

Evaluating Authentication Systems Using Bootstrap Confidence Intervals

Ruud M. Bolle, Nalini K. Ratha and Sharath Pankanti
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
{bolle, ratha, sharat}@us.ibm.com

Abstract

Reporting the matching performance of biometrics identification or authentication systems is a controversial issue and perhaps an issue that is not well understood. It seems apparent that False Reject Rates and False Accept Rates are the measurements to express system accuracy. The size of the database of biometrics signals which is used to obtain estimates of the error rates is another important parameter that should be reported along with the error estimates. This database size can be used to compute confidence intervals of the error rates. Further, some indication of the quality of the fingerprint impressions should also be given. This paper develops bootstrap confidence intervals for estimating the accuracy of the error estimates and bootstrap techniques for indicating the quality and diversity of the prints in the database. The size of the database is also related to the accuracy of the error estimates. Results of our experiments on a sample fingerprint database is reported.

1 Introduction

A biometrics signal is a pattern and authenticating a person with a fingerprint or other biometrics is in essence an issue of hypothesis testing. Let the stored fingerprint be presented by template P' and the acquired fingerprint by P . In terms of hypothesis testing, we have

$$\begin{aligned} H_0 : P = P', & \quad \text{the person is genuine} \\ H_1 : P \neq P', & \quad \text{the person is an impostor.} \end{aligned}$$

Often some similarity measure $s = Sim(P, P')$ is defined and H_0 is decided if $s \geq T$ and H_1 is decided if $s < T$. Then, deciding H_0 when H_1 is true gives a false accept and deciding H_1 when H_0 is true results in a false reject.

False Accept Rates (FAR) and False Reject Rates (FRR) are important intrinsic characteristics of a matcher. Given a matcher, in theory the FRR and FAR could be determined analytically. For instance,

in fingerprint authentication scenario, if all the sources of noise, such as sensor noise, feature noise and distortions between pairs of matching finger templates can be modeled then the error rates can be computed. Probably the most difficult issue here is the sources of noise introduced by imaging of the type of fingers found in the target population. Clearly, it is impossible to model all noise sources and one has to use statistical techniques to estimate the error rates. It is important to remember then that the reported error rates are estimates of the true error rates, and just that.

Statistical performance evaluation of biometrics authentication or identification systems in terms of FRR and FAR is a difficult issue, if addressed at all. All too often, error rates are not or poorly reported, or, worse, the systems are claimed to be 100% accurate. At least, the Equal Error Rate (where FRR = FAR) should be reported, but it is more desirable to report system accuracy with a Receiver Operating Curve (ROC) [1, 2]. This is a graph that expresses the relation between FRR and FAR when the matching threshold T is varied. However, the ROC in itself does not mean much. The database size of fingerprint impressions that is used to compute these statistics should be reported. Here, some indication of the quality of the impressions should be given, e.g., the conditions under which the prints are collected and a description of the subjects who are used for acquiring the database. Finally, it should be reported how accurate the estimates of the above statistics really are. Here, the accuracy of the estimates depends on the database size—the larger the size, the more accurate the estimate. (Other methods for evaluating authentication systems can be found in [4]).

All the above issues can be addressed by computing confidence intervals both on distributions and on distribution parameters. We use bootstrap techniques [3, 5] which lend themselves particularly well to computing confidence intervals to develop techniques for indicating the accuracy of biometrics systems.

In this paper, we describe a bootstrap-based technique for computing confidence intervals for system evaluation. Section 2 describes bootstrap technique and introduces the definitions used in this paper. Procedures for evaluating the validity and accuracy of the error estimates for biometrics-based authentication systems are presented in section 3. Results of our experiments are analyzed in section 4 and conclusions are presented in Section 5.

2 Performance evaluation and the bootstrap

With fingerprint as an example biometrics, we introduce the performance evaluation and bootstrap technique for computing confidence measures. It can be easily extended to other biometrics-based authentication systems. To evaluate a fingerprint authentication system, a set of matching fingerprint pairs $\mathbf{M}_0 = \{a_1, \dots, a_m\}$ and a set of non-matching pairs $\mathbf{M}_1 = \{b_1, \dots, b_n\}$ needs to be acquired. Here $m < n$ because collecting prints from subjects always results in more non-matching than matching pairs. (From this it immediately follows that the FAR can be estimated more accurately.) Matching these sets of pairs results in m scores of matching fingers and n scores of non-matching fingers. We denote these sets of scores by $\mathbf{X} = \{X_1, \dots, X_m\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_n\}$, respectively.

Let us concentrate on the set \mathbf{X} and assume that this is a sample of m numbers drawn from a population with distribution F . That is, $F(x) = \text{Prob}(X \leq x)$, with x being any real number. The function F is the probability distribution or cumulative distribution function of match scores s of matching pairs. We can study this sample \mathbf{X} in order to estimate a certain characteristic, $\theta(F)$, associated with F . Just as the distribution function F of matching scores, the distribution of $\theta(F)$ is of unknown form. A statistic, $T = T(\mathbf{X})$ may be used to estimate $\theta(F)$ from the data \mathbf{X} . (Here we assume that T is unbiased.)

What is needed then is a measure of the statistical accuracy of the point estimator $T(\mathbf{X})$. This because, in general, the estimator $T = T(\mathbf{X})$ is not equal to $\theta(F)$ and we would like to get some idea of the statistical properties of the error $T(\mathbf{X}) - \theta(F)$. In other words, we are interested in how much importance should be given to T . One way to achieve this is to compute the $(1 - \alpha)100\%$ confidence interval for $T(\mathbf{X})$ in the form $[q^*(\alpha/2), q^*(1 - \alpha/2)]$ where $q^*(\alpha/2)$ and $q^*(1 - \alpha/2)$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\text{Dist}_T(x) = \text{Prob}(T(\mathbf{X}) \leq x)$. The bootstrap principle then prescribes sampling with replacement the set \mathbf{X} B times, amounting to the sets \mathbf{X}_i^* , $i = 1, \dots, B$ and calculating many estimates $T_i^* = T_i^*(\mathbf{X}_i^*)$, $i = 1, \dots, B$. The

bootstrap then amounts to what essentially is a counting exercise. Probability distributions can be computed by counting the percentage of estimates T^* that are smaller than x ; quantiles q^* can be computed by determining the value x for which a percentage q^* of the T^* is smaller than x .

In this paper, we are interested in estimating the probability distribution $F(x)$ at some point x_o , that is, a characteristic of F , $\theta(F) = F(x_o)$. Here we have to do the best with what we have, i.e., the observed sample $\mathbf{X} = (X_1, \dots, X_m)$, since we do not have the whole population X . The sample population \mathbf{X} has distribution \hat{F} , which is called empirical distribution and is defined as

$$\hat{F}(x) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}(X_i \leq x) = \frac{1}{m} (\# X_i \leq x)$$

which puts equal mass $1/m$ at each observation x_i .

What we are interested in is $\theta(F) = F(x_o)$, but all that we can obtain is an estimate $\hat{F}(x_o)$. This estimate *is itself* a random variable which has distribution $\hat{G}(y) = \text{Prob}(Y \leq y) = \text{Prob}(F(x_o) \leq y)$. We can obtain a bootstrap confidence interval for $\hat{F}(x_o)$ by sampling with replacement from \mathbf{X} as follows.

1. Calculate the estimate $\hat{F}(x_o)$ from the sample \mathbf{X} .
2. *Resampling.* Create a bootstrap sample $\mathbf{X}^* = \{X_1^*, \dots, X_m^*\}$ by sampling \mathbf{X} with replacement.
3. *Bootstrap estimate.* Calculate $\hat{F}^*(x_o)$ from \mathbf{X}^* .
4. *Repetition.* Repeat steps 2-3 B times (B large), resulting in $\hat{F}_1^*(x_o), \hat{F}_2^*(x_o) \dots \hat{F}_B^*(x_o)$.

The distribution \hat{G}^* of $\hat{F}(x_o)$ is then given by

$$\hat{G}^*(y) = \frac{1}{B} \sum_{i=1}^B \mathbf{1}(\hat{F}_i^*(x_o) \leq y) = \frac{1}{B} (\#\hat{F}^*(x_o) \leq y)$$

To obtain a bootstrap estimate of a confidence interval of $\hat{F}(x_o)$, we sort the bootstrap estimates into increasing order to obtain $\hat{F}_{(1)}^*(x_o) \leq \hat{F}_{(2)}^*(x_o) \leq \dots \leq \hat{F}_{(B)}^*(x_o)$. A $(1 - \alpha)100\%$ bootstrap confidence interval is then $(\hat{F}_{(q_1)}^*(x_o), \hat{F}_{(q_2)}^*(x_o))$, where $q_1 = \lfloor B\alpha/2 \rfloor$ the integer part of $B\alpha/2$ and $q_2 = B - q_1 + 1$.

The bootstrap supplies very powerful methods for estimating the distribution of estimates of parameters or characteristics of unknown distributions. This, in turn, supplies estimates of confidence intervals. In the following section we give some examples of the power of the bootstrap for evaluating fingerprint matching scores.

The bootstrap is valid for i.i.d. samples. Clearly, in many cases with small numbers of mated pairs, the fingerprint data may not be identically distributed; in other cases the data may not be independently distributed. The issue of not identically distributed or weakly dependent data can be addressed [3]. For the moment we assume that the i.i.d. assumption is true.

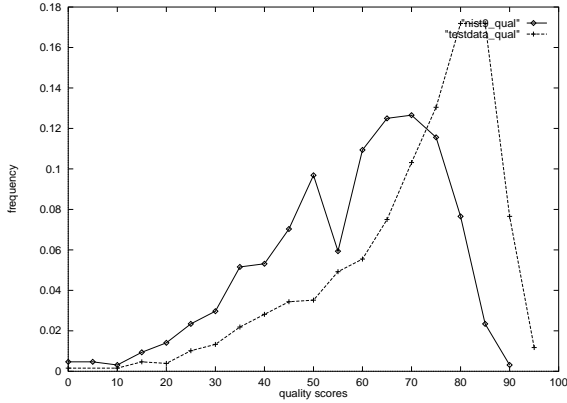


Figure 1: Image quality plots.

3 Procedures

In this section, procedures are proposed for evaluating the validity and accuracy of the error rate estimates of biometrics classifiers. We further discuss some tools for analyzing the properties of the fingerprint set that is used for error rate estimation. This is an attempt to characterize fingerprint test sets in terms of “difficulty,” and to characterize such sets for different kinds of populations.

For this purpose, we collected four samples of each finger and two fingers for 150 persons. That forms our data set for the experiments described below.

Test set characterization

Although automatic quality assessment of a fingerprint image is highly subjective, average quality of the fingerprint images in a data set is a practical index for comparing two data sets. Still a better method is to compute the histogram of image quality for the test dataset with respect to a known standard public database. Several distance measures can be used to find the distances between two histogram of image qualities. We plot the quality of 600 NIST 9 database images as the standard reference data set. The histograms of the two datasets (i.e., NIST 9 and the test dataset) are shown in Fig. 1.

There are several other methods of characterizing the test set. For example, if we have $i = 1, \dots, N$ subjects and $j = 1, \dots, M$ fingerprints per subject, we can generate $m \leq M(M - 1)/2$ matching scores $S_i = \{s_{i1}, s_{i2}, \dots, s_{im}\}$ per individual. Now, for each subject i , we compute bootstrap estimates for the mean μ and σ , $\hat{\mu}_i^*$ and $\hat{\sigma}_i^*$ respectively. Additionally, confidence intervals $[\hat{\mu}_i^-, \hat{\mu}_i^+]$ and $[\hat{\sigma}_i^-, \hat{\sigma}_i^+]$ can be established.

A graphical analysis of $\hat{\mu}$ versus $\hat{\sigma}$, and $\hat{\mu}$ versus the $\hat{\mu}$ confidence interval and the $\hat{\sigma}$ (as shown in Fig. 2)

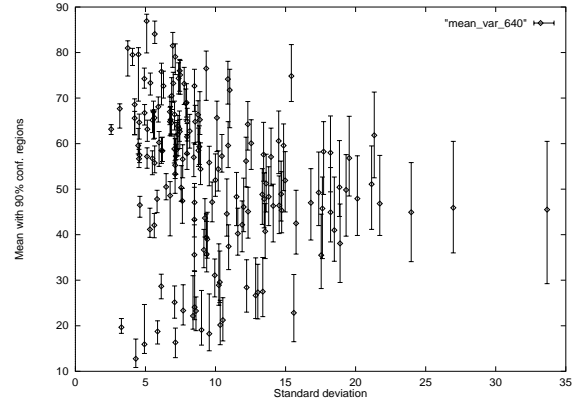


Figure 2: Mean vs. Standard deviation plot for the test data.

An ideal matcher will generate perfect scores with zero standard deviation for all matching pairs. In practice, poor quality matched pairs result in smaller scores and a significant spread in the matched scores. The spread itself may be caused by due to poor imaging conditions and/or inconsistent impression to impression variation. A statistic (e.g., $\sum \frac{\mu}{\sigma}$) could characterize the quality of the database.

The equal error rate

Given m matching scores and n scores non-matching pairs, we have an empirical distribution for the matching scores $\hat{F}(s)$ and an empirical distribution $\hat{G}(s)$ for the non-matches. The EER is given by s' with $(1 - \hat{G}(s')) = \hat{F}(s')$. From this we can put confidence intervals on the FAR and FRR at the EER, denoted FAR_{ee} and FRR_{ee} . That is, $[(1 - \max \hat{G}^*), (1 - \min \hat{G}^*)]$ is the confidence interval for FAR_{ee} , while $[\min \hat{F}^*, \max \hat{F}^*]$ is the confidence interval for FRR_{ee} . The EER for the test data set is shown in Fig. 3.

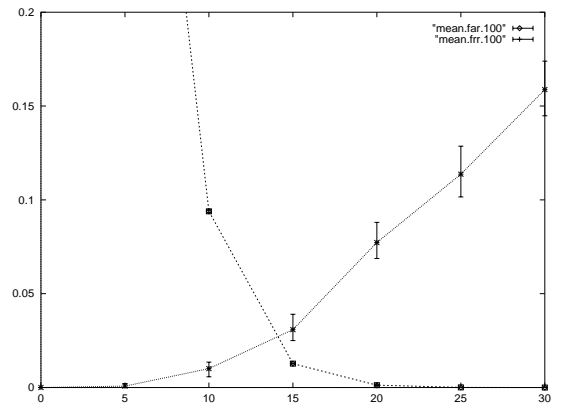


Figure 3: FAR vs. FRR and the Equal Error Rate point with confidence boxes for the test data set with 1000 bootstrap iterations.

Receiver operating curves

Similarly as for the confidence interval for the FAR and FRR at the EER, confidence intervals for both rates can be computed for any $s = T$. Hence, it is possible to plot ROC with a horizontal confidence interval for each FRR and a vertical confidence rate at each FAR. A sample ROC for the test data set is shown with 1000 bootstrap iterations in Fig. 4.

Characterizations in Fig. 3 and Fig. 4 not only depict performance of the system but also give an idea of how much variation in the performance is to be expected. A large variation in performance generally indicates poor sampling and/or poor matching model.

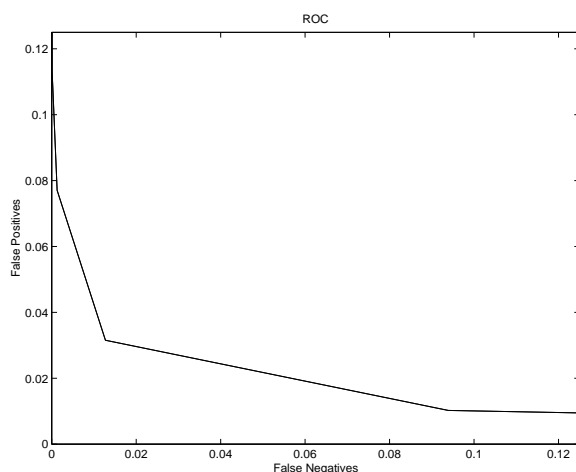


Figure 4: ROC with confidence boxes for the test data set with 1000 bootstrap iterations.

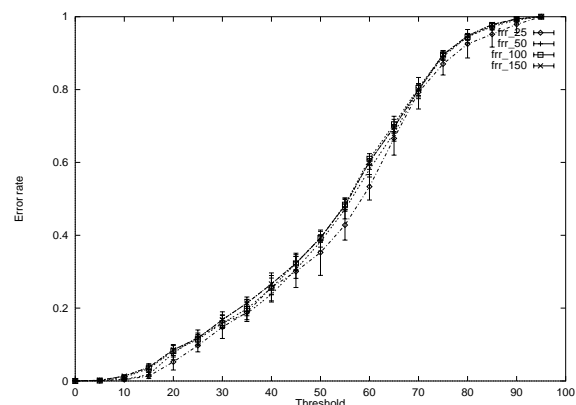


Figure 5: Confidence regions with increase in the number of samples in the test data. 25, 50, 100, and 150 fingerprints were randomly selected from the test data and error distributions were estimated from each sample using 1000 bootstrap iterations.

4 Results

As noted earlier, for the FAR, we have many more points. Hence the confidence region is narrow for FAR. For FRR, we always fewer samples than the mismatched pairs. Hence, the confidence regions are wider. As the number of samples increases, the confidence regions at a threshold asymptotically reaches the true confidence region achievable with the dataset. This is shown in the Fig. 5. The confidence prescribed by a given application dictates the number of samples to be used for evaluation.

5 Conclusions

Bootstrap techniques give powerful ways to develop confidence intervals and to perform hypothesis testing. The beauty of these techniques is that they are completely non-parametric. That is, no assumptions have to be made about the underlying forms of distribution functions.

We have used bootstrap techniques to establish two important goals. First, we use the bootstrap to estimate parameters of distributions of matching scores from collections of matches from a single person. This allows us to graphically view these distributions which gives insights into the complexity of the test sets of fingerprints. Secondly, we develop bootstrap techniques to put confidence intervals (or upper and lower bound) on distribution functions. We apply this to the distribution of genuine persons (matches) and intruders (non-matches). This allows us to determine confidence intervals for equal error rates and for receiver operating curves.

References

- [1] B. Germain et al. Issues in large scale automatic biometric identification. In *IEEE Workshop on Automatic Identification Advanced Technologies*, pages 43–46, Stony Brook, NY, November 1996.
- [2] W.W. Peterson, T.G. Birdsall, and W.C. Fox. The theory of signal detectability. *Transactions of the IRE*, PGIT-4:171–212, April 1954.
- [3] D.M. Politis. Computer-intensive methods in statistical analysis. *IEEE Signal Processing*, 15(1):39–55, January 1998.
- [4] J.L. Wayman. A scientific approach to evaluating biometric systems using mathematical methodology. In *CardTech SecureTech*, pages 477–492, Orlando, FL, May 1997.
- [5] A.M. Zoubir and B. Boashash. The bootstrap and its application in signal processing. *IEEE Signal Processing*, 15(1):56–76, January 1998.