G{}^2(s)]. \quad (13)$$

And the probability B_{IIG}, that one randomly chosen genuine match will get higher score than two randomly chosen impostor matches, can be expressed as

$$B_{IIG} = \sum_{s=s\,\min}^{s\,\max} P_G(s) \times [Q_I{}^2(s) + Q_I(s) \times P_I(s) + \frac{1}{3} \times P_I{}^2(s)]. \quad (14)$$

Finally, the SE of the AUC can be analytically estimated as

$$\hat{SE}_A(A) = \sqrt{\frac{1}{N_G N_I} \times [\hat{A}(1-\hat{A}) + (N_G - 1)(B_{GGI} - \hat{A}^2) + (N_I - 1)(B_{IIG} - \hat{A}^2)]}. \quad (15)$$

4 The bootstrap method to compute the estimated $\hat{SE}_B(A)$ of AUC

The uncertainty of AUC in terms of SE and 95% CI can also be estimated using the nonparametric two-sample bootstrap [1–2, 4–8]. With the i.i.d. assumption, the algorithm is as follows.

Algorithm (Nonparametric two-sample bootstrap)

1: **for** i = 1 **to** B **do**

2: select N_G scores randomly WR from **G** to form a set {new N_G genuine scores}$_i$

3: select N_I scores randomly WR from **I** to form a set {new N_I impostor scores}$_i$

4: {new N_G genuine scores}$_i$ & {new N_I impostor scores}$_i$ => statistic $\hat{A}_i$

5: **end for**

6: $\{\hat{A}_i \mid i = 1, \ldots, B\} \Rightarrow \hat{SE}_B$ and $(\hat{Q}_B(\alpha/2), \hat{Q}_B(1-\alpha/2))$

7: **end**

where B is the number of two-sample bootstrap replications and WR stands for “with replacement”. As shown from Step 1 to 5, this algorithm runs B times. In the i-th iteration, N_G scores are randomly selected WR from the raw genuine score set **G** shown in Eq. (4) to form a new set of scores, and N_I scores are randomly selected WR from the raw impostor score set **I** shown in Eq. (5) to form a new set of scores. Then in Step 4 from these two new sets of scores the i-th bootstrap replication of the estimated $\hat{AUC}$, i.e., $\hat{A}_i = \hat{AUC}_i$, is generated using Eq. (8).

Finally, as indicated in Step 6, after B iterations, the bootstrap distribution $\{\hat{A}_i \mid i = 1, \ldots, B\}$ formed by the B replications of AUC is generated. From this distribution, the $\hat{SE}_B$ estimated

by the sample standard deviation, and the $(1-\alpha)100\%$ $\hat{\text{CI}}(\hat{Q}_B(\alpha/2), \hat{Q}_B(1-\alpha/2))$ estimated by the $\alpha/2$ 100% and $(1-\alpha/2)$ 100% quantiles at the significance level α can be obtained [6]. While computing the quantile, Definition 2 in Ref. [22] is adopted, i.e., the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. For the 95% $\hat{\text{CI}}$, α is set to be 0.05.

The remaining issue is to determine how many iterations the bootstrap algorithms need in order to reduce the bootstrap variance and ensure the accuracy of the computation [6–8]. In our applications of ROC analysis, such as in biometrics and speaker recognition evaluation, the sizes of datasets are tens to hundreds of thousands of scores. Our statistics of interest are mostly probabilities or a weighted sum of probabilities rather than a simple sample mean. Most importantly our data samples of scores have no parametric model to fit. So, the bootstrap variability was re-studied empirically, and the number of bootstrap replications needed for our applications was determined to be 2000 [4].

5 The probability distribution of the bootstrap estimated $\hat{\text{SE}}_B(A)$ of AUC

Due to the stochastic nature of the bootstrap method, different runs can produce different results. Some results may be more probable and others less so. The bootstrap estimated $\hat{\text{SE}}_B(A)$ of AUC constitute a probability distribution. Such a distribution, $\mathbf{SE_B(A)} = \{\hat{\text{SE}}_{B\,i}(A) \mid i = 1, \ldots, L\}$, can be generated by running the above algorithm multiple times. Subsequently, the mean, median, 68.27% CI (i.e., 1 σ) and 95% CI (i.e., 1.96 σ) of the distribution can be estimated.

To determine the number of iterations L, two image matching algorithms, A of high accuracy and B of low accuracy were taken as examples. The number of iterations L was set to vary from 100 up to 500 at intervals of 100. Then the minimum, maximum, and range of L estimated $\hat{\text{SE}}_B(A)$ of AUC were calculated and are shown in Table 1. Across the five different numbers of iterations, for high-accuracy Algorithm A, they round to 0.00013, 0.00014, and 0.00001, respectively; and for low-accuracy Algorithm B, they round to 0.00046, 0.00050, and 0.00004 (with one slight exception), respectively. This indicates that the discrepancies in the results from 100 runs to 500 runs are small.

Further, in order to obtain a statistically meaningful estimated $\hat{\text{CI}}$, the number of estimated $\hat{\text{SE}}_B(A)$ of AUC, i.e., the number of iterations L, must be rather large. For instance, generally speaking, there are only about two instances located outside the 95% $\hat{\text{CI}}$ in each tail of the distribution for L = 100, whereas there are about 12 instances for L = 500. Therefore, for each matching algorithm, 500 estimates of $\hat{\text{SE}}_B(A)$ of AUC will be generated to represent a probability distribution.

Here are two examples. The distributions of 500 bootstrap estimated $\hat{\text{SE}}_B(A)$ of AUC for the high- and low-accuracy Algorithms A and B, respectively, are shown in Figure 3, where the red triangle stands for the analytical result, the blue diamonds are the two bounds of the 68.27% CI, and the green circles represent the two bounds of the 95% CI. It is shown in Figure 3 that Algorithm A has less dispersed values than Algorithm B, and for both

algorithms the analytically estimated $S\hat{E}_A(A)$ of AUC is very close to the mean as well as the median of the distribution (see Algorithms 3 and 14 in Table 2 of Section 7.1).

6 The 14 different datasets

In this article, the 14 datasets generated by 14 image matching algorithms, respectively, are taken as examples. These 14 datasets are quite different. In Figure 4 are shown the distributions of genuine scores and impostor scores for algorithms A (left) and B (right). Besides differences such as scoring methods using integers or real numbers in different ranges, the shapes of the score distributions, etc., Algorithm A has much less overlap between the genuine-score distribution and the impostor-score distribution than Algorithm B does. This indicates why Algorithm A is more accurate than Algorithm B as pointed out in Section 5 [1, 12].

7 Results

7.1 The analytical results and bootstrap results

As mentioned above, the 14 algorithms have different image matching accuracies. The larger the estimated $A\hat{U}C$ is, the more accurate the algorithm is. For each algorithm, the analytically estimated $S\hat{E}_A(A)$ using the Mann-Whitney statistic is unique; but the bootstrap estimated $S\hat{E}_B(A)$ constitute a probability distribution described by the estimated mean, median, 68.27% $C\hat{I}$, 95% $C\hat{I}$, and its own $S\hat{E}_B(SE)$ estimated by the sample standard deviation of the distribution. Along with the estimated $A\hat{U}C$, all these quantities are shown in Table 2. In this table, Algorithms 3 and 14 are Algorithms A and B employed in Section 5, and all distributions were generated by 500 runs.

To highlight the differences, seven decimal places are shown. It may be noted that most of the analytical estimators $S\hat{E}_A(A)$ fall within the estimated 95% $C\hat{I}$ of the distributions of $S\hat{E}_B(A)$; the exceptions are Algorithms 1, 7, and 11. For these three algorithms, there are huge stand-alone peaks at the lowest impostor score, which occupy 98.54%, 97.15%, and 80.02% of the impostor population, respectively. For the other matching algorithms, a stand-alone peak does not occupy more than 50% of the population when it exists. The randomness of resampling scores may be affected by such a huge stand-alone peak of score distributions.

7.2 The comparison of the two types of results using relative errors

The comparison between the bootstrap results and the analytical result can be quantified by the relative error δ of one of the quantities such as the mean, the median, the upper bound and lower bound of the 68.27% $C\hat{I}$, as well as the upper bound and lower bound of the 95% $C\hat{I}$ of the probability distribution of the bootstrap estimated $S\hat{E}_B(A)$ of AUC with respect to the unique analytically estimated $S\hat{E}_A(A)$ of AUC that is computed using Eq. (15). When using the estimated $C\hat{I}$, the larger of the two relative errors using the two bounds of the CI is employed. These relative errors in each scenario are denoted by δ_{mean}, δ_{median}, δ_{68}, and δ_{95}, respectively. The relative error takes into account the impact of the magnitude of the analytical result.

All the relative errors (%) for 14 matching algorithms are shown in Table 3. The corresponding box diagrams of 14 relative errors in each scenario are depicted in Figure 5. It is obvious that there are three outliers corresponding to Algorithms 1, 7, and 11, respectively. This is consistent with the discussion in Section 7.1.

The SEs, created by random runs using the nonparametric two-sample bootstrap, that would be obtained more probably than others are those near the estimated mean and median, and those within the estimated 68.27% CÎ of the distribution of estimated $S\hat{E}_B(A)$ (see Figure 3). The bootstrap estimators of SE can fall in between 68.27% CÎ and 95% CÎ with probability about 27%. And the relative errors δ_{mean}, δ_{median}, and δ_{68} may be more probable than the relative error δ_{95}.

Moreover, it is shown in Figure 5 that the distribution in each of the four scenarios is skewed. Thus, the median of the distribution is more important than the mean. The estimated mean and median of the 14 relative errors (%) in each scenario are shown in Table 4, where the three outlier algorithms are included. Those excluding the three outliers are presented in Table 5.

When the three outliers are included, the worst relative error of $S\hat{E}_B(A)$ is 5.49% which is related to a bound of the 95% CI of the distribution, but the median of 14 relative errors δ_{median} is 0.30%. When the three outliers are excluded, they are 3.65% and 0.09%, respectively. As a result, the discrepancies between the estimated $S\hat{E}_B(A)$ computed using the nonparametric two-sample bootstrap and the analytically estimated $S\hat{E}_A(A)$ using the Mann-Whitney statistic are quite small, especially for those random bootstrap runs obtained more probably. This validates the nonparametric two-sample bootstrap method in ROC analysis on large datasets from the perspective of relative errors.

7.3 The comparison of the two types of results using hypothesis testing

The estimated 95% CÎs shown in Table 2 were all calculated using the quantile method as described in Section 4. They can also be computed by using "mean ± 1.96 × $S\hat{E}_B(SE)$" where $S\hat{E}_B(SE)$ is shown in Table 2, assuming that the probability distribution of the bootstrap estimated $S\hat{E}_B(A)$ of AUC is normal. These two types of 95% CÎs are matched at least up to the fifth decimal place for all 14 algorithms. For instance, for Algorithm 3, the 95% CÎ derived from the quantile method is (0.0001297, 0.0001377), while it is (0.0001296, 0.0001376) based on the assumption of normality. This suggests that the probability distributions of the bootstrap estimated $S\hat{E}_B(A)$ of AUC for each matching algorithm can be assumed to be normal [1]. The normality of the distribution can also be seen in Figure 3.

As a result, the one-algorithm hypothesis testing can be carried out on each of 14 matching algorithms to determine whether the difference between the estimated mean of the distribution of the bootstrap estimated $S\hat{E}_B(A)$ of AUC and the analytical $S\hat{E}_A(A)$, which is assumed to be a hypothesized value, is statistically significant. It was found that the two-tailed *p*-values of Algorithms 1, 7, and 11 (the three outliers) were close to zero, whereas those of Algorithms 8 and 9 were about 20% and all others were greater than 70%. This is consistent with the observations in Table 2, where the analytical $S\hat{E}_A(A)$ falls outside the estimated 95% CÎ for Algorithms 1, 7, and 11, between the 68.27% CÎ and the 95% CÎ for

Algorithms 8 and 9, and inside the 68.27% CÎ for all other algorithms. Hence, in general the difference is not statistically significant. And this validates the nonparametric two-sample bootstrap method in ROC analysis on large datasets from the perspective of hypothesis testing.

8 Conclusions and discussion

In order to validate the nonparametric two-sample bootstrap in ROC analysis on large datasets, the estimated SÊ of AUC was computed analytically using the Mann-Whitney statistic if the trapezoidal rule is employed; it was also calculated numerically using the nonparametric two-sample bootstrap method. The analytical approach is a deterministic process, and thus its estimated $S\hat{E}_A(A)$ is unique and treated as the ground truth. However, the bootstrap method is a stochastic process, and thus its estimators of $S\hat{E}_B(A)$ constitute a probability distribution.

The comparisons between the probability distribution of the bootstrap estimated SÊs of AUC and the unique analytically estimated SÊ were carried out using relative errors as well as the one-algorithm hypothesis testing. To take the variance of such a distribution into consideration, the estimated mean, median, 68.27% CÎ, and 95% CÎ of the distribution of estimated $S\hat{E}_B(A)$ of AUC were compared with the analytical $S\hat{E}_A(A)$. And the 14 different datasets generated by 14 image matching algorithms were taken as examples.

It was found from all these analyses that the discrepancies between the bootstrap estimated $S\hat{E}_B(A)$ and the analytically estimated $S\hat{E}_A(A)$ are quite small, especially for those random bootstrap runs obtained more probably. Thus the bootstrap results match the analytical result very well.

As a consequence, this validates the nonparametric two-sample bootstrap method for computing the uncertainties of measures in ROC analysis on large datasets. As pointed out in Section 1, the analytical method is not appropriate and the nonparametric two-sample bootstrap method must be employed in many cases while computing the uncertainties of measures. Therefore, this validation provides a sound foundation for applying the bootstrap method to computing uncertainties of measure in the cases where the analytical method cannot be used.

As shown in Section 7.1, the randomness of resampling scores may be affected by a huge stand-alone peak of score distributions. The objective of creating such a peak at the lowest (and/or highest) score is to separate the distributions of genuine scores and impostor scores as far as possible so as to increase the matching accuracy [1, 12]. Nevertheless, the worst relative errors (up to 15.60%), as shown in Table 3, occur when such peaks are found. Though they are relatively large in comparison with others in the table, these relative errors are acceptable in real numerical computation.

For considerations of time complexity, the nonparametric two-sample bootstrap algorithm to compute the SE of the AUC may be partitioned as follows: 1) Randomly resample WR the raw genuine scores and impostor scores, respectively; 2) Sort the two new score sets; 3) Compute the two new discrete probability distributions of the scores; 4) Calculate the

bootstrap replication of AUC; 5) Repeat B = 2000 times and then estimate the SE of the AUC. The average running time of the quicksort algorithm is O(n log n), where n is the length of the data array. However, the time spent in Parts 3 and 4 depends also on the total number of possible scores in the scoring method after converting them to integers, i.e., the range from s_{min} to s_{max} (see Section 2).

Different algorithms employ different scoring methods using integers or real numbers in different ranges, which can be seen in Figure 4. And to obtain more accurate computation, as many score digits as possible were kept while converting to integers. Therefore, the total running time of computing the SE of the AUC is algorithm-dependent. For the 14 algorithms employed in this article, the average running time was 149 seconds; the fastest was 72 seconds; and the slowest was 219 seconds. Further, as discussed in Section 1, sample sizes larger than the ones used in this paper are not needed in order to guarantee the computation accuracy [18]. Therefore, in terms of running time, the bootstrap method is certainly feasible and reliable.

All the tests performed in this article were on large datasets with tens or hundreds of thousands of genuine scores and of impostor scores. A simple test on small medical datasets from Ref. [14] was also conducted, in which there were only 54 genuine scores and 58 impostor scores for both Modality 1 and Modality 2. The test was based on a random run of the bootstrap method rather than on generating a distribution of estimated $\hat{SE}_B(A)$. However, the number of bootstrap replications was set to be 2000, as discussed in Sections 1 and 4. For Modality 1, the estimated $\hat{AUC}$ was 0.882822, the analytical $\hat{SE}_A(A)$ was 0.032606, and the bootstrap $\hat{SE}_B(A)$ was 0.031943. Thus, the relative error was 2.03%. For Modality 2, they were 0.930236, 0.026434, and 0.025059, respectively. Hence, the relative error was 5.20%. These are small relative errors.

References

1. Wu JC, Martin AF, Kacker RN. Measures, uncertainties, and significance test in operational ROC analysis. J Res Natl Inst Stand Technol. 2011; 116(1):517–537. [PubMed: 26989582]
2. Wu JC, Martin AF, Greenberg CS, Kacker RN, Stanford VM. Significance test with data dependency in speaker recognition evaluation. Proceedings of SPIE. 2013; 8734:87340I.
3. Linnet K. Comparison of quantitative diagnostic tests: type I error, power, and sample size. Statistics in Medicine. 1987; 6:147–158. [PubMed: 3589244]
4. Wu JC, Martin AF, Kacker RN. Bootstrap variability studies in ROC analysis on large datasets. Communications in Statistics – Simulation and Computation. 2014; 43(1):225–236.
5. Efron B. Bootstrap methods: Another look at the Jackknife. Ann Statistics. 1979; 7:1–26.
6. Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
7. Hall P. On the number of bootstrap simulations required to construct a confidence interval. Ann Statist. 1986; 14(4):1453–1462.
8. Efron B. Better bootstrap confidence intervals. J Amer Statist Assoc. 1987; 82(397):171–185.
9. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning. 2009; 77:103–123.
10. Flach P, Hernandez-Orallo J, Ferri C. A coherent interpretation of AUC as a measure of aggregated classification performance. Proceedings of the 28th International Conference on Machine Learning. 2011:657–664.
11. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27:861–874.

12. Wu JC, Wilson CL. Nonparametric analysis of fingerprint data on large data sets. Pattern Recognition. 2007; 40(9):2574–2584.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
14. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:839–843. [PubMed: 6878708]
15. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psych. 1975; 12:387–415.
16. Noether, GE. Elements of nonparametric statistics. New York: John Wiley and Sons; 1967. p. 31-32.
17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44:837–845. [PubMed: 3203132]
18. Wu JC, Wilson CL. An empirical study of sample size in ROC-curve analysis of fingerprint data. Proceedings of SPIE. 2006; 6202:620207.
19. Wu JC, Martin AF, Greenberg CS, Kacker RN. Uncertainties of measures in speaker recognition evaluation. Proceedings of SPIE. 2011; 8040:804008.
20. Bandos AI, Rockette HE, Gur D. Resampling methods for the area under the ROC curve. Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning. 2006:1–8.
21. van der Waerden, BL. Mathematical Statistics. Berlin: Springer; 1969. p. 274p. 333-335.
22. Hyndman RJ, Fan Y. Sample quantiles in statistical packages. American Statistician. 1996; 50:361–365.

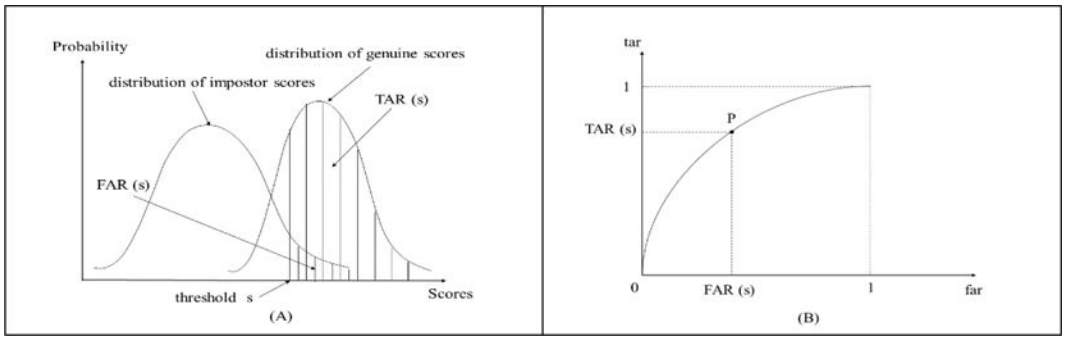

Figure 1.

(A): A schematic diagram of two distributions of continuous genuine scores and impostor scores, showing three related variables: TAR, FAR, and threshold. **(B):** A schematic drawing of an ROC curve constructed by moving the threshold from the highest score down to the lowest one.

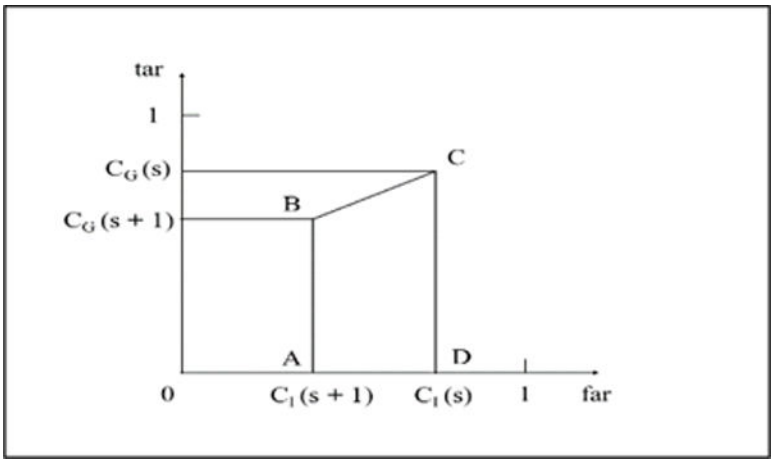

Figure 2.

A schematic drawing of a trapezoid at a score s formed by four points A, B, C, and D along with their coordinates in the far-tar coordinate system.

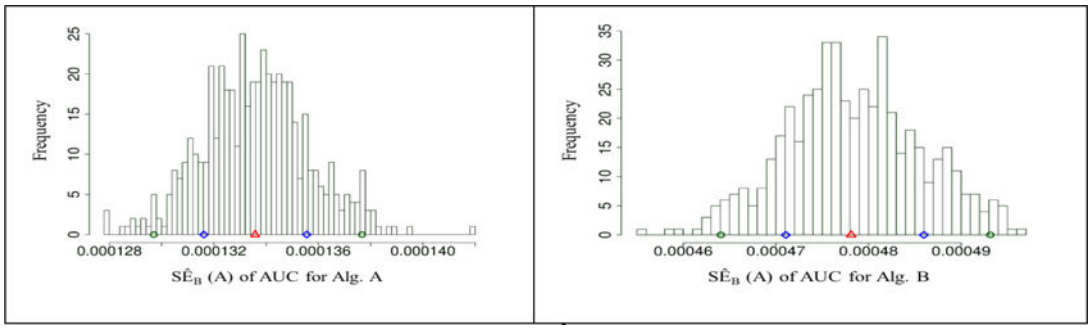

Figure 3.

The distributions of 500 bootstrap estimated $S\hat{E}_B(A)$ of AUC for high-accuracy Algorithm A (L) and low-accuracy Algorithm B (R). The red triangle stands for the analytical result, the blue diamonds are the two bounds of the 68.27% CI, and the green circles represent the two bounds of the 95% CI.

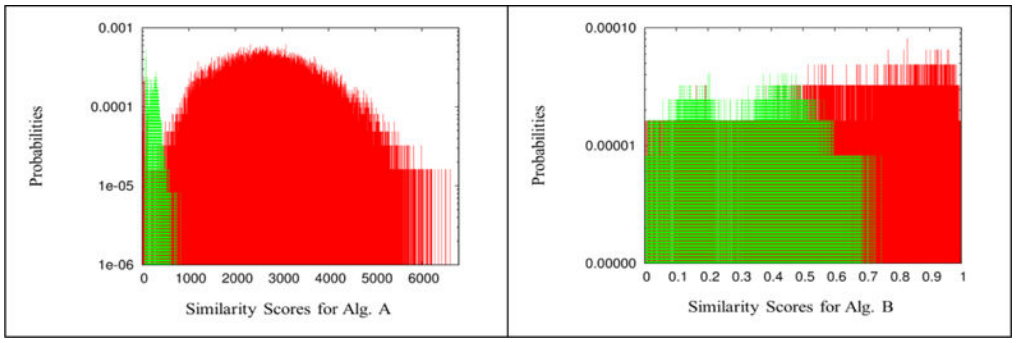

Figure 4.
Distributions of genuine scores (red) and impostor scores (green) for Algorithms A (left) and B (right).

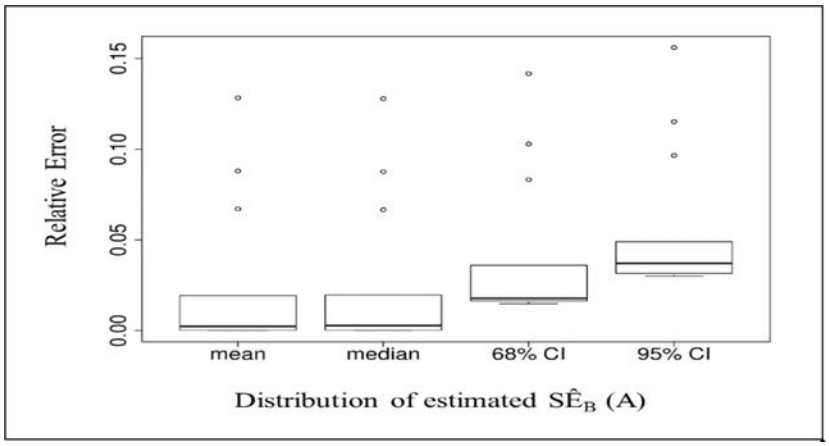

Figure 5.

The box diagrams of 14 relative errors of the estimated mean, median, 68.27% CÎ, and 95% CÎ of the probability distribution of $\hat{SE}_B(A)$, respectively. There are three outliers in each scenario.

Table 1

High-accuracy Algorithm A's and low-accuracy Algorithm B's minimum, maximum, and range of L bootstrap estimated $\hat{SE}_B(A)$ of AUC, while the number of bootstrap replications B was set to be 2000.

$\hat{SE}_B(A)$		Number of Iterations L				
		100	**200**	**300**	**400**	**500**
Alg. A	Min.	0.0001289	0.0001288	0.0001278	0.0001262	0.0001279
	Max.	0.0001393	0.0001413	0.0001395	0.0001385	0.0001418
	Range	0.0000104	0.0000125	0.0000118	0.0000123	0.0000140
Alg. B	Min.	0.0004595	0.0004560	0.0004560	0.0004574	0.0004560
	Max.	0.0004978	0.0004978	0.0005011	0.0004984	0.0004963
	Range	0.0000383	0.0000418	0.0000451	0.0000410	0.0000403

Table 2

The estimated AÛC, the analytical $S\hat{E}_A(A)$, and the estimated mean, median, 68.27% CÎ, 95% CÎ, and $S\hat{E}_B(SE)$ of the distribution of the bootstrap estimated $S\hat{E}_B(A)$ for 14 matching algorithms.

Alg.	AÛC	$S\hat{E}_A(A)$	Distribution of estimated $S\hat{E}_B(A)$				
			Mean	Median	68.27% CÎ	95% CÎ	$S\hat{E}_B(SE)$ (10^{-6})
1	0.9985568	0.0001242	0.0001083	0.0001083	(0.0001066, 0.0001100)	(0.0001048, 0.0001116)	1.71075
2	0.9982568	0.0001231	0.0001231	0.0001231	(0.0001209, 0.0001251)	(0.0001193, 0.0001274)	2.08399
3	0.9982322	0.0001336	0.0001336	0.0001336	(0.0001316, 0.0001356)	(0.0001297, 0.0001377)	2.05723
4	0.9973597	0.0001463	0.0001465	0.0001466	(0.0001442, 0.0001489)	(0.0001422, 0.0001507)	2.26842
5	0.9967486	0.0001695	0.0001695	0.0001695	(0.0001668, 0.0001723)	(0.0001641, 0.0001752)	2.78157
6	0.9943234	0.0002472	0.0002464	0.0002463	(0.0002427, 0.0002505)	(0.0002373, 0.0002541)	4.11347
7	0.9939199	0.0002670	0.0002435	0.0002436	(0.0002395, 0.0002473)	(0.0002362, 0.0002517)	3.88214
8	0.9929374	0.0002579	0.0002530	0.0002528	(0.0002486, 0.0002572)	(0.0002457, 0.0002607)	4.00288
9	0.9923011	0.0002656	0.0002605	0.0002606	(0.0002564, 0.0002645)	(0.0002526, 0.0002682)	4.05652
10	0.9914864	0.0002742	0.0002728	0.0002726	(0.0002685, 0.0002770)	(0.0002636, 0.0002815)	4.52074
11	0.9846023	0.0003928	0.0003664	0.0003666	(0.0003601, 0.0003725)	(0.0003548, 0.0003784)	6.01951
12	0.9845747	0.0004343	0.0004341	0.0004342	(0.0004279, 0.0004404)	(0.0004206, 0.0004480)	6.55578
13	0.9818637	0.0003910	0.0003912	0.0003914	(0.0003847, 0.0003974)	(0.0003781, 0.0004024)	6.21827
14	0.9729011	0.0004781	0.0004783	0.0004779	(0.0004711, 0.0004860)	(0.0004641, 0.0004931)	7.31221

Table 3

The relative errors (%) δ_{mean}, δ_{median}, δ_{68}, and δ_{95} using the estimated mean, median, 68.27% CÎ, and 95% CÎ of the probability distribution of SÊB(A), respectively, for 14 matching algorithms.

Alg.	Relative Errors (%) of SÊB(A)			
	δ_{mean}	δ_{median}	δ_{68}	δ_{95}
1	12.83	12.79	14.17	15.60
2	0.05	0.02	1.74	3.55
3	0.02	0.03	1.48	3.06
4	0.14	0.21	1.75	3.00
5	0.03	0.03	1.66	3.37
6	0.34	0.38	1.83	3.99
7	8.80	8.76	10.29	11.51
8	1.91	1.97	3.60	4.74
9	1.93	1.90	3.47	4.91
10	0.53	0.61	2.08	3.88
11	6.71	6.66	8.32	9.65
12	0.05	0.04	1.48	3.16
13	0.04	0.09	1.62	3.30
14	0.04	0.05	1.64	3.14

Table 4

The estimated mean and median of 14 relative errors (%) in each scenario if three outliers are included.

Include three outlines	Relative Errors (%) of $\hat{SE}B(A)$			
	δ_{mean}	δ_{median}	δ_{68}	δ_{95}
Mean	2.39	2.39	3.94	5.49
Median	0.24	0.30	1.79	3.71

Table 5

The estimated mean and median of 11 relative errors (%) in each scenario if three outliers are excluded.

Exclude three outlines	Relative Errors (%) of $\hat{SEB}(A)$			
	δ_{mean}	δ_{median}	δ_{68}	δ_{95}
Mean	0.46	0.48	2.03	3.65
Median	0.05	0.09	1.74	3.37