An Evaluation of Error Confidence Interval Estimation Methods

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 accuracy performance of pattern recognition systems (e.g., biometrics ID system) is a controversial issue and perhaps an issue that is not well understood [5, 7]. This work focuses on the research issues related to the oft used confidence interval metric for performance evaluation. Using a biometric (fingerprint) authentication system, we estimate the False Reject Rates and False Accept Rates of the system using a real fingerprint dataset. We also estimate confidence intervals of these error rates using a number of parametric (e.g., see [7]) and non-parametric (e.g., bootstrapping [1, 3, 6]) methods. We attempt to assess the accuracy of the confidence intervals based on estimate and verify strategy applied to repetitive random train/test splits of the dataset. Our experiments objectively verify the hypothesis that the traditional bootstrap and parametric estimate methods are not very effective in estimating the confidence intervals and magnitude of interdependence among data may be one of the reasons for their ineffective estimates. Further, we demonstrate that the resampling the subsets of the data samples (inspired from moving block bootstrap [4]) may be one way of replicating interdependence among the data; the bootstrapping methods using such subset *resampling may indeed improve the accuracy of the estimates. Irrespective of the method of estimation, the results show that the* $(1-\alpha)100\%$ *confidence intervals empirically estimated from the training set capture significantly smaller than* $(1-\alpha)$ *fraction of the estimates obtained from the test set.*

1. Introduction

Accuracy performance evaluation of biometrics authentication or identification systems in terms of false reject rate (FRR) and false accept rate FAR is a difficult issue. These error rates in themselves do not mean much. What also needs to be reported is the dataset size that is used to compute these statistics. Some indication should be given of the quality of the dataset, e.g., the conditions under which the data were collected and a description of the subjects that are used for acquiring the database. Finally, it should be reported how accurate the estimates of the above statistics really are. All the above issues can be addressed by computing confidence intervals both on distributions and on distribution parameters. In this work, we attempt understand the practical issues related to accurate estimation of the confidence intervals.

This paper is organized as follows. Section 2 introduces terminology and confidence intervals. Sections 3 and 4 summarize the methodology for estimating confidence intervals using parametric and non-parametric methods. Section 5 presents the experimental methodology used to test the accuracies of the confidence interval estimates. We also present the data used for the experiments and the experimental results in Section 5. In Section 6, we discuss the implications of our results.

2. Confidence Intervals for Error Estimates

Suppose we have a database DB of biometric samples acquired from $\mathcal{D}$ biometrics (meaning, these are real-world biometrics, $\mathcal{B}_1, ..., \mathcal{B}_{\mathcal{D}}$) from which d samples are acquired per biometric. The number $\mathcal{D}$ of biometrics $\mathcal{B}_i$, $i = 1, ..., \mathcal{D}$ may be larger than the number of subjects $\mathcal{P}$ that are used to collect the samples, since people may have more than one of the particular biometric (e.g., finger). In any case, the database contains $d\,\mathcal{D}$ biometric samples, and given a *biometric match engine*, one can compute the test score sets: a set of genuine (match) scores $\mathbf{X} = \{X_1, X_2, ..., X_M\}$ and a set of imposter (mismatch) scores $\mathbf{Y} = \{Y_1, Y_2, ..., Y_N\}$.

Matching mated pairs in DB, i.e., matching samples from the same biometric, gives the sample match score (genuine score) set $\mathbf{X}$; matching samples in DB from different identities (or biometrics) gives the mismatch (imposter) score set $\mathbf{Y}$. In this work, as a concrete example, we focus on fingerprint databases and fingerprint matchers to illustrate the subtleties of biometric error confidence interval estimation.

A biometric match engine is in theory completely specified by its $F(s)$, the genuine score distribution, and its $G(s)$,

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

the imposter score distribution. Equivalently, the biometric matcher is completely specified by $\mathrm{FRR}(T)$ and $\mathrm{FAR}(T)$.

When estimating the $\mathrm{FRR}(T)$ and $\mathrm{FAR}(T)$ at some operating point $T = T_o$, the immediate question is how accurately these estimates are because no matter how much data we acquire, we will never be able to estimate the FAR and FRR with 100% accuracy. We will only be able to estimate these error rates within a certain $(1-\alpha)100\%$ range, or confidence interval. Here α is the probability that the true value of the FAR or the FRR are outside the respective confidence intervals. The confidence intervals are a means to assess the accuracy of the estimates of the FAR or FRR; they are measures of how much belief one may attribute to the estimates. Let us first concentrate on estimating characteristics of the match score distributions F. The mean is one such characteristic of F that can be estimated from $\mathbf{X}$; another characteristic of F that can be estimated from $\mathbf{X}$ is the value of the distribution at x_o, $\hat{F}(x_o)$, this gives the estimate of $\mathrm{FRR}(T)$ at $T = x_o$. For example, the point estimate of F at x_o is given by

$$\begin{aligned} \hat{F}(x_o) = &\quad \mathrm{FRR}'(x_o) = \tfrac{1}{M}\textstyle\sum_{i=1}^{M} \mathbf{1}\,(X_i \le x_o) \\ = &\quad \tfrac{1}{M}\#(\,X_i \le x_o). \end{aligned} \tag{1}$$

3. Parametric confidence intervals

Let us define Z as a binomial random variable, the number of successes in M trials with probability of success $F(x_o) = Prob\,(X \le x_o)$ (i.e., success $\equiv$ $(X \le x_o)$). This random variable Z has mass distribution

$$P(Z = z) \quad = \begin{pmatrix} M \\ z \end{pmatrix} F(x_o)^z (1 - F(x_o))^{M-z},$$

where $z = 0, ..., M$. The expectation of Z, $\mathrm{E}(Z) = MF(x_o)$ and the variance of Z, $\sigma(Z) = MF(x_o)(1 - F(x_o))$. For large M, $\hat{F}(x)$ is normally distributed, with an estimate of the variance given by

$$\hat{\sigma}(x) = \sqrt{\frac{\hat{F}(x)(1-\hat{F}(x))}{M}}. \tag{2}$$

So, confidence intervals can be determined with percentiles of the normal distribution, e.g., a 90% interval of confidence is

$$-1.645\,\hat{\sigma}(x) < \hat{F}(x) < 1.645\,\hat{\sigma}(x) \tag{3}$$

Estimates $\hat{G}(y)$ for the probability distribution $G(x) = Prob(Y \le y)$ using a set of mismatch scores $\mathbf{Y}$ can be obtained in a similar fashion.

4. Non-parametric Confidence Interval Estimation

Let us assume the set $\mathbf{X}$ can be divided into $\mathcal{K}$ subsets $\mathbf{X} = \{\mathcal{X}_1, ..., \mathcal{X}_\mathcal{K}\}$.

A bootstrap estimate (see [2]) of a $(1-\alpha)100\%$ confidence interval for the estimate $\hat{F}(x_o)$ is obtained as follows:

1. Divide the set of match scores $\mathbf{X}$ into $\mathcal{K}$ subsets $\mathcal{X}_1, ...,$ $\mathcal{X}_\mathcal{K}$.
2. Many (B) times do:
   (a) Generate a bootstrap set $\mathbf{X}^\star$ by sampling $\mathcal{K}$ subsets with replacement from $\mathbf{X} = \{\mathcal{X}_1, ..., \mathcal{X}_\mathcal{K}\}$.
   (b) Compute the bootstrap estimate $\hat{F}^\star$ as
   $$\hat{F}^\star(x_o) = \frac{1}{M} \sum_{X_i \in \mathbf{X}^\star} \mathbf{1}\,(X_i \le x_o).$$
   This gives the set $\mathbf{F}^\star(x_o) = \{\hat{F}^\star_k(x_o), k = 1, ..., B\}$ of B bootstrap estimates.
3. Rank the estimates in $\mathbf{F}^\star(x_o)$:
   $$\begin{aligned} \mathbf{F}^\star(x_o) = \quad &\{\hat{F}^\star_{(1)}(x_o) \le \hat{F}^\star_{(2)}(x_o) \le ... \\ &\le \hat{F}^\star_{(B)}(x_o)\}. \end{aligned}$$
4. Eliminate the bottom $(\alpha/2)100\%$ and the top $(\alpha/2)100\%$ of estimates $\hat{F}^\star_{(k)}(x_o)$. The leftover set of estimates $\mathbf{F}^{\star\star}(x_o)$ with $B' = (1-\alpha)B$ elements gives the $(1-\alpha)100\%$ confidence interval for $\hat{F}(x_o)$.

The bootstrap sampling implicitly assumes that the data being sampled is i.i.d. and therefore, any violation of such assumption would result in inaccurate confidence intervals. In realistic (biometric) datasets, there is always significant dependence among the data. For example, the match scores generated from fingerprint impressions of a finger are not independent. Similarly, the match scores of involving different fingers of a person may be dependent. Note that the number and constitution of $\mathcal{K}$ subsets plays an important role in the estimation of confidence interval. Depending upon the magnitude of independence of each sample subset (w.r.t. other sample subsets), bootstrap resampling will be able to propagate the dependence in the data; consequently the confidence intervals will be more realistic. In this work, we have experimented with three different types of bootstrap sampling. First, each match score constitutes a (singleton) subset in itself. This is conventional bootstrap. In second case, we divide the match scores into $\mathcal{PD}$ subsets such that each subset contains match scores resulting from a single finger. We call this finger subset bootstrap. Finally, $\mathcal{P}$ subsets are constructed such that each subset consists of match scores involved with a single person only. This method of bootstrap is referred to as person subset bootstrap. Since the subsets in person bootstrap are relatively more independent than those in finger subset bootstrap, we expect that person susbset bootstrap should be able to better estimate the FRR confidence intervals. Similarly, finger

and person subsets should be able to estimate confidence intervals better than the conventional bootstrap.

The bootstrap confidence interval estimation concepts can be extended to the non-match scores as well in a straightforward fashion with one exception. Since the non-match scores involve two different fingers, it turns out that completely independent datasets cannot constructed without sacrificing portions of non-match scores. So, there is an option of either using all of the non-match score data and tolerating some amount of dependence among finger and person subsets or using very little fraction of the non-match score data while ascertaining subset data independence. In this work, we choose the former option.

5. Experiments

As mentioned elsewhere, the source of the inaccuracies in the error estimates of a matcher may be either inaccurate sampling of the target population or inaccuracies in the estimation procedure. There is no substitute for collection of the representative data and in order to arrive at the correct error estimates, a carefully designed data collection procedure must capture a representative sample of the biometric data. In this work, we assume that the data collected is representative and we attempt to compare the efficacy of different error estimation methods by sequestering a random portion of the biometric data. The non-sequestered data is first used to arrive at false positive and false negative error rate estimates and their respective confidence intervals using (i) parametric, (ii) conventional bootstrap, (iii) finger subset bootstrap, and (iv) person subset bootstrap methods. The accuracies of these confidence interval estimates is ascertained using the error rates estimated from the sequestered data. Because of the limited amount of data, the procedure of splitting the data into two independent (e.g., train and test) datasets is repeated.

1. Randomly split the number of IDs into two sets, A and B, each set containing identical number of IDs.
2. Use set A to compute the FAR_A and FRR_A confidence interval estimates.
3. Use set B to compute an estimate of FAR_B and FRR_B.
4. Check whether FAR_B estimate is within the confidence interval FAR_A and whether FRR_B estimate is within the confidence interval FRR_A.
5. By repeating steps 1-4 n number of times, obtain average estimates of probabilities $Prob(FAR_B \in CI\ of\ FAR_A)$ and $Prob(FRR_B \in CI\ of\ FRR_A)$.

We use a private data set. The data are acquired from $C = 114$ different fingers in 2 sessions 5 weeks apart. The subjects are approximately half adult males and half adult females in the age group 22-65. In each session, for each subject, 5 prints of the left and right index finger are acquired. Hence, the database contains a total of $1,140$ impressions, i.e., 10 prints of 114 fingers. The number of match scores m per finger is 90 and the number of non-match scores n per finger is $5,650$. ($M = 10,260$ and $N = 644,100$.)

The results of the experiments are summarized in Figures 1, 2 and Table 1.

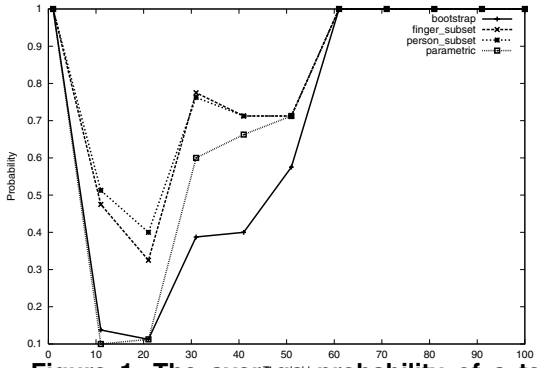

Figure 1. The average probability of a test data set FAR at a given threshold landing into the FAR Confidence intervals predicted from the training data using different estimation methods for a private data set.

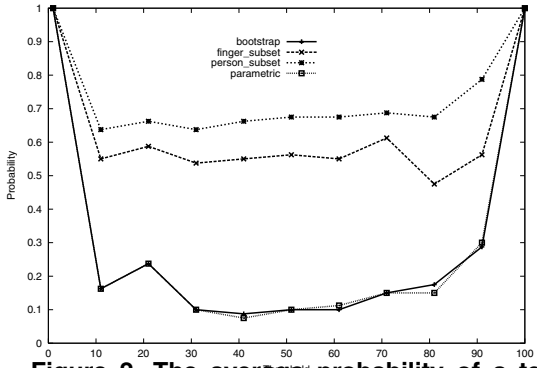

Figure 2. The average probability of a test data set FRR at a given threshold landing into the FRR Confidence intervals predicted from the training data using different estimation methods for private data set.

6. Discussion

From Table 1, it is readily observed that in a realistic situation, 90% confidence intervals estimated from the training set data do not capture 90% of the estimates from the

Error / Estimate	$FRR\,(\%)$	$FAR\,(\%)$
Parametric	76.80	32.26
Regular Bootstrap	76.49	36.49
Finger Subset	39.94	23.20
Person Subset	30.90	21.15

Table 1. On the average what percentage of times the 90% training confidence intervals failed to capture the test data for different methods of estimates based on a private dataset used in our experiments (see Figs. 1 and 2)? The ideal failure rate should be 10%.

test data. This is a surprising finding since the both the training and test data are sampled from the original database. It is also surprising that the discrepancy in the performance of the confidence intervals is conspicuously significant.

As is usual, the performance of the FRR confidence intervals is significantly inferior to the performance in the FAR confidence intervals. One reason for this is due to a signficantly smaller number of match samples available to estimate FRR than the the non-match samples available to estimate FAR. Another reason for this hiatus in performance is due to the larger variance of the match score distribution than in the non-match score distribution.

Indeed, the confidence intervals estimated using true subset bootstrap methods (e.g., finger and person subset) are significantly better than those estimated using conventional bootstrap or parametric methods. This is mostly because the parametric and conventional bootstrap methods cannot effectively model the interdependence among the data and consequently underestimate the confidence intervals.

In other words, there surely is statistical dependence among match scores $X_1, X_2, ...$ (and mismatch scores) because of the way test databases are collected. Subsequent finger impressions are obtained by successive dabbing of the finger on an input device. That is, given a first impression I of a finger plus an additional two impressions I_t and $I_{t+\Delta}$ of the same finger the match scores $X_i = s(I, I_t)$ and $X_{i+1} = s(I, I_{t+\Delta})$ are dependent. There are additional sources responsible for the dependence of the scores that are due to other subtleties of the collection process of test data sets or the subject population. In general, fingerprint image formation is a complex process and a function of many random variables (finger pressure, finger moisture, etc.), for a given individual many of these random variables are dependent from one impression to the next.

The way the traditional bootstrap sets $\mathbf{X}^\star$ are obtained from the original set of match scores $\mathbf{X}$ *does not* replicate this dependence among the X_i and there is less interdependence among match (and non-match) scores in bootstrap set $\mathbf{X}^\star$. Therefore the bootstrap estimates $\overline{X}_1{}^\star, ..., \overline{X}_B{}^\star$ have lower variance than would be the case if the match scores $\mathbf{X}$ are independent. Resampling the *subsets* of samples can alleviate this problem and typically, meaningful subset resampling can replicate the data interdependence in the bootstrap resample and facilitate the accuracy of the estimates.

Also, there is relatively smaller improvement in CI performance going from finger subset to person subset. This indicates that the person subsets model relatively less interdependence among data than the finger subset. At least one could infer that both of the subsets model similar types of data interdependence in this particular test situation.

Note further that use of subset bootstrap results a significantly more conspicuous improvement in FRR confidence interval performance than that in FRR confidence interval. Due to abundance of non-match data, the FAR confidence intervals can in general be more reliably estimated than the FRR confidence intervals and the FAR confidence interval estimation error is very small, irrespective of the method of the estimation. Consequently, there is smaller scope for improvement in FAR confidence intervals.

References

[1] K. Cho, P. Meer, and J. Cabrera. Performance assessment through bootstrap. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(11):1185–1198, Nov. 1997.

[2] B. Efron. Bootstrap methods: Another look at the Jackknife. *Ann. Statistics*, 7:1–26, 1979.

[3] A. K. Jain, R. C. Dubes, and C.-C. Chen. Bootstrap techniques for error estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(5):628–633, Sept. 1987.

[4] R. Liu and K. Singh. Moving blocks Jackknife and Bootstrap capture weak dependence. In R. LePage and L. Billard, editors, *Exploring the Limits of the Bootstrap*, pages 225–248, New York, NY, 1992. John Wiley & Sons, Inc.

[5] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. An introduction to evaluating biometric systems. *IEEE Computer*, 33(2):56–63, 2000.

[6] S. M. Weiss. Small sample error rate estimation for k-NN classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(3):285–289, Mar. 1991.

[7] J. L. Wayman. Confidence interval and test size estimation for biometric data. In *Proc. IEEE AutoID'99*, pages 177–184, Oct. 1999.