NISTIR 7733

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

Jin Chu Wu
Alvin F. Martin
Raghu N. Kacker

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7733

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

Jin Chu Wu
Alvin F. Martin
Raghu N. Kacker

October 2010

U.S. Department of Commerce
Gary Locke, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Director

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

Jin Chu Wu[*a], Alvin F. Martin[a] and Raghu N. Kacker[b]
[a]Information Access Division, [b]Applied and Computational Mathematics Division,
Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract – Sampling variability results in uncertainties of measures. The nonparametric two-sample bootstrap method has been used to compute uncertainties of measures in receiver operating characteristic (ROC) analysis on large datasets, such as the standard error (SE) of the equal error rate in biometrics, the SE of a detection cost function in speaker recognition evaluation, and others. Specifically, the SE of the area under ROC curve (AURC) can be computed analytically using the Mann-Whitney statistic. It can also be calculated using the nonparametric two-sample bootstrap method. The analytical result could be treated as a ground truth. The relative errors of bootstrap-method results with respect to the analytical-method results using different matching algorithms were examined, and they were quite small. Hence, this validates the nonparametric two-sample bootstrap method applied in ROC analysis on large datasets.

Index Terms -- ROC analysis, bootstrap, area under ROC curve, uncertainty, standard error, biometrics, speaker recognition.

[*] Tel: + 301-975-6996; fax: + 301-975-5287. E-mail address: jinchu.wu@nist.gov.

1 Introduction

Sampling variability results in uncertainties for all measurements. That is, multiple sample sets are collected under the similar conditions, and then statistical measures will vary across. Indeed, the measurement uncertainty quantification is a very important issue. Hence, when evaluating and comparing the performance of algorithms, the measurement uncertainties must be taken into account [1-3].

Receiver operating characteristic (ROC) analysis is an important statistical technique in many areas. The nonparametric two-sample bootstrap method has been used to compute uncertainties of measures in operational ROC analysis on large datasets, such as the standard error (SE) of the equal error rate (EER) in biometrics, the SE of a detection cost function in speaker recognition evaluation, etc., based on our extensive bootstrap variability studies on large datasets [2]. The detection cost function is defined as a weighted sum of probabilities of type I error and type II error [4]. It has been hard to calculate uncertainties of these statistics of interest without using the two-sample bootstrap method [3].

The area under an ROC curve (AURC) is an important metric in ROC analysis [5, and references therein]. The AURC corresponds to the probability of correctly identifying which of the two stimuli is more likely than the other. It measures the overall ROC curve rather than the performance at a particular operational point on the ROC curve. Moreover, if it is computed using the trapezoidal rule, the AURC is equivalent to the Mann-Whitney statistic that is formed by matching scores, namely, genuine and impostor scores in biometrics, or target and non-target scores in speaker recognition evaluation, etc. Hence, the variance of the Mann-Whitney statistic can be utilized as the variance of the AURC. In other words, the SE of AURC can be computed analytically. This analytical approach is a deterministic process and thus the result is unique.

The Mann-Whitney statistic is asymptotically normally distributed, regardless of the distributions of matching scores, thanks to the Central Limit Theorem. Thus, the Z statistic formulated in terms of two AURCs along with their SEs and correlation coefficient is subject to the standard normal distribution and can be used to test the significance of the difference of two ROC curves [6, 7, and references therein].

On the other hand, the SE of AURC can also be calculated using the nonparametric two-sample bootstrap method [1-3, 8-10]. Unlike the analytical approach using the Mann-Whitney statistic, the bootstrap method is a stochastic process. In other words, the result will change for different runs. Thus, rather than dealing with a single measure of the SE of AURC in the analytical approach, the results derived from the bootstrap method constitute a probability distribution. Some results may be more probable and others less. However, if the analytical result is treated as the ground truth and if the relative errors of the bootstrap results with respect to the analytical results are not large, this can validate nonparametric two-sample bootstrap method used in computing the uncertainties of some statistics of interest on large datasets, which cannot be calculated otherwise.

The nonparametric two-sample bootstrap method is particularly of interest in operational ROC analysis on large datasets. The two samples are referred to as a set of genuine (i.e., target) scores

and a set of impostor (i.e., non-target) scores. They constitute two distributions. An ROC curve is characterized by the relative relationship between these two distributions [5, 11]. These two distribution functions are indeed interrelated by the algorithm that generates them. In other words, the performance of a matching algorithm is affected not only by genuine matching but also by impostor matching. All statistics of interest in ROC analysis are influenced by the combined impact of these two samples.

Furthermore, it was shown in our previous studies that 1) these two distributions usually do not have well defined parametric forms; 2) the shapes of these two distributions may be considerably different for the same algorithm; and 3) the distributions may vary substantially from algorithm to algorithm in a way that differentiates algorithms in terms of matching accuracy [5]. This suggests that the nonparametric statistical analysis be appropriate for analyzing such data. Thus, the empirical distribution is assumed for each of the observed scores.

As is well-known, the bootstrap method assumes that an independent and identically distributed (i.i.d.) random sample of size n is drawn from a population with its own probability distribution. Our large government data bases used for developing similarity scores in fingerprint technology were randomly collected from real practice rather than using multiple acquisitions and thus had no dependencies. Thus, the random sample is assumed to be i.i.d. in our work.

With the i.i.d. assumption, the objects of a nonparametric two-sample bootstrap are individuals in the sample [2, 3]. Otherwise, the bootstrap objects are the subsets of the sample into which the sample is grouped based on data dependencies caused by multiple biometric acquisitions [12, 13]. This can preserve the dependencies among the data. However, everything else in the bootstrap method remains intact. Of course, how the sample is grouped into subsets will have impact on the bootstrap results. As a matter of fact, from the statistical point of view, the sample should be collected as randomly as possible in test design.

The number of bootstrap replications is a very important parameter in bootstrap method. In order to reduce the bootstrap variance and ensure the accuracy of the computation in our applications where the size of data samples is large, the statistics of interest are probabilities, and no normality assumption can be made for distributions of similarity scores, the bootstrap variability was empirically studied extensively [2, 14]. As a result of our study, the appropriate number of bootstrap replications was determined to be 2000 in our applications.

In this article, 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 in fingerprint applications, if the numbers of similarity scores get larger than these, the measurement accuracy will improve little [15]. The research was carried out by applying Chebyshev's inequality to two metrics: the AURC and the true accept rate (TAR) at an operational false accept rate (FAR). All similarity scores were converted to integers if they were not already. Hence, the probability distribution functions of the similarity scores were all discrete, and thus the ROC curve was not a smooth curve [5].

The analytical method using the Mann-Whitney statistic to compute the estimated $\hat{SE}_A$ (A) of AURC along with the formulations of discrete distribution functions of genuine scores and

impostor scores is shown in Section 2. The algorithm of the nonparametric two-sample bootstrap method for calculating the estimated $\hat{SE}_B$ (A) of AURC and how to generate a probability distribution of $\hat{SE}_B$ (A) are provided in Section 3. The relative errors used for comparison are presented in Section 4. The results of the analytical method and the results of the bootstrap method, as well as a comparison of these two types of results, are offered in Section 5, involving 14 different fingerprint-image matching algorithms[1] used as examples. Finally, conclusions and discussion can be found in Section 6.

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

It is assumed that the trapezoidal rule is employed while computing AURC, and thus the AURC is equivalent to the Mann-Whitney statistic directly formed from the discrete genuine and impostor scores. Further, the variance of the Mann-Whitney statistic can be computed analytically. Hence, it can be utilized as the variance of AURC [5, and references therein]. First are the formulations of distribution functions.

2.1 The formulations of discrete distribution functions of genuine and impostor scores

All similarity scores were converted into integers if they were not, as mentioned in Section 1. Thus, without loss of generality, the similarity scores generated by an algorithm are expressed inclusively using the integer score set $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$, running consecutively from the lowest score s_{min} to the highest score s_{max}.

The genuine score set is denoted as

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

where N_G is the total number of genuine scores. And the impostor score set is expressed as

$$\mathbf{I} = \{ n_i \mid n_i \in \{s\} \text{ and } i = 1, \ldots, N_I\} , \quad (2)$$

where N_I is the total number of impostor scores.

These two sets of similarity scores constitute two discrete probability distribution functions, respectively. Let P_i (s), where $s \in \{s\}$ and $i \in \{G, I\}$, denote the empirical probabilities of the genuine scores and the impostor scores at a score s, respectively. It may very well be that some of them are zeroes at some scores in the set {s}. Nonetheless, the two distribution functions can be expressed, respectively, as

$$\boldsymbol{P_i} = \{ P_i (s) \mid \forall\, s \in \{s\} \text{ and } \sum_{\tau=s\,min}^{s\,max} P_i (\tau) = 1 \} , i \in \{G, I\} . \quad (3)$$

The cumulative discrete probability distribution functions of genuine scores and impostor scores are defined in this article to be the probabilities cumulated from the highest score s_{max} down to the integer score s, and are expressed as

[1] Specific hardware and software products identified in this report 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.

$$\boldsymbol{C_i} = \{ C_i(s) = \sum_{\tau=s}^{s\max} P_i(\tau) \mid \forall\, s \in \{s\} \} \,,\ i \in \{G, I\} \,, \tag{4}$$

where C_i (s), $i \in \{G, I\}$, are the cumulative probabilities of genuine scores and impostor scores at a score s, respectively.

2.2 Compute the estimated AÛRC

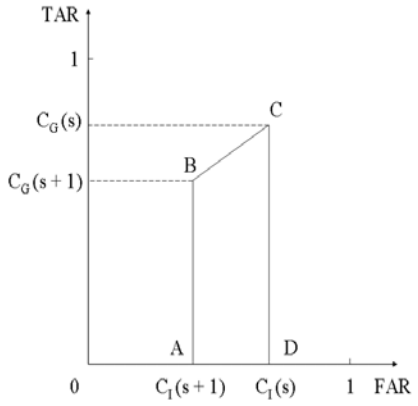

Figure 1 A schematic drawing of four points A, B, C, and D along with their coordinates in the FAR-and-TAR coordinate system. They form a trapezoid at a score s, and BC is a segment of an ROC curve.

After conversion of similarity scores to integers, the distributions of genuine scores and impostor scores are all discrete. As a result, the ROC curve is no longer a smooth curve. While cumulating probabilities of genuine scores and impostor scores from the highest similarity score, respectively, an ROC curve can go horizontally, vertically, inclined toward upper right, or stay where it is for each decrement of score, depending on whether $P_I(s)$ and/or $P_G(s)$ are nonzero or not. Thus, the AURC consists of a set of trapezoids, each of which is built by a rectangle and a triangle in general. The trapezoid can be reduced to a rectangle, a vertical line, or a point.

Without loss of generality, a trapezoid is shown in Figure 1. In the FAR-and-TAR coordinate system, at a score $s \in \{s\}$, by including zero-frequency scores, a trapezoid 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), in clockwise direction, assuming C_I (s_{max} + 1) = C_G (s_{max} + 1) = 0. This boundary condition corresponds to the origin of the FAR-and-TAR coordinate system, and will be applied throughout the following discussion. 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., $\sum_{\tau=s+1}^{s\max} P_G(\tau)$) create a rectangle. As a consequence, the estimated AÛRC can be calculated as,

$$\begin{aligned}\hat{A} &= \sum_{s=s\max}^{s\min} \text{trapezoid}(s) \\ &= \sum_{s=s\max}^{s\min} \text{triangle}(s) + \sum_{s=s\max}^{s\min} \text{rectangle}(s) \\ &= \sum_{s=s\max}^{s\min} P_I(s) \times \left[\frac{1}{2} \times P_G(s) + \sum_{\tau=s+1}^{s\max} P_G(\tau) \right]\end{aligned} \tag{5}$$

Note that the summation runs consecutively in the descending order from s_{max} to s_{min}, including zero-frequency scores, and $\sum_{\tau=s\max+1}^{s\max} = 0$ is assumed according to the above boundary condition. This notation will be applied throughout the following discussion.

2.3 Relate AURC to the Mann-Whitney statistic

In order to relate AURC to the Mann-Whitney statistic, the order relations among similarity scores are established as follows. All the N_I scores in the impostor score set **I** in Eq. (2) are compared with all the N_G scores in the genuine score set **G** in Eq. (1). It counts 1, ½, or zero depending whether an impostor score s_I is less than, equal to, or greater than a genuine score s_G. This rule can be expressed as

$$\mathbf{R}(s_G, s_I) = \begin{cases} 1 & \text{if } s_I < s_G \\ ½ & \text{if } s_I = s_G \\ 0 & \text{if } s_I > s_G \end{cases} \tag{6}$$

After converting probabilities of genuine and impostor scores in Eq. (5) back to frequencies and by including zero-frequency scores, the first term in Eq. (5) shows the total number of score pairs in which the impostor score is equal to the genuine score, weighted by ½ and divided by $N_G N_I$. And the second term in Eq. (5) represents the total number of score pairs in which the impostor score is less than the genuine score, weighted by 1 and divided by $N_G N_I$. This term is the so called "the number of inversions" in a sequence formed by impostor and genuine scores [16].

Finally, the estimated AÛRC can be re-written as

$$\hat{A} = \frac{1}{N_G N_I} \times \sum_{s_G=1}^{N_G} \sum_{s_I=1}^{N_I} \mathbf{R}(s_G, s_I) \tag{7}$$

Except for the coefficient, this is exactly the Mann-Whitney statistic formed by the genuine and impostor scores. As a consequence, the variance of AURC can be obtained by computing the variance of the Mann-Whitney statistic.

2.4 Compute the estimated $S\hat{E}_A$ (A) of AURC

The variance of the Mann-Whitney statistic can be computed analytically and it is utilized as the variance of AURC. To do so, two more cumulative probability distribution functions are required. One is

$$\boldsymbol{Q_G} = \{\, Q_G(s) = \sum_{\tau=s+1}^{s\max} P_G(\tau) \mid \forall\, s \in \{s\} \,\}\,. \tag{8}$$

The other one is

$$\boldsymbol{Q_I} = \{\, Q_I(s) = \sum_{\tau=s\min}^{s-1} P_I(\tau) \mid \forall\, s \in \{s\} \,\} \tag{9}$$

where another boundary condition $\sum_{\tau=s\min}^{s\min-1} = 0$ is assumed. Note that the cumulation of probabilities is taken place from s_{max} down to s + 1 with respect to genuine scores in Eq. (8), and from s_{min} up to s – 1 on impostor scores in Eq. (9).

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

$$B_{GGI} = \sum_{s=s\min}^{s\max} P_I(s) \times [Q_G^2(s) + Q_G(s) \times P_G(s) + \frac{1}{3} \times P_G^2(s)\,] \tag{10}$$

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

$$B_{IIG} = \sum_{s=s\min}^{s\max} P_G(s) \times [Q_I^2(s) + Q_I(s) \times P_I(s) + \frac{1}{3} \times P_I^2(s)\,] \tag{11}$$

Finally, the analytical estimator of SE of AURC can be computed as

$$\hat{SE}_A(A) = \text{square root}\,\{\, \frac{1}{N_G N_I} \times [\, \hat{A}(1-\hat{A}) + (N_G - 1)(B_{GGI} - \hat{A}^2) + (N_I - 1)(B_{IIG} - \hat{A}^2)\,]\,\} \tag{12}$$

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

3.1 The algorithm of the nonparametric two-sample bootstrap [1-3, 8-10]

The estimated uncertainties in terms of SE and 95 % confidence interval (CI) can also be computed using the nonparametric two-sample bootstrap. Assuming the data set is i.i.d., the bootstrap objects are individuals in the data set, rather than subsets of the sample into which the sample data are grouped according to data dependencies, as mentioned in Section 1. With such an assumption, the algorithm of the nonparametric two-sample bootstrap is as follows.

***Algorithm I* (Nonparametric two-sample bootstrap)**

1: **for** i = 1 **to** B **do**
2: select N_G scores randomly WR from **G** to form a set $\{\text{new } N_G \text{ genuine scores}\}_i$
3: select N_I scores randomly WR from **I** to form a set $\{\text{new } N_I \text{ impostor scores}\}_i$
4: $\{\text{new } N_G \text{ genuine scores}\}_i$ & $\{\text{new } N_I \text{ impostor scores}\}_i$ => statistic $\hat{A}_i$
5: **end for**
6: $\{\hat{A}_i \mid i = 1, ..., B\} \Rightarrow \hat{SE}_B$ 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". The original genuine score set **G** with N_G scores shown in Eq. (1) and the original impostor score set **I** with N_I scores shown in Eq. (2) are generated by a matching algorithm. 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 original genuine score set **G** to form a new set of N_G genuine scores, N_I scores are randomly selected WR from the original impostor score set **I** to form a new set of N_I impostor scores, and then in Step 4 from these two new sets of similarity scores the i-th bootstrap replication of the estimated AÛRC, i.e., $\hat{A}_i = A\hat{U}RC_i$, is generated using Eq. (5).

Finally, after B iterations, as indicated in Step 6, from the set $\{\hat{A}_i \mid i = 1, ..., B\}$, the estimator of the SE, denoted by $S\hat{E}_B$, i.e., the sample standard deviation of the B replications, and the estimators of the $\alpha/2$ 100 % and (1 - $\alpha/2$) 100 % quantiles of the bootstrap distribution, denoted by $\hat{Q}_B(\alpha/2)$ and $\hat{Q}_B(1-\alpha/2)$, at the significance level α can be calculated [10]. The Definition 2 of quantile in Ref. [17] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. Thus, $(\hat{Q}_B(\alpha/2), \hat{Q}_B(1-\alpha/2))$ stands for the estimated bootstrap (1 - α) 100 % CÎ. If 95 % CÎ is of interest, then α is set to be 0.05.

The remaining issue is to determine how many iterations this bootstrap algorithm needs to run, i.e., what the number of the nonparametric two-sample bootstrap replications is, in order to reduce the bootstrap variance and ensure the accuracy of the computation. As stated in Section 1, based on our empirical bootstrap variability studies, the appropriate number of bootstrap replications B in our applications was determined to be 2000 [2, 14].

3.2 Generate a probability distribution of $S\hat{E}_B$ (A)

The analytical approach of computing the estimated $S\hat{E}_A$ (A) of AURC using the Mann-Whitney statistic is a deterministic process, and thus the analytical solution is unique. On the other hand, as pointed out in Section 1, the bootstrap method is a stochastic process. It can generate different results from different runs, and some results may be more probable and others less. Hence, the bootstrap estimators $S\hat{E}_B$ (A) of AURC constitute a probability distribution. Here is an Algorithm for generating such a distribution.

Algorithm II **(Generating a probability distribution)**

1: **for** i = 1 **to** L **do**
2: **for** j = 1 **to** B **do**
3: select N_G scores randomly WR from **G** to form a set {new N_G genuine scores}$_j$
4: select N_I scores randomly WR from **I** to form a set {new N_I impostor scores}$_j$
5: {new N_G genuine scores}$_j$ & {new N_I impostor scores}$_j$ => statistic $\hat{A}_{ij}$
6: **end for**
7: $\{\hat{A}_{ij} \mid j = 1, \ldots, B\} \Rightarrow S\hat{E}_{Bi}$ (A)
8: **end for**
9: **SE_B (A)** = $\{S\hat{E}_{Bi}$ (A) | i = 1, …, L } => estimators of mean, median, 68.27 % CI, & 95 % CI

10: **end**

where L is the number of Monte Carlo iterations and B is the number of bootstrap replications. As a matter of fact, in Algorithm II, from Step 2 to 7 is the Algorithm I shown in Section 3.1, which computes the i-th $\hat{SE}_{B\,i}$ (A) of AURC using the nonparametric two-sample bootstrap and runs L iterations as indicated in Step 1. As shown in Step 9, L estimated $\hat{SE}_{B\,i}$ (A) of AURC are created and form a set $\mathbf{SE_B\ (A)}$. Subsequently from this set, the estimated mean, median, 68.27 % CI and 95 % CI of the distribution of estimated $\hat{SE}_B$ (A) of AURC can be calculated. The CI can be obtained using the Definition 2 of quantile in Ref. [17].

$\hat{SE}_B$ (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 1 High-accuracy Algorithm A's and low-accuracy Algorithm B's minimum, maximum, and range of L estimated $\hat{SE}_B$ (A) of AURC, where the number of iterations L was set to be from 100 up to 500 at intervals of 100, while the number of bootstrap replications B was set to be 2000.

In Algorithm II, the number of bootstrap replications B was set to be 2000, as discussed above. Then the next question is how to determine the number of iterations L. Two fingerprint-image matching algorithms, high-accuracy A and low-accuracy B, were taken to be examples. The number of iterations L was set to be from 100 up to 500 at intervals of 100. Then the minimum, maximum, and range of L estimated $\hat{SE}_B$ (A) of AURC were calculated and are shown in Table 1. If the accuracy is up to the 5^{th} decimal place, the minimum, maximum, and range of $\hat{SE}_B$ (A) of AURC for high-accuracy Algorithm A across all five different numbers of iterations are rounded to 0.00013, 0.00014, and 0.00001, respectively; and those for low-accuracy Algorithm B are rounded to 0.00046, 0.00050, and 0.00004 (except one), respectively. As a result, the discrepancies among the results from 100 runs to 500 runs are small.

Further, in order to obtain statistically meaningful estimated $\hat{CI}$, the number of $\hat{SE}_B$ (A) of AURC, i.e., the number of iterations L, must be quite large. For instance, in order to obtain 95 % $\hat{CI}$, there are only about two instances located in each end of the distribution for 100 $\hat{SE}_B$ (A), however there are about 12 instances for 500 $\hat{SE}_B$ (A). As a consequence, the number of iterations was set to be 500. In other words, for each matching algorithm, 500 estimated $\hat{SE}_B$ (A) will be generated to constitute a probability distribution, and each of 500 estimators is computed using the nonparametric two-sample bootstrap.

The distribution of estimated $\hat{SE}_B$ (A) of AURC for the matching Algorithm A is shown in Figure 2, where the red triangle stands for the analytical result, the blue diamonds are the two bounds of 68.27 % CI, and the green circles represent the two bounds of 95 % CI. It is also shown in Figure 2 that for Algorithm A the analytical estimator $\hat{SE}_A$ (A) of AURC is very close

to the mean as well as the median of the distribution of 500 estimated $\hat{SE}_B$ (A), which are all approximately equal to 0.0001336 (see Table 2 in Section 5.1 referred to as Algorithm 3).

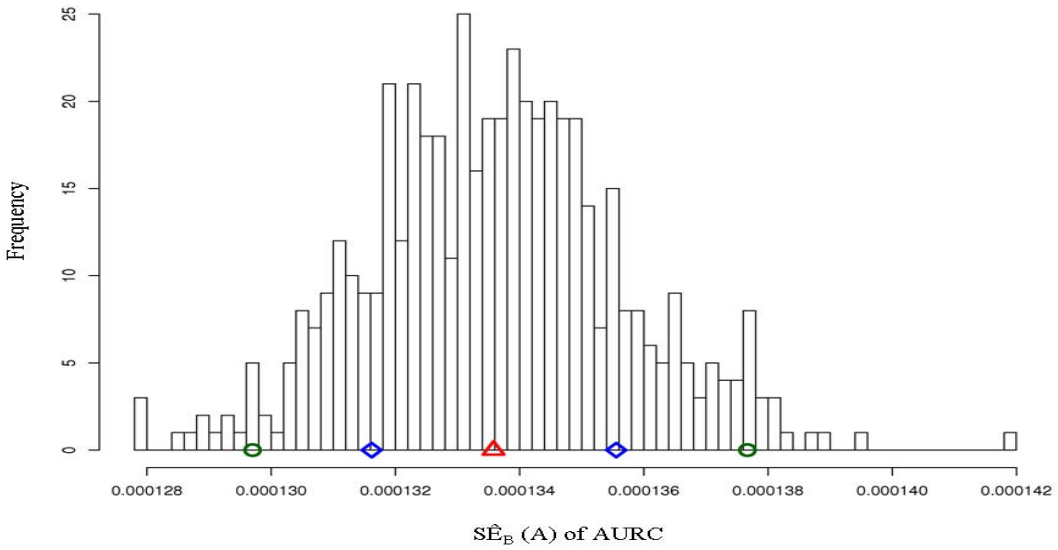

Figure 2 The distribution of 500 estimators of $\hat{SE}_B$ (A) of AURC computed using the nonparametric two-sample bootstrap for the matching Algorithm A. The red triangle stands for the analytical result, the blue diamonds are the two bounds of 68.27 % CI, and the green circles represent the two bounds of 95 % CI.

4 The relative errors used for comparison

While comparing the bootstrap results with the unique analytical estimator $\hat{SE}_A$ (A) of AURC, the comparison is quantified by the relative error in order to take into account the impact of the magnitude of the analytical result. The estimated relative error $\hat{\eta}$ is defined as

$$\hat{\eta} = | \hat{X} - \hat{SE}_A (A) | / \hat{SE}_A (A) \times 100\ \% \quad (13)$$

where $\hat{SE}_A$ (A) is the analytical estimator of SE of AURC computed using Eq. (12) in Section 2.4, and $\hat{X}$ is one of estimated quantities which describe the probability distribution of bootstrap estimated $\hat{SE}_B$ (A) of AURC.

As pointed out in Section 1, the bootstrap method is a stochastic process. While performing the comparisons involving a distribution, it is not enough to just pick *one* bootstrap result from a random run. In order to take account of the variance of stochastic process, not only should the estimated mean and median of the distribution be compared with the analytical result, but the upper bound and lower bound of 68.27 % CI (corresponding to one standard deviation) as well as the two bounds of 95 % CI (corresponding to 1.96 standard deviation) of the distribution should also be compared. While comparing the estimated $\hat{CI}$ with the analytical result $\hat{SE}_A$ (A) of AURC, the larger one between the two relative errors using the upper bound and the lower bound of CI, respectively, will be employed. Notice that with probability about 27 % the bootstrap estimators of SE can fall in between 68.27 % CI and 95 % CI of the estimated $\hat{SE}$s.

The relative errors of the estimated $\hat{SE}_B$ (A) of AURC with respect to the analytical estimated $\hat{SE}_A$ (A) using the estimated mean, median, 68.27 % CÎ, and 95 % CÎ of the distribution of $\hat{SE}_B$ (A) are denoted by $\hat{\eta}_\mu$, $\hat{\eta}_\nu$, $\hat{\eta}_\xi$, and $\hat{\eta}_\zeta$, respectively, in the following text.

5 Results

5.1 The analytical results and bootstrap results

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

Table 2 The estimated AÛRC, the unique analytical $\hat{SE}_A$ (A), and the estimated mean, median, 68.27 % CÎ, and 95 % CÎ of the probability distribution of estimated $\hat{SE}_B$ (A) for 14 matching algorithms. The distribution was generated by 500 runs.

To show both analytical results and bootstrap results, 14 fingerprint-image matching algorithms were taken as examples. The estimated AÛRC, the unique analytical $\hat{SE}_A$ (A), and the estimated mean, median, 68.27 % CÎ, and 95 % CÎ of the probability distribution of estimated $\hat{SE}_B$ (A) for 14 matching algorithms are shown in Table 2. The distribution was generated by 500 runs. Some matching algorithms are of relatively high accuracy and some are of relatively low accuracy, as indicated by their estimated AÛRC. The larger the estimated AÛRC is, the more accurate the

matching algorithm is [5, and references therein]. In Table 2, Algorithms 3 and 14 are Algorithms A and B employed in Section 3.2, respectively.

In order to show the difference, seven decimal places were kept. Indeed, in our real computation, many more decimal places were kept in the intermediate steps of calculations. It is noticed that most analytical estimators $\hat{SE}_A$ (A) fall in the estimated 95 % C$\hat{I}$ of the distributions of $\hat{SE}_B$ (A), except Algorithms 1, 7, and 11. This is related to the characteristics of the distributions of genuine scores and impostor scores.

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 impostor population. For other matching algorithms, if there is a stand-alone peak, it does not occupy larger than 50 % of the population. These extremely large peaks of impostor distribution at the lowest score can cause a very large portion of ROC curve at the top part to be formed by a long straight line segment (i.e., the ROC curve jumps from one point to the next one by a large distance). They might impede bootstrap to function well.

As indicated in Section 3.1, the estimated C$\hat{I}$ in Table 2 were all obtained using the Definition 2 of quantile in Ref. [17]. In the meantime, they can also be computed by assuming that the probability distribution of $\hat{SE}_B$ (A) for each matching algorithm is normal. The estimated S$\hat{E}$s of the distribution of $\hat{SE}_B$ (A) for Algorithms 1 through 14 are 0.00000171, 0.00000208, 0.00000206, 0.00000227, 0.00000278, 0.00000411, 0.00000388, 0.00000400, 0.00000406, 0.00000452, 0.00000602, 0.00000656, 0.00000622, and 0.00000731, respectively.

The estimated 95 % C$\hat{I}$s calculated in these two ways do match at least up to the fifth decimal place. Generally speaking, the more accurate the matching algorithms are, the more decimal places they do match. For example, for high-accuracy Algorithm 2, the estimated 95 % C$\hat{I}$ using the quantile method is (0.0001193, 0.0001274) as shown in Table 2, and the 95 % C$\hat{I}$ assuming normal distribution is (0.0001190, 0.0001272) using the estimated mean 0.0001231 and the estimated S$\hat{E}$ 0.00000208. This indicates that the distributions of the estimated $\hat{SE}_B$ (A) of AURC can be assumed to be normal.

5.2 The comparison of two types of results using relative error

In Table 3 are shown the relative errors (%) $\hat{\eta}_\mu$, $\hat{\eta}_\nu$, $\hat{\eta}_\xi$, and $\hat{\eta}_\zeta$ of $\hat{SE}_B$ (A) with respect to the analytical estimated $\hat{SE}_A$ (A) using the estimated mean, median, 68.27 % C$\hat{I}$, and 95 % C$\hat{I}$ of the distribution of estimated $\hat{SE}_B$ (A) of AURC, respectively, for 14 matching algorithms. The corresponding box diagrams of relative errors of 14 matching algorithms are depicted in Figure 3. It is obvious that there are three outliers that correspond to Algorithms 1, 7, and 11, respectively. This is consistent with the discussion in Section 5.1.

For those random runs using the nonparametric two-sample bootstrap, the results of SEs that would be obtained more probably than others are those at the estimated mean, median, and within the 68.27 % CI of the distribution of estimated $\hat{SE}_B$ (A). As discussed in Section 4, the bootstrap estimators of SE can fall in between 68.27 % CI and 95 % CI with probability about 27

%. In other words, the relative errors $\hat{\eta}_{\mu}$, $\hat{\eta}_{\nu}$, and $\hat{\eta}_{\xi}$, defined in Section 4, may be more probable than the relative error $\hat{\eta}_{\zeta}$.

Alg.	Relative Errors (%) of $\hat{SE}_B$ (A)			
	$\hat{\eta}_{\mu}$	$\hat{\eta}_{\nu}$	$\hat{\eta}_{\xi}$	$\hat{\eta}_{\zeta}$
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 3 Relative errors (%) $\hat{\eta}_{\mu}$, $\hat{\eta}_{\nu}$, $\hat{\eta}_{\xi}$, and $\hat{\eta}_{\zeta}$ of $\hat{SE}_B$ (A) using the estimated mean, median, 68.27 % CÎ, and 95 % CÎ of the distribution of $\hat{SE}_B$ (A), respectively, for 14 matching algorithms.

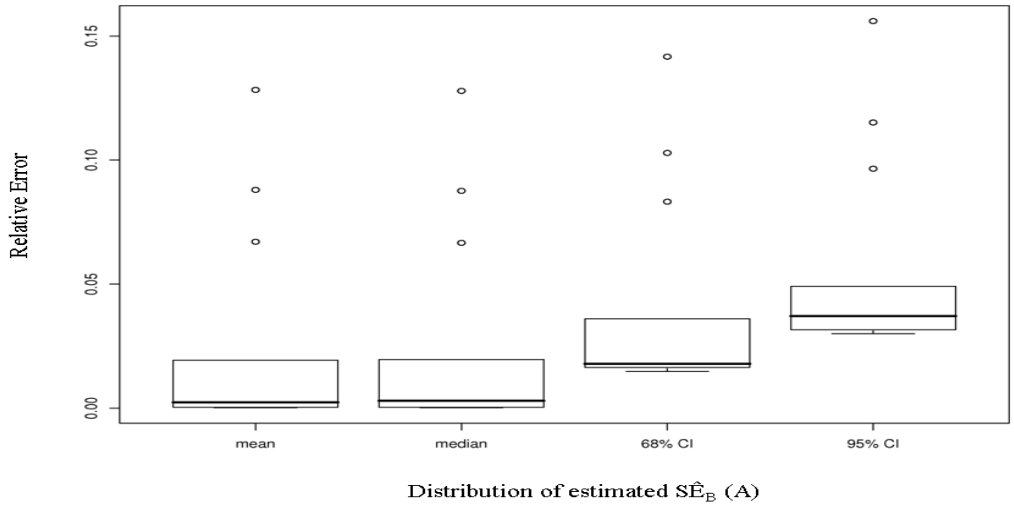

Figure 3 Box diagrams of 14 relative errors of $\hat{SE}_B$ (A) using the estimated mean, median, 68.27 % CÎ, and 95 % CÎ of the distribution of estimated $\hat{SE}_B$ (A), respectively. There are three outliers.

Include three outlines	Relative Errors (%) of $\hat{SE}_B$ (A)			
	$\hat{\eta}_\mu$	$\hat{\eta}_\nu$	$\hat{\eta}_\xi$	$\hat{\eta}_\zeta$
Mean	2.39	2.39	3.94	5.49
Median	0.24	0.30	1.79	3.71

Table 4 The estimated mean and median of 14 relative errors (%) of $\hat{SE}_B$ (A) using the estimated mean, median, 68.27 % CI, and 95 % CI of the distribution of $\hat{SE}_B$ (A), respectively, if three outliers are included.

Exclude three outlines	Relative Errors (%) of $\hat{SE}_B$ (A)			
	$\hat{\eta}_\mu$	$\hat{\eta}_\nu$	$\hat{\eta}_\xi$	$\hat{\eta}_\zeta$
Mean	0.46	0.48	2.03	3.65
Median	0.05	0.09	1.74	3.37

Table 5 The estimated mean and median of 11 relative errors (%) of $\hat{SE}_B$ (A) using the estimated mean, median, 68.27 % CI, and 95 % CI of the distribution of $\hat{SE}_B$ (A), respectively, if three outliers are excluded.

Moreover, it is shown in Figure 3 that all four distributions are skewed. Thus, the median of the distribution is more important than the mean. Hence, the estimated mean and median of 14 relative errors (%) of $\hat{SE}_B$ (A) using the estimated mean, median, 68.27 % CI, and 95 % CI of the distribution of estimated $\hat{SE}_B$ (A), respectively, are shown in Table 4, where three algorithms as outliers are included. Those excluding three outliers are presented in Table 5.

If including three outliers, the worst relative error of $\hat{SE}_B$ (A) is 5.49 % that is related to a bound of the 95 % CI of the distribution, but the median of 14 relative errors $\hat{\eta}_\nu$ using the median of the distribution of estimated $\hat{SE}_B$ (A) for each matching algorithm is 0.30 %. If excluding three outliers, they are 3.65 % and 0.09 %, respectively. As a result, the discrepancies between the estimated $\hat{SE}_B$ (A) computed using the nonparametric two-sample bootstrap and the analytically estimated $\hat{SE}_A$ (A) using the Mann-Whitney statistic are quite small especially for those random bootstrap runs obtained more probably. Subsequently, this validates the two-sample bootstrap method on large datasets.

6 Conclusions and discussion

The estimated $\hat{SE}$ of AURC was computed analytically using the Mann-Whitney statistic if the trapezoidal rule is employed, as well as numerically using the nonparametric two-sample bootstrap method. The analytical approach is a deterministic process, and thus its estimated $\hat{SE}_A$ (A) is unique. However, the bootstrap method is a stochastic process, and thus its estimators of $\hat{SE}_B$ (A) constitute a distribution. In order to take the variance of such a process into consideration, the estimated mean, median, 68.27 % $\hat{CI}$, and 95 % $\hat{CI}$ of the distribution of estimated $\hat{SE}_B$ (A) of AURC are compared with the analytical $\hat{SE}_A$ (A) for each matching algorithm. While comparing an estimated $\hat{CI}$ with the analytical result, the relative error is defined to be the larger one between using the upper bound and the lower bound of $\hat{CI}$.

14 matching algorithms, including three outliers, were taken as examples. Therefore, in each case, i.e., using mean, median, 68.27 % CÎ and 95 % CÎ, respectively, 14 relative errors were generated. The mean and median of such 14 relative errors were created as well. All such means and medians, with or without three outliers, were presented. It was found 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.

As a consequence, this validates the two-sample bootstrap method on large datasets. In the meantime, the nonparametric two-sample bootstrap was carried out with the i.i.d. assumption for the datasets in this article. Hence, it shows again that our large government data bases used for developing similarity scores in fingerprint technology have no dependencies. As a matter of fact, from the statistical point of view, the sample should be collected as randomly as possible in test design.

The one-algorithm hypothesis testing was carried out on each of 14 matching algorithms to determine whether the difference between the estimated mean of the distribution of estimated $S\hat{E}_B$ (A) of AURC and the analytical $S\hat{E}_A$ (A) as a hypothesized value is statistically significant, since the distribution can be assumed to be normal as discussed in Section 5.1 [1]. It was found that the two-tailed p-values of Algorithms 1, 7, and 11 (three outliers) were close to zero, 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 68.27 % CÎ and 95 % CÎ for Algorithms 8 and 9, and inside 68.27 % CÎ for all other algorithms. Hence, generally speaking, the difference is not real.

An extremely large stand-alone peak of distribution of similarity scores, which occupies a very large portion of population, can impede the bootstrap functioning well, as shown in Section 5. This might be because the randomness of resampling similarity scores from such a distribution could be affected by the huge stand-alone peak. The objective of creating such a peak at the lowest (and/or highest) similarity score is to separate the distributions of genuine scores and impostor scores as far as possible so as to increase the matching accuracy [5, 11]. This is one of techniques employed by some matching algorithms. Nevertheless, the worst relative error 15.60 % that is related to a bound of the 95 % CI of the distribution as shown in Table 3 is relatively large in comparison with others in the table, but it is acceptable in real numerical computation.

All the tests performed in this article were on large datasets with tens and hundreds of thousands of genuine scores and impostor scores. A simple test on small medical datasets from Ref. [7] was also conducted, in which there were 54 genuine scores and 58 impostor scores for both Modality 1 and 2. It was based on a random run of bootstrap method rather than generating a distribution of estimated $S\hat{E}_B$ (A). However, the number of bootstrap replications was set to be 2000, as discussed in Section 1. For Modality 1, the estimated $A\hat{U}RC$ was 0.882822, the analytical $S\hat{E}_A$ (A) was 0.032606, and the bootstrap $S\hat{E}_B$ (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 %. They are all small relative errors.

In comparison of datasets, it seems that the larger the dataset, the more accurate the bootstrap method. For small datasets, the statistics of interest employed in operational ROC analysis, such as TAR, EER, detection cost function, etc., can lose statistical meaning anyway, because of the small numbers of genuine scores and impostor scores. Under such circumstances, the metric AURC can be used and its estimated $S\hat{E}$ can be computed analytically.

For large datasets, from the operational perspective, the metrics, such as TAR, EER, detection cost function, etc., must be employed. And as pointed out in Section 1, it is hard to calculate uncertainties of such statistics of interest without using the nonparametric two-sample bootstrap method. Therefore, the validation of such an approach on large datasets provides a foundation for computing uncertainties in operational ROC analysis.

References

1. J.C. Wu, A.F. Martin, R.N. Kacker, C.R. Hagwood, Significance test in operational ROC analysis, in Biometric Technology for Human Identification VII, Proceedings of SPIE Vol. 7667, 76670I (2010).
2. J.C. Wu, Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap, NISTIR 7449, National Institute of Standards and Technology, September, 2007.
3. J.C. Wu, Operational measures and accuracies of ROC Curve on large fingerprint data Sets, NISTIR 7495, National Institute of Standards and Technology, May, 2008.
4. J.C. Wu, A.F. Martin, C.S. Greenberg, R.N. Kacker, Measurement uncertainties in speaker recognition evaluation, NISTIR 7722, National Institute of Standards and Technology, September, 2010.
5. J.C. Wu, C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (9) (2007) 2574-2584.
6. J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.
7. J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology, 148 (1983) 839-843.
8. B. Efron, Bootstrap methods: Another look at the Jackknife. Ann. Statistics, 7:1-26, 1979.
9. B. Efron, Better bootstrap confidence intervals, J. Amer. Statist. Assoc. 82 (397) (1987) 171-185.
10. B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
11. J.C. Wu, M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, in Biometric Technology for Human Identification IV, Proceedings of SPIE Vol. 6539, 65390N (2007).
12. R.Y. Liu, K. Singh, Moving blocks jackknife and bootstrap capture weak dependence. Exploring the limits of bootstrap, ed. by LePage and Billard. John Wiley, New York, 1992.
13. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, Guide to Biometrics, Springer, New York, 2003 pp. 269-292.

14. J.C. Wu, A.F. Martin, R.N. Kacker, Further studies of bootstrap variability for ROC analysis on large datasets, NISTIR 7730, National Institute of Standards and Technology, October, 2010.
15. J.C. Wu, C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, in Biometric Technology for Human Identification III, Proceedings of SPIE Vol. 6202, 620207 (2006).
16. B.L. van der Waerden, Mathematical Statistics, Springer, Berlin, 1969 p. 274 and pp. 333–335.
17. R.J. Hyndman, Y. Fan, Sample quantiles in statistical packages, American Statistician 50 (1996) 361-365.