NISTIR 7449

Studies of Operational Measurement of ROC Curve on Large Fingerprint Data Sets Using Two-Sample Bootstrap

Jin Chu Wu

NIST
National Institute of Standards and Technology

NISTIR 7449

Studies of Operational Measurement of ROC Curve on Large Fingerprint Data Sets Using Two-Sample Bootstrap

Jin Chu Wu

National Institute of Standards and Technology
Gaithersburg, MD 20899

September 2007

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
William Jeffrey, Director

Studies of Operational Measurement of ROC Curve on Large Fingerprint Data Sets Using Two-Sample Bootstrap

Jin Chu Wu[*]

Image Group, Information Access Division, Information Technology Laboratory

National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

From the operational perspective, on large fingerprint data sets, a receiver operating characteristic (ROC) curve is usually measured by the true accept rate (TAR) of the genuine scores given a specified false accept rate (FAR) of the impostor scores. The ties of genuine and/or impostor scores at a threshold can often occur on large fingerprint data sets, and how to determine the TAR at an operational FAR is provided. The accuracy of the measurement of TAR at a specified FAR for an ROC curve is explored using the nonparametric two-sample bootstrap. The variability of the estimates of standard error and lower bound and upper bound of 95% confidence interval of two-sample bootstrap distribution of the statistic TARs on large fingerprint data sets is extensively studied empirically. Thereafter, the number of two-sample bootstrap replications is determined. Both high-accuracy and low-accuracy fingerprint-image matching algorithms are taken as examples.

Keywords: Receiver operating characteristic (ROC) curve; Fingerprint matching; Nonparametric two-sample bootstrap; Variability; Standard errors; Confidence interval

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

1. Introduction

In many biometric applications, the performances of, for example, algorithms, etc. are evaluated by a receiver operating characteristic (ROC) curve. For instance, concerning the analysis of fingerprint data on large data sets, comparing two different fingerprint images of the same subject generates genuine score, and matching two fingerprint images of two different subjects creates impostor score. Sometimes, both of them are referred to as similarity scores in this article. The cumulative probabilities of genuine and impostor scores from the highest similarity score to a specified similarity score (i.e., threshold) are defined as the true accept rate (TAR) and the false accept rate (FAR), respectively. For continuous similarity scores, all these elements are schematically depicted in Figure 1 (A). Inside the similarity score range, among three variables TAR, FAR and threshold, one determines the other two. An ROC curve is constructed by moving the threshold from the highest similarity score down to the lowest similarity score [1]. An ROC curve in the FAR-and-TAR coordinate system is schematically drawn in Figure 1 (B).

Different fingerprint-image matching algorithms employ different scoring systems. In reality, nonetheless, different scoring systems can be converted to integral scores, if they are not [1]. Therefore, all similarity scores dealt with in this article are integral scores, and thus the distribution functions explored in this article are all discrete probability distribution functions and an ROC curve is no longer a smooth curve. The empirical distribution is assumed for each of the observed similarity scores. An ROC curve can be measured either by invoking the area under ROC curve [1, and references therein], or by using the TAR (or 1 – TAR) at a specified FAR from the operational perspective [2].

The area under an ROC curve, first, is equal to the probability of correctly identifying which is more likely than the other in the two stimuli under investigation, and it measures the overall ROC curve. Second, this area, if it is computed using the trapezoidal rule, is equivalent to the Mann-Whitney statistic that is formed, in the case of fingerprint data, by genuine and impostor scores. Hence, the variance of the Mann-Whitney statistic can be utilized as the variance of the area. That is, the measure of the area under an ROC curve can always be accompanied with a standard error. Third, because the Mann-Whitney statistic is asymptotically normally distributed regardless of the distributions of genuine and impostor scores thanks to the Central Limit Theorem, the Z statistic formulated in terms of areas under two ROC curves along with their variances and the correlation coefficient is subject to the standard normal distribution and can be used to test the significance of the difference of these two ROC curves.

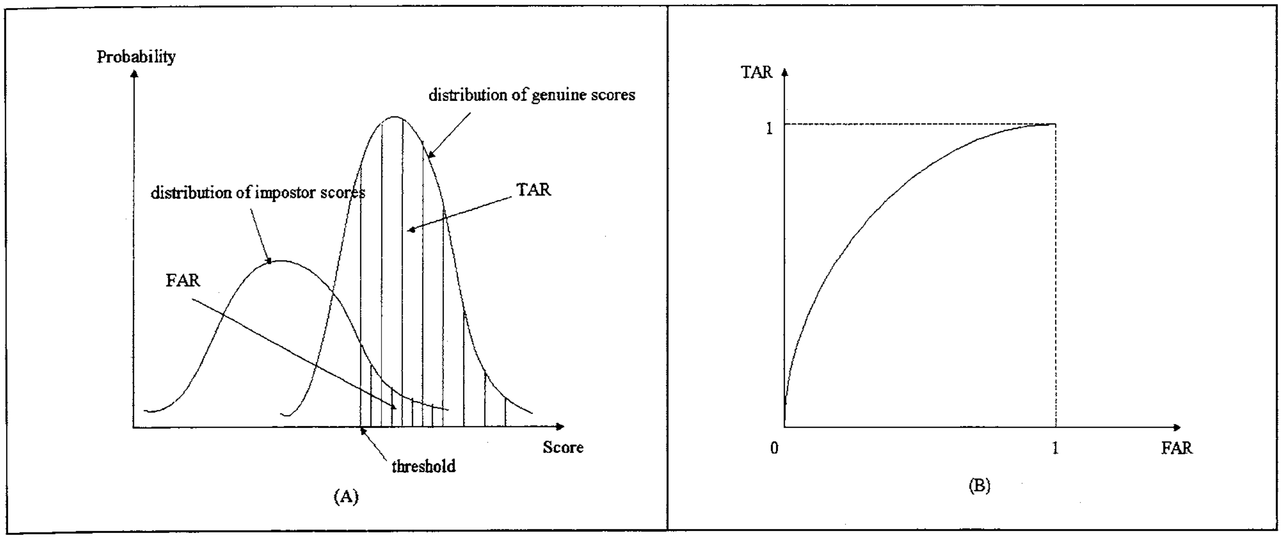
Probability
distribution of genuine scores
distribution of impostor scores
TAR
FAR
Score
threshold
(A)
TAR
1
0
1
FAR
(B)

Figure 1 (A): A schematic diagram of distributions of continuous genuine scores and impostor scores, showing TAR, FAR, and threshold. **(B):** A schematic drawing of an ROC curve constructed by moving the threshold from the highest similarity score down to the lowest similarity score.

Regarding the TAR at a specified FAR, which is used as a metric to evaluate an ROC curve from the operational perspective, the ties of genuine and/or impostor scores at a threshold can often occur on large fingerprint data sets. Under such circumstances, how to compute the TAR at a specified FAR will be presented in this article. The TAR at a specified FAR for an ROC curve is always an approximate measurement. Then, a question arises: How accurate is such an operational measurement? In the medical applications, many studies were done [3-7]. But, their data sets are usually far smaller than those in the fingerprint applications. On large fingerprint data sets, so far, there has been no study about measuring the accuracy associated with this metric. In the literature [8], authors dealt with the situation, in which the threshold was specified and thus 1 - TAR and FAR varied.

As having extensively investigated [1], after executing different fingerprint-image matching algorithms on large data sets, it was revealed that there is usually no underlying parametric distribution function for genuine and impostor scores, the distribution of genuine scores is considerably different from the distribution of impostor scores in general, and the distributions vary substantially from algorithm to algorithm in such a way that makes algorithms so different in terms of qualities. This suggests that the nonparametric analysis is pertinent to evaluating fingerprint-image matching algorithms on large-scale data sets.

An ROC curve is determined by the distribution function of genuine scores as well as the distribution function of impostor scores. More explicitly, an ROC curve is characterized by the relative relationship between these two distributions [1,9]. Further, the distribution of genuine scores and the distribution of impostor scores are interrelated by the algorithm that generates these two distributions. In other words, the performance of a fingerprint-image matching algorithm is determined not only by its ability of executing the genuine matching but also by its ability of implementing the impostor matching. In the medical applications, two possible corrections to the variance that was estimated under the assumption of binomial distribution were proposed [3,5].

As a result, the issue of computing the TAR at a specified FAR and measuring its accuracy for an ROC curve is a two-distribution issue other than one-distribution issue. Thus, in this article, the nonparametric two-sample bootstrap rather than the one-sample bootstrap is employed [10-13]. Here, the two samples are a set of genuine scores and a set of impostor scores, respectively. The statistic of interest is the TAR at an operational FAR under the combined impact of these two samples. The FAR is set to be 0.001 in this article [9,14]. The total number of genuine scores is a little over 60 000 and the total number of impostor scores can reach as high as about 120 000 [14].

For bootstrap methods, one of very important parameters is the number of bootstrap replications. It is intrinsically related to the variability of some feature, such as standard error, lower bound and upper bound of confidence interval, etc., of two-sample bootstrap distribution of the statistic of interest. And it also depends on the underlying distributions of, for instance, similarity scores in fingerprint application and what the statistic of interest is [11-13]. As studied before [1], it is absolutely inappropriate to assume the normality for distributions of similarity scores generated by fingerprint-image matching algorithms. The statistic of interest in our case is the TAR at a specified FAR rather than a simple sample mean. Moreover, the sizes of fingerprint similarity scores are much greater than those that have been encountered in other applications, such as medical application, etc.. Therefore, the issue of how many two-sample bootstrap replications are needed to meet accuracy requirement in the fingerprint application needs to be investigated.

Not only the standard error but also the 95% confidence interval of the TAR at a specified FAR can be computed using the nonparametric two-sample bootstrap. It can have many applications in the biometrics. For instance, while evaluating fingerprint-image matching algorithms, once a criterion of the TAR at a specified FAR is set, those algorithms whose 95% confidence intervals of the TARs at a specified FAR are higher than the criterion would be accepted at least at 95% confidence level.

The discrete distribution functions of genuine and impostor scores along with ROC curve are explored in Section 2. The method of computing TAR at a specified FAR for an ROC curve, including that the ties of similarity scores occur at the threshold, is presented in Section 3. An algorithm of calculating standard error and confidence interval of the TAR at a specified FAR using the nonparametric two-sample bootstrap is provided in Section 4. The variability of two-sample bootstrap estimates on large fingerprint data sets and thus the number of bootstrap replications are investigated in Section 5. The results of four fingerprint-image matching algorithms[1] are shown in Section 6, among which two are of high accuracy and two are of low accuracy. Finally, conclusion and discussion can be found in Section 7.

2. The discrete distribution functions of genuine and impostor scores and ROC curve

It is supposed that all similarity scores are represented in integers [1]. Without loss of generality, for any matching algorithm, the scoring system can be expressed inclusively using the integral

[1] These tests were performed for the Department of Homeland Security in accordance with section 303 of the Border Security Act, codified at 8 U.S.C. 1732. 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.

score set $\{s\} = \{s_{\min}, s_{\min}+1, \ldots, s_{\max}\}$, running consecutively from the minimum score $s_{\min}$ to the maximum score $s_{\max}$. To make the presentation clear, in the following text, the symbol "$\forall\ s \in \{s\}$" indicates that s takes all integral scores from $s_{\min}$ up to $s_{\max}$ in the ascending order, and the symbol "$\forall\ s \in \{\bar{s}\}$" means that s takes all integral scores from $s_{\max}$ down to $s_{\min}$ in the descending order.

While executing a matching algorithm over a fingerprint-image data set, the genuine score set, generated by comparing two different fingerprint images of the same subject, is denoted as

$$\mathbf{G} = \{\, m_i \mid m_i \in \{s\} \text{ and } \forall\ i \in \{1, \ldots, N_G\}\} \,, \tag{1}$$

where N_G is the total number of genuine scores. Note that genuine score m_i may not exhaust all members in the integral score set {s}. In addition, some of the comparisons may very well share the same integral value. Therefore, the genuine score set **G** can be partitioned into pairwise-disjoint subsets $\{\mathbf{G}_m\}$, in each of which members possess the same integral score $m \in \{s\}$. And the genuine score set **G** is the union of all these subsets $\{\mathbf{G}_m\}$.

Let P_G (m) denote the empirical probability at a genuine score m corresponding to the subset $\mathbf{G}_m$. To deal with the whole spectrum of the scores by including zero frequencies, the discrete probability distribution function of the genuine scores can be expressed as

$$\boldsymbol{P}_{\mathbf{G}} = \{\, P_G\,(s) \mid \forall\ s \in \{s\} \text{ and } \sum_{\tau = s\min}^{s\max} P_G\,(\tau) = 1 \,\} \,. \tag{2}$$

And the cumulative discrete probability distribution function of the genuine scores can be computed by moving the threshold one integral score at a time from the highest score $s_{\max}$ down to the lowest score $s_{\min}$. Thus, it can be expressed as

$$\boldsymbol{C}_{\mathbf{G}} = \{\, C_G\,(s) = \sum_{\tau = s}^{s\max} P_G\,(\tau) \mid \forall\ s \in \{\bar{s}\}\,\} \,, \tag{3}$$

where C_G (s) is the cumulative probability of the genuine scores, i.e., the TAR, at the integral score s from the highest score $s_{\max}$.

By analogy with the genuine scores, the impostor score set, created by matching two fingerprint images of two different subjects, is expressed as

$$\mathbf{I} = \{\, n_i \mid n_i \in \{s\} \text{ and } \forall\, i \in \{1, \ldots, N_I\}\} \,, \tag{4}$$

where N_I is the total number of impostor scores. The discrete probability distribution function of the impostor scores can be formulated in terms of the empirical probability P_I (s) as

$$\boldsymbol{P}_{\mathbf{I}} = \{\, P_I\,(s) \mid \forall\, s \in \{s\} \text{ and } \sum_{\tau=s\min}^{s\max} P_I\,(\tau) = 1 \,\} \,. \tag{5}$$

And the cumulative discrete probability distribution function of the impostor scores can be expressed as

$$\boldsymbol{C}_{\mathbf{I}} = \{\, C_I\,(s) = \sum_{\tau=s}^{s\max} P_I\,(\tau) \mid \forall\, s \in \{\bar{s}\,\} \,\} \,, \tag{6}$$

where C_I (s) is the cumulative probability of the impostor scores, i.e., the FAR, at the integral score s from the highest score s_{max}.

An ROC curve, constructed based on the cumulative discrete probability distribution functions of the genuine and impostor scores, is defined in this article as a curve connecting $s_{max} - s_{min} + 1$ points, $\{\, (C_I\,(s), C_G\,(s)) \mid \forall\, s \in \{\bar{s}\,\} \,\}$, in the FAR-and-TAR coordinate system, and extending to the origin of the coordinate system. The fingerprint-image matching algorithm is designed in such a way that an ROC curve always starts from the origin of the FAR-and-TAR coordinate system, ends at the point (1, 1), and is above the straight line from the origin to (1, 1). Overlap of points (C_I (s), C_G (s)) can occur, while both P_I (s) and P_G (s) are zero. An ROC curve goes horizontally, vertically, or inclined upper-rightwards at the score s, depending on whether only P_I (s) is nonzero, or only P_G (s) is nonzero, or both of them are nonzero, respectively.

Except at scores where both P_I (s) and P_G (s) are zero, such a precise ROC curve provides the same information as that nonzero P_I (s) and nonzero P_G (s) provide. The precise ROC curve uniquely and accurately represents the cumulative discrete probability distribution functions of the genuine and impostor scores. Moreover, such an ROC curve is constructed directly from the original data, after converting to integral scores if necessary, without any assumption regarding their distributions. Investigating ROC curve of genuine and impostor scores is a way to discover how the discrete probability distribution functions of the genuine and impostor scores are related to each other, and thus how well/bad the fingerprint-image matching algorithm works.

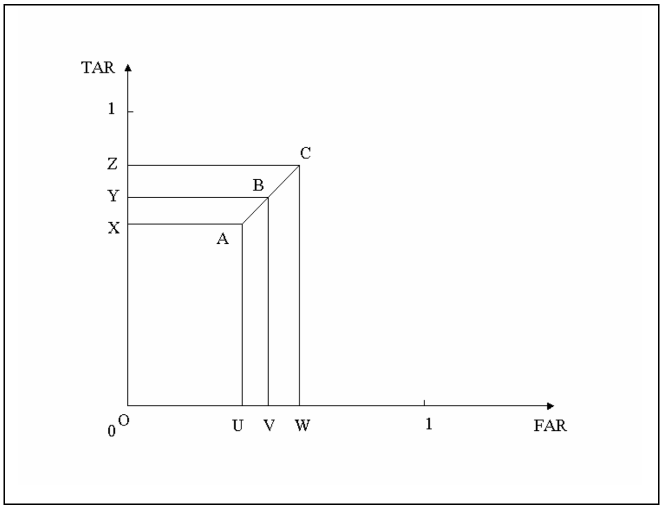

Figure 2 A schematic diagram of a line segment AC from similarity score t + 1 to the threshold t on an ROC curve in the FAR-and-TAR coordinate system. $\overline{OV}$ is equal to the specified FAR = f.

3. Determine the TAR Value at a Specified FAR for an ROC Curve

Assume that all similarity scores are converted to integral scores. Given a FAR = f where $0 < f < 1$, without loss of generality, the threshold t is defined to satisfy

$$C_I(t + 1) < f \text{ and } C_I(t) \geq f , \qquad (7)$$

where both t and (t + 1) $\in$ {s}. As a consequence, the probability of impostor scores at the threshold t, $P_I(t) = C_I(t) – C_I(t + 1)$, is always positive. If the probability of genuine scores at the threshold t, $P_G(t)$, is not equal to zero, it is schematically depicted in Figure 2 that AC represents a line segment from similarity score t + 1 to the threshold t on an ROC curve in the FAR-and-TAR coordinate system. Therefore, Point A is at ($C_I(t + 1)$, $C_G(t + 1)$), and Point C is at ($C_I(t)$, $C_G(t)$). In Figure 2, $\overline{OV}$ is set to be the specified FAR = f.

Theorem The estimated TAR at a specified FAR = f is

$$T\hat{A}R(f) = C_G(t + 1) + P_G(t) * \frac{f - C_I(t+1)}{P_I(t)} . \qquad (8)$$

Proof If $P_G(t)$ is not equal to zero, as shown in Figure 2, $\overline{OU} = C_I(t + 1)$, $\overline{OW} = C_I(t)$, $\overline{OX} = C_G(t + 1)$, $\overline{OZ} = C_G(t)$, and $\overline{OV} = f$. And $P_i(t) = C_i(t) – C_i(t + 1)$, where $i \in \{I, G\}$, due to

Eqs. (3) and (6). Therefore, $\hat{TAR}(f) = \overline{OY}$ is the one as shown in Eq. (8). If P_G (t) is equal to zero, the validity of Eq. (8) is obvious.

Eq. (8) indicates that if P_G (t) is not equal to zero, the ratio of ($\hat{TAR}(f) - C_G$ (t + 1)) to P_G (t) must be equal to the ratio of (f – C_I (t + 1)) to P_I (t); otherwise, the TAR is the same as C_G (t + 1). And the ratio is always in (0, 1] because of Eq. (7). In the practice of testing and evaluating different fingerprint-image matching algorithms executed on different qualities of data sets, which generates large sizes of genuine scores and impostor scores, it is found that ties of similarity scores at the threshold can often occur. That is, P_G (t) and/or P_I (t) can be relatively large. In some cases, the impact of the second term in Eq. (8) is not negligible. Hence, simply choosing C_G (t + 1) (i.e., neglecting this term) or C_G (t) = C_G (t + 1) + P_G (t) to be $\hat{TAR}(f)$ is inappropriate concerning the accuracy of the evaluation.

4. Compute Standard Error and Confidence Interval Using the Nonparametric Two-Sample Bootstrap [10-13]

As pointed out in Section 1, an ROC curve is determined by the relative relationship between the distribution function of genuine scores and the distribution function of impostor scores. Moreover, these two distributions are interrelated by the fingerprint-image matching algorithm that generates them. The statistic of interest in our practice from the operational perspective is the TAR at a specified FAR under the combined impact of these two distributions. Hence, the estimates of standard error and 95% confidence interval of the TAR at a specified FAR are computed using the nonparametric two-sample bootstrap rather than one-sample bootstrap.

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. If some kinds of dependence occur among individuals in the sample, for instance, time dependence, origination dependence, etc., the sample may need to be regrouped into subsets according to the dependency while resampling in the bootstrap [13]. Certainly, how to regroup the sample into subsets will have impact on the bootstrap results. In this article, the random sample is assumed to be i.i.d.. If this assumption is violated, then the bootstrap objects are the subsets of the sample rather than the individuals in the sample in order to preserve the dependency; however everything else in the bootstrap method remains intact.

The statistic of interest $T\hat{A}R(f)$ is estimated using Eq. (8). The algorithm of computing the accuracy of the estimator $T\hat{A}R(f)$ is as follows.

Algorithm I

```
for $i \leftarrow 1$ to B
        select $N_G$ scores with replacement from $N_G$ genuine scores $\rightarrow$ {new $N_G$ genuine scores}$_i$
        select $N_I$ scores with replacement from $N_I$ impostor scores $\rightarrow$ {new $N_I$ impostor scores}$_i$
        {new $N_G$ genuine scores}$_i$ & {new $N_I$ impostor scores}$_i$ $\rightarrow$ statistic $T\hat{A}R_i(f)$, as FAR = f

$\{ T\hat{A}R_i(f) \mid i = 1, ..., B \} \rightarrow S\hat{E}_B(f)$ and/or $(\hat{Q}_B(\alpha/2, f), \hat{Q}_B(1 - \alpha/2, f))$
```

where B is the number of two-sample bootstrap replications, N_G and N_I are the numbers of genuine and impostor scores, respectively, and $T\hat{A}R_i(f)$ represents the ith bootstrap replication at a specified FAR = f derived using Eq. (8). $S\hat{E}_B(f)$ denotes the estimator of the unbiased standard deviation. At the significance level α, $\hat{Q}_B(\alpha/2, f)$ and $\hat{Q}_B(1 - \alpha/2, f)$ are the $\alpha/2$ 100% and $(1 - \alpha/2)$ 100% quantiles of the distribution formed by all members in the set $\{T\hat{A}R_i(f) \mid i = 1,\ldots,B\}$, respectively. The Definition 2 in Ref. [15] is adopted in this article. That is, the sample quantile is obtained by inversing the empirical distribution function of sample with averaging at discontinuities. Thus, $(\hat{Q}_B(\alpha/2, f), \hat{Q}_B(1 - \alpha/2, f))$ stand for the approximate bootstrap $(1 - \alpha)$ 100% confidence interval. If 95% confidence interval is of interest, then α is set to be 0.05.

5. Variability of Two-Sample Bootstrap Estimates on Large Fingerprint Data Sets and the Number of Bootstrap Replications

1) Variability of Two-Sample Bootstrap Estimates

As pointed out in the literature [11-13], the substantial bootstrap variance is caused by the sampling variability as well as the bootstrap resampling variability. The former is because the sample size is less than the population size, and the latter is because the number of bootstrap replications is not infinite. In the meantime, the bootstrap variance results in the variances of, for example, standard error and confidence interval (i.e., its lower bound and upper bound) of the distribution formed by bootstrap replications of the statistic of interest. As a consequence, these variances can be functions of the sample size as well as the number of bootstrap replications. On the other hand, the sample size and the number of bootstrap replications can be determined by studying the variances of standard error and confidence interval of the bootstrap-replication distribution.

In our case, the bootstrap is a two-sample bootstrap. The sample sizes include both the total number of genuine scores N_G and the total number of impostor scores N_I. The impact of sample sizes N_G and N_I on the accuracy of measurement of ROC curves in the analysis of fingerprint data on large data sets was investigated in Ref. [14]. The studies were carried out in terms of both the area under an ROC curve and the TAR at an operational FAR using Chebyshev's inequality in combination with simple random sampling. Thus, it is assumed in this article that these two sample sizes are fixed as stated in Section 1.

As pointed out in Section 1, there is usually no underlying parametric distribution function for genuine and impostor scores, the distributions of genuine scores and impostor scores are considerably different in general, and the distribution functions are substantially different from algorithm to algorithm [1]. As a result, it is absolutely inappropriate to assume normal distribution for genuine scores and impostor scores generated by fingerprint-image matching algorithms. In addition, the statistic of interest in our fingerprint applications is the TAR at a specified FAR rather than a simple sample mean. Moreover, the sizes of fingerprint similarity scores are much greater than those that have been encountered in other applications, such as medical application, etc.. Therefore, the variances of standard error and confidence interval of the bootstrap-replication distribution on large fingerprint data sets will be investigated empirically by executing the two-sample bootstrap iteratively with respect to a fixed number of bootstrap replications. Thereafter, the number of bootstrap replications can be determined.

2) Compute Coefficients of Variation

To take into account the impact of the mean value while dealing with the variability of any estimate, the coefficient of variation (CV) is invoked. The algorithm of empirically computing CVs for standard error, lower bound and upper bound of confidence interval is as follows.

Algorithm II

```
for i ← 1 to L
    for j ← 1 to B
        Algorithm I (two-sample bootstrap)
    $\{ \hat{TAR}_j(f)_i \mid j = 1, ..., B \} \rightarrow \hat{SE}_B(f)_i, \hat{Q}_B(\alpha/2, f)_i, \hat{Q}_B(1-\alpha/2, f)_i$
$\{\hat{SE}_B(f)_i, \hat{Q}_B(\alpha/2, f)_i, \hat{Q}_B(1-\alpha/2, f)_i \mid i = 1, ..., L\} \rightarrow \hat{CV}_{B,L}(\kappa), \kappa = \mathbf{SE}_{B,L}(\mathbf{f}), \mathbf{Q}_{B,L}(\boldsymbol{\alpha}/2, \mathbf{f}), \mathbf{Q}_{B,L}(1-\boldsymbol{\alpha}/2, \mathbf{f})$
```

where L is the number of iterations and B is the number of bootstrap replications.

For a fixed number of bootstrap replications B, after L iterations of executing two-sample bootstrap, the following three sets are generated,

$$\begin{aligned} \mathbf{SE}_{B,L}(\mathbf{f}) &= \{ \hat{SE}_B(f)_i \mid \forall\, i \in \{1, \cdots, L\} \}, \\ \mathbf{Q}_{B,L}(\boldsymbol{\alpha}/2, \mathbf{f}) &= \{ \hat{Q}_B(\alpha/2, f)_i \mid \forall\, i \in \{1, \cdots, L\} \}, \\ \mathbf{Q}_{B,L}(1-\boldsymbol{\alpha}/2, \mathbf{f}) &= \{ \hat{Q}_B(1-\alpha/2, f)_i \mid \forall\, i \in \{1, \cdots, L\} \}. \end{aligned} \tag{9}$$

Hence, three CVs of standard error, lower-bound and upper-bound of confidence interval, respectively, are,

$$\hat{CV}_{B,L}(\kappa) = \frac{\sqrt{\hat{VAR}_{B,L}(\kappa)}}{\hat{E}_{B,L}(\kappa)}, \text{ where } \kappa = \mathbf{SE}_{B,L}(\mathbf{f}), \mathbf{Q}_{B,L}(\boldsymbol{\alpha}/2, \mathbf{f}), \mathbf{Q}_{B,L}(1-\boldsymbol{\alpha}/2, \mathbf{f}). \tag{10}$$

Here, the three CVs are functions of the number of bootstrap replications B and the number of iterations L, besides the significance level α and the FAR f. Therefore, the number of bootstrap replications B can be determined by the tolerable CV. Then, the question is: How many iterations L are required for a fixed number of bootstrap replications B? This issue will also be dealt with empirically.

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.047524	0.034664	0.027754	0.023912	0.021570
	Max.	0.054346	0.039866	0.031685	0.026866	0.023686
	Range	0.006822	0.005202	0.003931	0.002954	0.002116
CVLB	Min.	0.000062	0.000044	0.000036	0.000030	0.000026
	Max.	0.000067	0.000047	0.000041	0.000037	0.000031
	Range	0.000005	0.000003	0.000005	0.000007	0.000005
CVUB	Min.	0.000054	0.000041	0.000032	0.000030	0.000026
	Max.	0.000062	0.000044	0.000036	0.000032	0.000030
	Range	0.000008	0.000003	0.000004	0.000002	0.000004

Table 1 High-accuracy Algorithm 1's minimum, maximum, and range of CVSEs, CVLBs, and CVUBs, while the number of replications B was from 200 up to 1000 for every 200. For each fixed number of replications B, the minimum, maximum, and range of CVs were generated from 10 estimates of CVs as the number of iterations L ran from 100 up to 1000 for every 100.

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.021218	0.018613	0.017951	0.016331	0.016040
CVLB	0.000027	0.000024	0.000023	0.000023	0.000020
CVUB	0.000024	0.000023	0.000022	0.000020	0.000019

Table 2 High-accuracy Algorithm 1's CVSEs, CVLBs, and CVUBs, while the number of replications B was from 1200 up to 2000 for every 200 and the number of iterations was fixed at 500.

3) Determine the Number of Iterations and Results of Three Coefficients of Variation

Two fingerprint-image matching algorithms are employed.[2] Among them, Algorithm 1 is of high accuracy, and Algorithm 2 is of low accuracy. The significance level was set to be 5% and the operational FAR = f was specified at 0.001. The estimates of three CVs for standard error, lower bound and upper bound of 95% confidence interval are denoted by CVSE, CVLB, and CVUB, respectively. In Table 1, it shows high-accuracy Algorithm 1's minimum, maximum, and range of CVSEs, CVLBs, and CVUBs, respectively, while the number of replications B was set to be from 200 up to 1000 for every 200. For each fixed number of replications B, the minimum, maximum, and range of CVs were generated from 10 estimates of CVs as the number of iterations L ran from 100 up to 1000 for every 100, respectively.

[2] The algorithms are proprietary. Hence, they cannot be disclosed.

Num. of replications B		200	400	600	800	1000
CVSE	Min.	0.056895	0.037193	0.031792	0.026763	0.024033
	Max.	0.062609	0.043167	0.034696	0.030500	0.026695
	Range	0.005714	0.005974	0.002904	0.003737	0.002662
CVLB	Min.	0.000941	0.000677	0.000519	0.000473	0.000442
	Max.	0.001052	0.000734	0.000627	0.000526	0.000478
	Range	0.000111	0.000057	0.000108	0.000053	0.000036
CVUB	Min.	0.001068	0.000685	0.000637	0.000532	0.000488
	Max.	0.001171	0.000838	0.000738	0.000611	0.000544
	Range	0.000103	0.000153	0.000101	0.000079	0.000056

Table 3 Low-accuracy Algorithm 2's minimum, maximum, and range of CVSEs, CVLBs, and CVUBs, while the number of replications B was from 200 up to 1000 for every 200. For each fixed number of replications B, the minimum, maximum, and range of CVs were generated from 10 estimates of CVs as the number of iterations L ran from 100 up to 1000 for every 100.

Num. of replications B	1200	1400	1600	1800	2000
CVSE	0.023673	0.022299	0.021272	0.018918	0.017705
CVLB	0.000457	0.000397	0.000354	0.000331	0.000318
CVUB	0.000445	0.000429	0.000420	0.000389	0.000389

Table 4 Low-accuracy Algorithm 2's CVSEs, CVLBs, and CVUBs, while the number of replications B was from 1200 up to 2000 for every 200 and the number of iterations was fixed at 500.

It is observed from Table 1 that the maximal CVs for the standard error are relatively not too small but are getting smaller as the number of replications B increases. The ranges of 10 estimates of CVs for the standard error change from about 0.007 down to 0.002. The maximal CVs for lower bound and upper bound of 95% confidence intervals are less than 0.00007, and the ranges are not greater than 0.000008. As a result, to obtain the estimate of CV at a fixed number of replications B which is higher than 1000, the number of iterations L does not need to run from 100 up to 1000 for every 100. Therefore, the number of iterations L was fixed at 500, while the number of replications B varied from 1200 up to 2000 for every 200. The corresponding estimates of CVs are shown in Table 2.

The CVs for low-accuracy Algorithm 2 are all greater than those for high-accuracy Algorithm 1, correspondingly. This is consistent with the phenomena observed in our previous studies

[1,9,14]. Hence, the tolerances for low-accuracy algorithms should be set larger than those for high-accuracy algorithms if necessary. Nonetheless, the tendency of changes of CVs with respect to the number of replications B as well as the number of iterations L for low-accuracy algorithm remains the same as the trend for high-accuracy algorithm. As shown in Table 3, which has the same structure as Table 1, for low-accuracy Algorithm 2, the ranges of 10 CVs for the standard error vary from about 0.006 down to 0.003. The maximal CVs for lower bound and upper bound of 95% confidence intervals are less than 0.0012, and the ranges are less than 0.0002. Thus, the number of iterations L could also be set at 500 while creating Table 4.

4) Further Investigation of Three Coefficients of Variation

Further investigation is taken over the cases generated by 500 iterations while the number of two-sample bootstrap replications B was set to be 2000 for high-accuracy Algorithm 1 and low-accuracy Algorithm 2. These two cases are shown in the last column of Table 2 and Table 4, respectively. The corresponding six histograms of standard error, lower bound and upper bound of 95% confidence interval for two algorithms are shown in Figure 3. The means, standard errors, CVs, and 95% confidence intervals of these six distributions are listed in Table 5.

With such numbers of iterations and replications, as shown in Table 5, for Algorithm 1, the mean, standard error, and thus the CV are 0.000331, 0.0000053, 0.016040 for the standard error; 0.992617, 0.0000198, 0.000020 for the lower bound; and 0.993913, 0.0000192, 0.000019 for the upper bound of 95% confidence interval, respectively. For Algorithm 2, the mean, standard error, and thus the CV are 0.003474, 0.0000615, 0.017705 for the standard error; 0.789746, 0.0002514, 0.000318 for the lower bound; and 0.804121, 0.0003124, 0.000389 for the upper bound of 95% confidence interval, respectively.

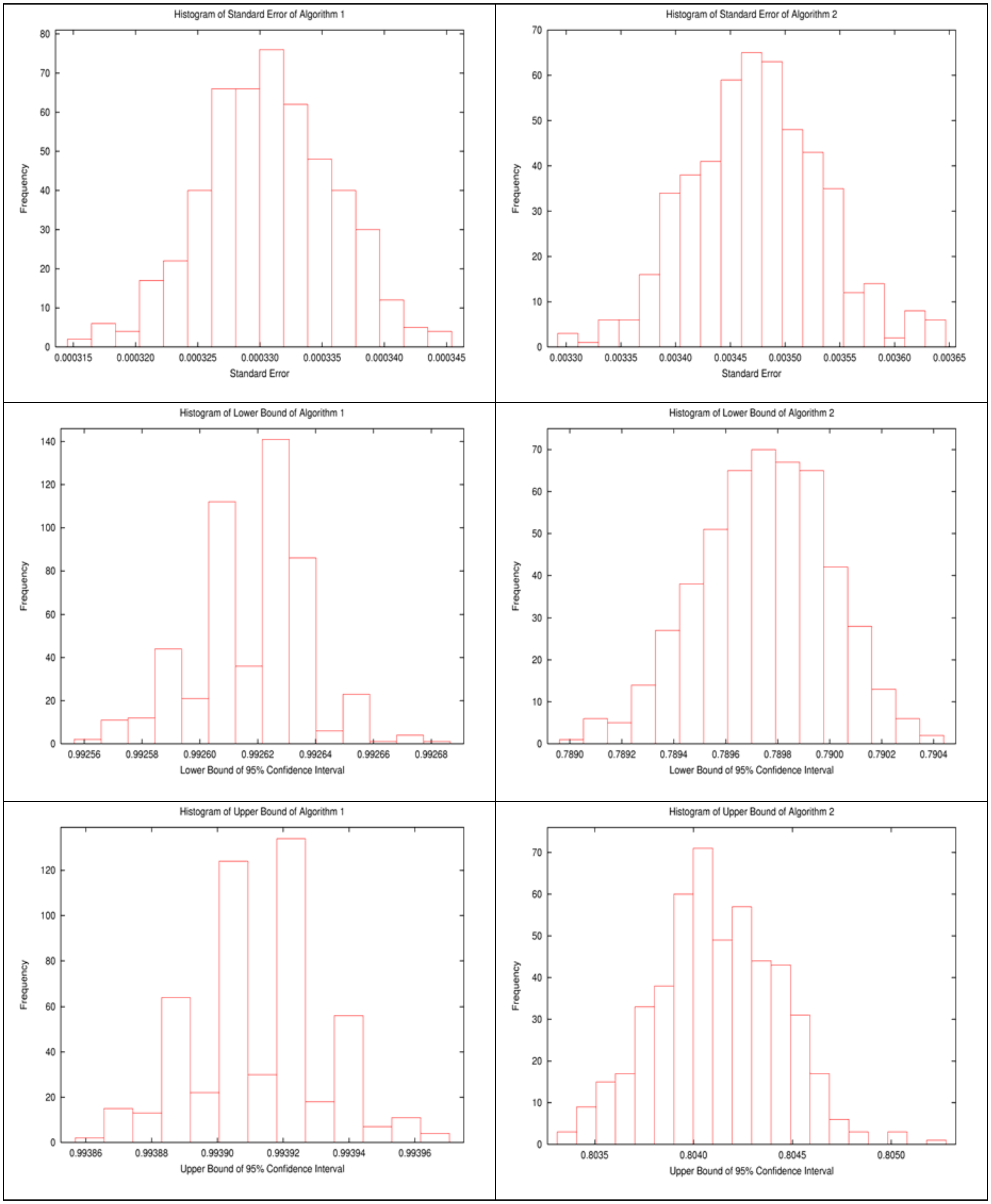

Figure 3 Histograms of standard errors, lower bounds and upper bounds of 95% confidence intervals, generated by 500 iterations, while the number of replications B was set to be 2000, for high-accuracy fingerprint-image matching Algorithm 1 (left column) and low-accuracy Algorithm 2 (right column).

Algorithm		Mean	SE	CV	95% Confidence interval
1	Standard error	0.000331	0.0000053	0.016040	(0.000320, 0.000341)
	Lower bound	0.992617	0.0000198	0.000020	(0.992575, 0.992654)
	Upper bound	0.993913	0.0000192	0.000019	(0.993873, 0.993954)
2	Standard error	0.003474	0.0000615	0.017705	(0.003362, 0.003618)
	Lower bound	0.789746	0.0002514	0.000318	(0.789244, 0.790220)
	Upper bound	0.804121	0.0003124	0.000389	(0.803522, 0.804700)

Table 5 Means, standard errors (SE), CVs, and 95% confidence intervals of distributions of standard errors, lower bounds and upper bounds of 95% confidence intervals for Algorithm 1 and 2, respectively, generated by 500 iterations, while the number of replications B was set to be 2000.

Thus, it is demonstrated that the distribution of standard errors is of less dispersion than the distributions of lower bounds and upper bounds of 95% confidence intervals, respectively, regardless of whether the accuracy of algorithm is high or low. However, the mean of standard errors is much less than 1, and on the contrary the means of lower bounds and upper bounds of 95% confidence intervals are very close to 1 for high-accuracy algorithms and quite close to 1 for low-accuracy algorithms. This causes that the CV for standard errors is much larger than the CVs for lower bounds and upper bounds of 95% confidence intervals for each algorithm. As a consequence, the tolerance for CV of standard errors may be set larger than those for CVs of two bounds of 95% confidence intervals.

The 95% confidence intervals shown in Table 5 were computed using the Definition 2 of quantile in Ref. [15]. They do match the 95% confidence intervals by at least five decimal places for high-accuracy Algorithm 1 and four decimal places for low-accuracy Algorithm 2, which are calculated if the distributions generated by 500 iterations while the number of bootstrap replications is fixed at 2000 are assumed to be normal.

5) Determine the Number of Bootstrap Replications

All CVs shown in Table 1 through Table 4 are depicted in Figure 4 through Figure 6. They are CVSEs, CVLBs, and CVUBs for Algorithms 1 and 2, respectively. As stated in Section 5.3, for each fixed number of bootstrap replications running from 200 to 1000 for every 200, all CVs were referred to the maximal CVs taken from 10 CVs that were generated while the number of iterations L was set to be from 100 up to 1000 for every 100. For higher numbers of bootstrap replications, all CVs were created while the number of iterations L was fixed at 500.

As illustrated in Figure 4, all CVs for standard errors of Algorithm 1 and 2, respectively, decrease as the number of replications B increases. If the tolerance is set to be at 0.02, 1400 two-sample bootstrap replications are sufficient for high-accuracy Algorithm 1, and 1800 replications are enough for low-accuracy Algorithm 2. This is consistent with what was learned before [1,9,14], that is, to achieve the same level of accuracy, high-accuracy fingerprint-image matching algorithms always require less execution than low-accuracy algorithms do.

The CVs for lower bound and upper bound of 95% confidence interval for Algorithm 1 are shown in Figure 5. As indicated in Section 5.4, the tolerances for CVs of lower bounds and upper bounds of 95% confidence intervals should be set smaller. Hence, if the tolerance is set to be at 0.000025, 1400 replications can meet the requirement. Those for Algorithm 2 are depicted in Figure 6. As pointed out in Section 5.3, the tolerance for low-accuracy algorithms could be set larger. Thus, if the tolerance is set to be at 0.000450, 1400 replications can satisfy the restriction.

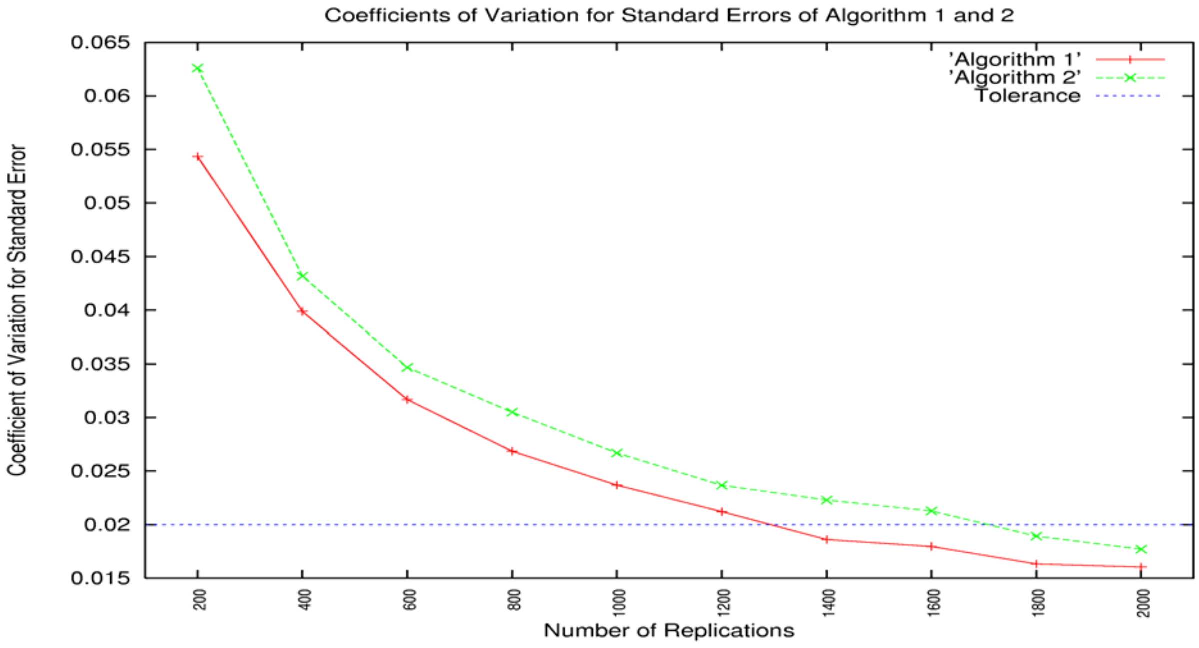

Figure 4 CVs of standard errors for high-accuracy Algorithm 1 and low-accuracy Algorithm 2, respectively, as a function of the number of replications, along with the tolerance line that is set to be at 0.02.

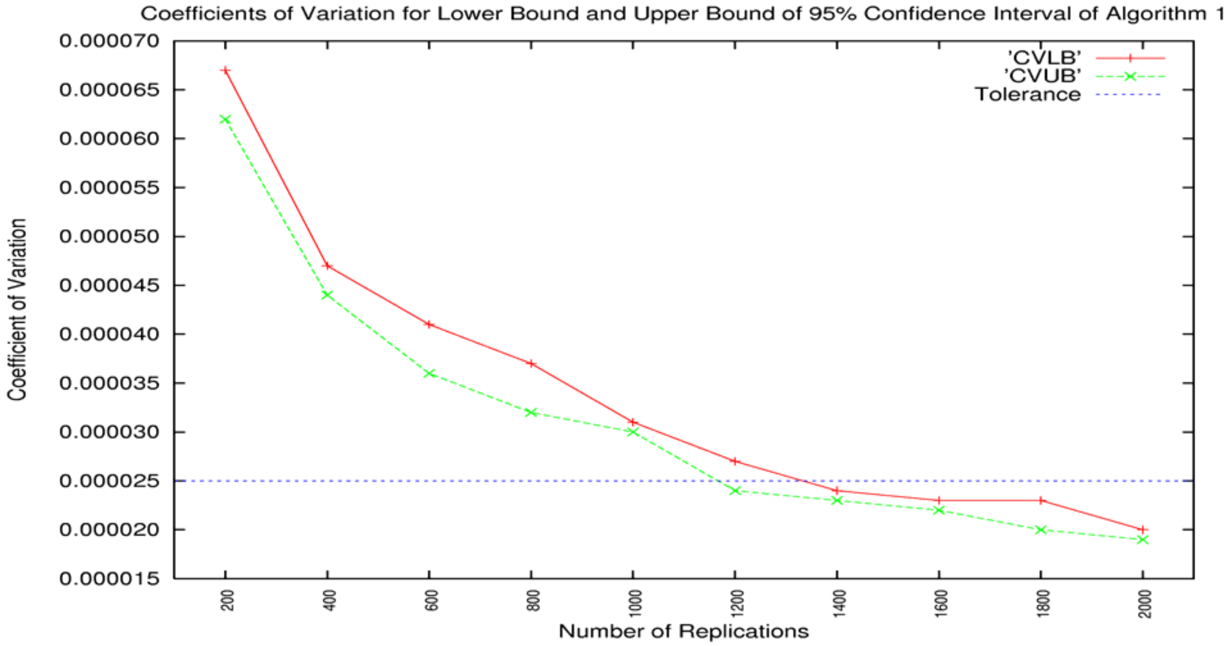

Figure 5 CVs of lower bound and upper bound of 95% confidence interval for high-accuracy Algorithm 1, respectively, as a function of the number of replications, along with the tolerance line that is set to be at 0.000025.

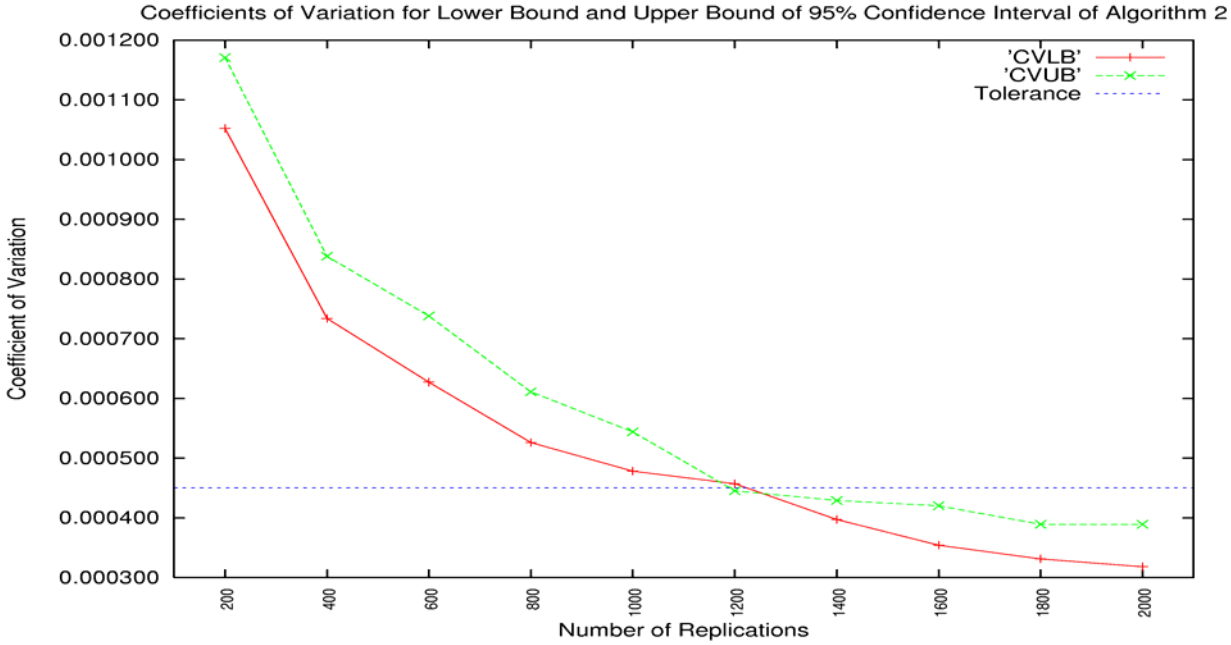

Figure 6 CVs of lower bound and upper bound of 95% confidence interval for low-accuracy Algorithm 2, respectively, as a function of the number of replications, along with the tolerance line that is set to be at 0.000450.

Although the tolerance set for the CVs for standard error is much larger than those for lower bound and upper bound of 95% confidence interval, the tolerance 0.02 for CVs is acceptable concerning our fingerprint application. Therefore, 1400 replications are sufficient for high-accuracy algorithm, and 1800 replications are enough for low-accuracy algorithm. To reconcile numbers of replications for different qualities of fingerprint-image matching algorithms, and further to be more conservative, it is suggested that 2000 two-sample bootstrap replications be required in order to achieve certain statistical accuracy.

Algorithm	$\hat{TAR}(f)$	$\hat{SE}$	95% Confidence interval
1	0.993255	0.000333	(0.992589, 0.993905)
3	0.994322	0.000337	(0.993679, 0.994962)
2	0.796753	0.003452	(0.789878, 0.803920)
4	0.871757	0.002265	(0.867283, 0.876270)

Table 6 The estimates of TARs, standard errors (SE), and 95% confidence intervals for high-accuracy Algorithms 1 and 3, and low-accuracy Algorithms 2 and 4, respectively, while the number of two-sample bootstrap replications was set to be 2000 and FAR was specified at 0.001.

6. Results

Two more fingerprint-image matching algorithms are taken as examples.[3] Among them, one is of high accuracy named as Algorithm 3, and the other is of low accuracy called as Algorithm 4. In Table 6 shown are the estimates of TARs, standard errors (SE), and 95% confidence intervals for high-accuracy Algorithms 1 and 3, and low-accuracy Algorithms 2 and 4, respectively, while the number of two-sample bootstrap replications is set to be 2000 and FAR is specified at 0.001. The 95% confidence intervals were calculated using the Definition 2 of quantile in Ref. [15]. They do match the 95% confidence intervals up to the fourth decimal place for high-accuracy Algorithms 1 and 3 and the third decimal place for low-accuracy Algorithms 2 and 4, if the distribution of 2000 bootstrap replications of the statistic $\hat{TAR}(f)$ for each algorithm is assumed to be normal. It is found that the higher the accuracy of the algorithm is, the smaller the standard error is, and

[3] The algorithms are proprietary. Hence, they cannot be disclosed.

thus the narrower the 95% confidence interval is. This is consistent with the observations in Ref. [1,9,14].

For Algorithms 1 and 2, while the number of bootstrap replications B is fixed at 2000, 500 iterations had been run and the resultant 95% confidence intervals of standard error, lower bound and upper bound of 95% confidence interval, respectively, are shown in Table 5. The results shown in Table 6 were generated by a random run (i.e., which is not one of the above 500 runs) of each algorithm while the number of bootstrap replications was fixed at 2000. It is worth pointing out that for Algorithms 1 and 2, the standard errors, lower bounds and upper bounds of 95% confidence intervals shown in Table 6 all fall in the corresponding 95% confidence intervals shown in Table 5. For instance, 0.000333, the estimate of the standard error of Algorithm 1 in Table 6, falls in the 95% confidence interval (0.000320, 0.000341) of the standard error for Algorithm 1 in Table 5. Noticeably, these 95% confidence intervals shown in Table 5 are very narrow with respect to different qualities of algorithms. This indicates that the computation in this article is quite self-consistent.

7. Conclusion and Discussion

On large fingerprint data sets, the ties of genuine and/or impostor scores at a threshold can often occur. The method of calculating the estimator $\hat{TAR}(f)$ at a specified FAR from operational perspective for an ROC curve generated by a fingerprint-image matching algorithm was provided. The two-sample bootstrap was applied to computing the accuracy of the measure $\hat{TAR}(f)$ at a fixed FAR in terms of standard error and confidence interval. The variability of two-sample bootstrap with respect to large fingerprint data sets was extensively studied empirically. It is suggested that 2000 two-sample bootstrap replications be sufficient to meet the acceptable tolerances that are set, respectively, for the CVs of standard error, lower bound and upper bound of 95% confidence interval, as well as for both high-accuracy and low-accuracy fingerprint-image matching algorithms. Finally, four algorithms, among which two were high-accuracy and two were low-accuracy, were taken as examples. The same approach can be applied to investigating the estimate of threshold and its accuracy.

In this article, the statistic of interest is TAR at a specified FAR. In some literature [2], to measure an ROC curve, the false non-match rate (FNMR), which is equal to 1 – TAR given the same FAR, was employed. It is trivial to show that under the same conditions (i.e., with respect to the same series of two bootstrap samples selected with replacement from genuine scores and impostor scores, respectively, as described in Algorithm I of Section 4) the standard error of FNMR is the same as that of TAR, and the lower bound and upper bound of 95% confidence interval for FNMR can be obtained by interchanging two bounds for TAR and subtracting them from 1, respectively.

In terms of FNMR, two bounds of 95% confidence intervals are very close to 0 for high-accuracy algorithms and quite close to 0 for low-accuracy algorithms. Such a behavior is different from that in terms of TAR, and this can have impact on CVs, as pointed out in Section 5.4. Under the same conditions of Table 5, generated by 500 iterations while the number of replications was set to be 2000, the CVs of lower bound and upper bound of 95% confidence interval were 0.003152 and 0.002687 for Algorithm 1, and 0.001595 and 0.001196 for Algorithm 2, respectively. Therefore, the assertion that the number of two-sample bootstrap replications is 2000 is still valid if FNMR is invoked. These numbers also indicate that the CVs for two bounds of 95% confidence intervals increase while using FNMR as opposed to TAR. To determine the uncertainties of the FNMR measure, everything else related to TAR presented in this article holds good for FNMR.

Different accuracies of fingerprint-image matching algorithms have different standard errors and thus different confidence intervals under the same circumstances. Generally speaking, the higher the accuracy of the algorithm is, the smaller the standard error is, and the narrower the confidence interval is. It is for sure that there is no universal standard error which can hold good for all algorithms. In addition, the higher-accuracy algorithms have less bootstrap variance as well.

As pointed out in Section 5.1, the variance of two-sample bootstrap is also caused by the sample size. In our fingerprint applications, the total number of genuine scores was a little over 60 000 and the total number of impostor scores was as high as about 120 000. What if the number of similarity scores changes? As demonstrated in our previous studies [14], if the sizes of similarity scores get larger than what are used here, the accuracy of the measure of statistic $\hat{TAR}(f)$ at a fixed FAR will not obtain substantial improvement. Indeed, after the sizes get greater than a certain level, there is little improvement in accuracy. If the sample sizes in other biometric applications are different from the ones used here as well as the normality cannot be assumed for the population from which the sample is selected, etc., the number of bootstrap replications may need to be reinvestigated. Nonetheless, the empirical methodology developed in this article should remain the same.

The alternative measure of statistical accuracy for the estimator $\hat{TAR}(f)$ besides standard error and confidence interval is using bias. A bootstrap estimated bias can be defined as the absolute difference between the average of bootstrap replications and the estimate of the statistic of interest [13]. If the bias is too large in comparison with the standard error, then $\hat{TAR}(f)$ might not be an appropriate estimator. For the examples as shown in Section 6, the means out of 2000 bootstrap replications for Algorithms 1, 3, 2, and 4 are 0.993264, 0.994324, 0.796746, and 0.871733, respectively. Thus, the corresponding bootstrap estimated biases are 0.0000081, 0.0000019, 0.0000063, and 0.0000237. And the ratios of the bootstrap estimated bias to the standard error for these four algorithms are 0.024, 0.006, 0.002, and 0.010, respectively. Indeed, these ratios are substantially small.

Using the area under an ROC curve as a metric to evaluate the performance of an algorithm and thus the variance of the Mann-Whitney statistic to measure the standard error of area is a deterministic process. However, using the two-sample bootstrap to compute the accuracy of the measure from the operational perspective is a stochastic process of a Monte Carlo simulation. Therefore, standard error, lower bound and upper bound of 95% confidence interval of the statistic of interest may fluctuate every time when they are calculated by a random run of two-sample bootstrap. Nonetheless, as has been studied in Sections 5.4 and 6, such standard error, lower bound and upper bound of 95% confidence interval may fall into the confidence intervals with 95% probability, which are generated by many iterations of executions of two-sample bootstrap. Moreover, these confidence intervals are so narrow from the practical point of view.

As pointed in Section 1, if the area under an ROC curve is used to measure the performance of fingerprint-image matching algorithms, the Z statistic can be used to test the significance of the difference between two ROC curves. In this article, from the operational perspective, the metric of evaluating matching algorithms is using the statistic $\hat{TAR}(f)$ at a specified FAR. So far, standard error and confidence interval have been studied. However, the associated significance test has not been investigated yet. In order to compare two fingerprint-image matching algorithms or compare an algorithm against a criterion, the 95% confidence interval can be invoked to some extent. The work of the significance test in this regard is underway.

References

1. J.C. Wu, C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (2007) 2574-2584.
2. R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, A.K. Jain, Performance evaluation of fingerprint verification systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (1) 2006 3-18.
3. K. Linnet, Comparison of quantitative diagnostic tests: type I error, power, and sample size, Statistics in Medicine 6 (1987) 147-158.
4. D. Mossman, Resampling techniques in the analysis of non-binormal ROC data, Medical Decision Making, 15 (4) (1995) 358-366.
5. R.W. Platt, J.A. Hanley, H. Yang, Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test, Statistics in Medicine, 19 (3) (2000) 313-322.
6. G. Campbell, General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, Statistics in Medicine, 13 (1994) 499-508.
7. K. Jensen, H.-H. Muller, H. Schafer, Regional confidence bands for ROC curves, Statistics in Medicine, 19 (4) (2000) 493-509.
8. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, Guide to Biometrics, Springer, New York, 2003 pp. 269-292.
9. J.C. Wu, M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, Proc. of SPIE Vol. 6539, Biometric Technology for Human Identification IV, Edited by S. Prabhakar, A.A. Ross, April 2007.

10. B. Efron, Bootstrap methods: Another look at the Jackknife. Ann. Statistics, 7:1-26, 1979.

11. P. Hall, On the number of bootstrap simulations required to construct a confidence interval, Ann. Statist. 14 (4) (1986) 1453-1462.

12. B. Efron, Better bootstrap confidence intervals, J. Amer. Statist. Assoc. 82 (397) (1987) 171-185.

13. B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993 pp. 271-282.

14. J.C. Wu, C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, Proc. of SPIE Vol. 6202, Biometric Technology for Human Identification III, Edited by P.J. Flynn, S. Pankanti, April 2006.

15. R.J. Hyndman, Y. Fan, Sample quantiles in statistical packages, American Statistician 50 (1996) 361-365.