Guidelines for Appropriate Use of Simulated Data for Bio-Authentication Research

Yan Ma
Department of Statistics
West Virginia University
Morgantown, WV USA
yma@stat.wvu.edu

Michael Schuckers
Department of Mathematics,
Computer Science and Statistics
St. Lawrence University
Canton, NY USA

Bojan Cukic
Lane Department of Computer
Science and Electrical Engineering
West Virginia University
Morgantown, WV USA

Abstract

In this paper, we outline a framework for appropriate and proper usage of simulated data for biometric authentication. Currently, there are no formal guidelines concerning the use of simulated data in the biometric authentication literature. Some have suggested the usage of simulated or synthetic data while others have advised against it. Our position is that there is a place for simulation data in biometrics research but that such implementations need to meet certain requirements. To that end we describe conditions under which it is reasonable to use such data, as well as criteria for evaluating the appropriateness of a data generation methodology. This criteria is that models for generation of artificial data should be flexible, consistent and parsimonious. Along with justifying these criteria, we illustrate how simulated data might be used to evaluate a classifier.

1. Introduction

Simulated data is an artificial set of numbers which are not observed, but randomly drawn from a population with given characteristics. Simulated data allows us to work with data we do not possess. Generated data saves time and money relative to the collection of observed data, regardless of how large the data set is. However, any analysis based on the artificial data is depends implicitly on the propriety of the proposed model. To ensure proper usage of simulated data for biometric authentication, certain criteria should be satisfied. Currently there are no guidelines concerning the use of simulated data in the biometric authentication literature. Here, we present a set of robust criteria for the use of synthetic biometric data. Buetter *et al.* [26] discuss the current state of synthetic image usage. Although there is some controversy over the use of simulated biometric data, we think that there is a place for such data in biometrics research; however, such implementations need to be robust and conceptually grounded. The remainder of this paper is organized as follows. Section 2 contains a taxonomy of biometric data. Our criteria for appropriate use and modelling of synthetic data are outlined in Section 3. Section 4 provides an illustration of the utility and the goals of this paper. Finally we conclude with some remarks in Section 5.

2. Biometric Data

Biometrics and biometric data are ambiguous terms that can have multiple meanings depending on the audience. As a consequence, it is important to begin by specifying what we mean by data. Here biometrics refers to technologies for measuring and analyzing human body characteristics such as fingerprints, eye retinas/irises, voice patterns, facial images, and hand measurements, especially for authenticating an individual [8]. We will call those biological measurements *biometric data*. As stated above the term, biometric data is ambiguous and encompasses a wide range of concepts and ideas. Thus, we delineate here four types of biometric data: images, feature measurements, matching scores and decision data. Each of these four types of 'data' have received some attention from the biometric authentication

community. Each can be simulated/generated by computer algorithms resulting in synthetic biometrics.

- **Images** Fingerprint and other biometric devices consist of a scanning device which converts the scanned information into digital form. Technology to generate synthetic biometrics such as fingerprint, face and iris currently exists. Several examples of artificial generation of this type of data can be found in Cappelli *et al.* [27] or Beuttner *et al.* [**?**] or Cui *et al.* [4]. Generation of such images from templates has received much attention lately, see [1, 28]
- **Feature Measurements** Feature measurements are the measurements (intereye distance, hamming codes, number of fingerprints minutiae, etc.) made on the corresponding raw images. We are currently unaware of any attempts to directly generate synthetic feature measurements though those could easily be generated from synthetic images.
- **Matching Scores** A biometric matcher produces a matching score that either indicating the "closeness" of the input feature vector with the stored template feature vector. User authentication decision is then made based on such score(s). A critical assumption is that the data generation distribution is an accurate model of how the matching scores from a particular biometric device really are distributed. Wein and Baveja provide a recent example of this in [11]. In that paper, they use gamma distributions to generate intra-person similarity matching scores and log-normal distributions to simulate inter-person matching scores.
- **Decision Data** Every biometric device is able to make its own decision according to its own feature vector. In the context of biometrics authentication, the decision data is binary – "reject or accept as a genuine user". Schuckers [29] applied Beta-binomial synthetic decision data in his paper and did a goodness-of-fit test for to show that observed decision data followed a Beta-binomial distribution. Similarly, Schuckers *et al.* [30] used synthetic decision data to evaluate different confidence intervals methodologies that have been proposed for estimation of error rates.

Understanding the type of data that is being considered is crucial to determining the appropriateness of the data generation methodology.

3. Guidelines for Using Simulated Data

In this section, we lay out guidelines for using simulated data. The primary motivations for the use of synthetic or simulated data is a lack of observed data of a specific kind. Simulated data is capable of filling a void where more information is desired. Thus, such data can be used to test the performance of asymptotic methods on small samples, to test the scalability of systems, and to test limits of statistical methodology. All of these tests are possible with actually observed data; however, carrying out such tests are often impractical for reasons of time or money or both.

Two circumstances are worth considering here. First, when no actual data is available. In some situations, the real data is not available. But we possess some "expert knowledge" about how the real system works. We could generate artificial data based on such knowledge. Second, it is commonly the case that some actual data is available; however, either the type of data or the amount of data is lacking in some way. For example, we may want to test the performance of a particular assessment methodology but the actual data fails to display all possible parameter combinations for the methodology or other possible constraints including the need to evaluate a theoretical model in a situation where experimental test would be too costly or dangerous.

Thus, the primary impetus for using simulated data is a paucity of data of a particular type. Additionally, any simulation model should be an acceptable representation of the real system under study, [16]. And the artificially generated data should have the "appearance" of being received from the actual experiment. If inappropriate simulated data is used, the study will not produce meaningful results. Therefore, it is important to know when their usage is suitable. In what follows, we propose three criteria: flexibility, consistency and parsimony for the propriety of a simulation mechanism.

3.1. Flexibility

Simulated data are often drawn from a hypothetical population with stated characteristics. Such characteristics are described by parameter values or by a range of parameter values in the simulation model. In many applications of artificial data, we assume the behavior of simulated system can be explained by one or more probability density functions (pdf's). It is important that the distributions being considered for random generation have enough parameters and that these parameters are robust enough to model the data under study. Then, the simulation study can proceeded via sampling from the pdf's [25].

3.2. Parsimony

The purpose of simulation is to understand the behavior of a real world process or evaluate strategies for the operation of the process. Models that are adequate descriptions of reality can save money and time. Often the real world models tend to be complex and sometimes their evaluation is not feasible or become a time consuming task, conse-

quently, simple mathematical models can not be constructed to represent them. One goal of any data generation procedure should be that it is as simple as the data will allow but no simpler. This leads to a delicate balancing act. We want to represent the variation in the population with models capable of describing that variability but without extraneous and unnecessary parameters.

3.3. Consistency and Goodness-of-Fit

Having found a data generation model that is both flexible and simple, it is necessary to assess how well that model fits that data. In any good statistical approach the data analysis must mirror the data collection. Similarly, simulated data must be consistent with the behavior of the system as well as the data collection of the process it aims to simulate. If this is not the case, the simulation study may result in unreliable output. To validate the fitted models,an assessment as to how well the model matches any available observed data needs to be undertaken. (Clearly, if no observed data is available then assessing consistency is an impossible task.)

In the statistics literature, these methods generally fall under the heading of "Goodness-of-Fit" test, [32]. The basic idea for a goodness of fit test is that we are comparing observed data to a proposed population model usually a specific distribution or family of distributions. If the proposed population model is 'far', in a statistical sense, from the observed data then we reject that particular distribution as not an appropriate fit for the data. On the other hand, if the observed data is sufficiently 'close' then we fail to reject the proposed model and conclude that it is consistent with the observed data. Several approaches to goodness-of-fit can be found in the statistics literature including Chi-squared tests, Kolmogorov-Smirnov tests and Anderson-Darling tests. Schuckers [29] applies a goodness-of-fit test to biometric decision data. There are also graphical statistical methods for making comparisons between observed data and proposed distributions. The most popular among these is the QQ-plot or quantile-quantile plot. Daugman [31] uses such a plot to compare the distribution of match scores to a theoretical model.

4. An Illustration

In this section, we present an examples of both the appropriate uses of distribution models for generation of synthetic data as well as the utility of such methods. We demonstrate the ability to model match scores from three modalities: hand, finger and face. These models then serve as generators for the multi-modal random forest evaluation. We will use observed data from [12] to illustrate the consistency criteria we developed above. Thus, we fit models to the observed data and the goodness-of-fit is tested. This is followed by a description of a Random Forests algorithm implemented using the synthetic data.

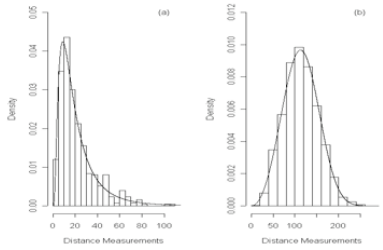

Figure 1. Histogram of facial image matching scores with the density of fitted distributions superimposed. (a). Genuine matching scores (b). Imposter matching scores.

A database of multimodal matching scores was collected at Michigan State University[1]. There are 500 genuine scores — 10 scores each for 50 individuals — and 12,250 imposter scores — 5 scores each for 500×499 cross comparisons — obtained from each of three modalities: *face*, *fingerprint* and *hand geometry*. The details on data collection can be found in [12]. The *face* and *hand geometry* scores are distance scores while the *fingerprint* verification scores are similarity scores. For each of six sets, efforts are made to find the *best* fitted distribution. The final results are summarized as follows.

The histogram of facial image genuine matching scores ((a) in Fig. 1) indicates a positively skewed distribution, with measurements from 0.00 to 106.00 in the data. More than half of the scores are less than 20.00. Since the scores are distance measurements, the lower the score, the higher the likelihood of matching. Therefore, the peak of the distribution occurs in the lower end. A log-normal distribution with density function (1) is fitted to the distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(logx-\mu)^2}{2\sigma^2}}, x \in (-\infty, \infty) \qquad (1)$$

where μ and σ are the mean and standard deviation of the logarithm. 2.7939 and 0.7803 are the maximum likelihood estimates of the mean and standard deviation of the distribution on the log scale. The density of this fitted distribution superimposed to the histogram shows a good fit. Results for a goodness-of-fit test are given below. The histogram labelled (b) in Figure (1) shows a relatively symmetric distribution. The Weibull distribution with shape parameter α and scale parameter β is a reasonable candidate

1 Data used with permission

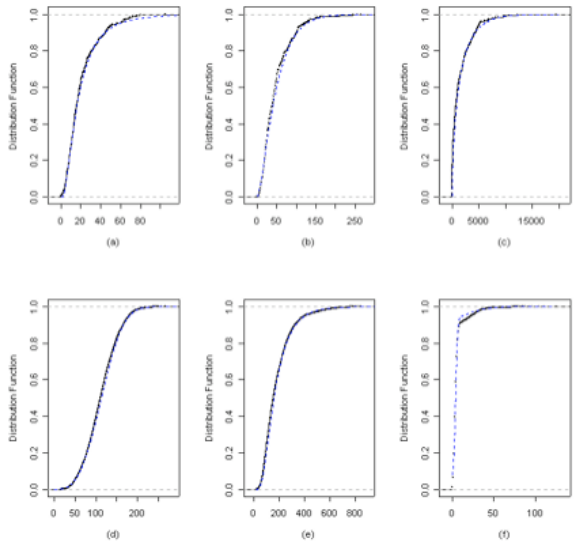

Figure 2. Empirical cumulative distribution and theoretical distribution (a). Face genuine (b). Hand geometry genuine (c). Fingerprint genuine (d). Face imposter (e). Hand geometry imposter (f). Fingerprint imposter

to model this data. The density is given by

$$f(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}e^{-(\frac{x}{\beta})^{\alpha}} \quad (2)$$

for x $>$ 0. The maximum likelihood estimates of the shape and scale parameters are 3.1627 and 127.4522, respectively. Similar graphs can me made of hand geometry distributions and the fingerprint distributions. Likewise, we can use estimation to calculate parameters for the models of the fingerprint and hand geometry population models. Note that for the hand geometry and facial match score data models with either two or three parameters were necessary and, thus, these models were both flexible and simple. (Again, a model with three parameters is relatively simple in a statistical modelling sense.) In the case of the fingerprint match score data, it was necessary to transform the genuine match scores and to fit the imposter match scores by use of a mixture of two Gaussian distributions.

To access the goodness of fit for the distributions fit to the above match score data, we first generate the empirical cumulative distribution function for each set and map it to the fitted distribution, then we carry out a statistical goodness-of-fit test.

Figure 4 shows the overlaid empirical distribution function and the population distribution. The fitted distribution is represented by dashed lines. It can be seen that the empirical distributions are all very close to the fitted distributions,

Modality		
Population	**Fitted Distribution**	***p*-value**
Face		
Genuine	Log-Normal	0.3030
Imposter	Weibull	0.1800
Hand Geometry		
Genuine	Gamma	0.1640
Imposter	Log-Normal	0.8830
Fingerprint		
Genuine (transformed)	Gamma	0.6960
Imposter	Truncated Mixture Normal	0.2010

Table 1. Match Score Goodness-of-Fit Test Summary

providing evidence of the goodness of fit of the distributions described in Table 4 for each set. We then perform a parametric bootstrap test based on the Kolmogorov-Smirnov test statistic. And this test is an application of the Monte Carlo approach. The fitted distribution parameters are estimated from the actual data, thus we are unable to assess the fitting directly based on the Kolmogorov-Smirnov tests. The bootstrap test is performed as follows.

- Start with the null hypothesis of data coming from a population with specific distribution models obtained above and the alternative is the opposite.
- Perform a one-sample Kolmogorov-Smirnov test, compute the calculated test statistic, t_{calc}.
- Generate random numbers of the same size as the actual data set from the fitted distribution.
- Perform Kolmogorov-Smirnov test see if the random sample generated in step 3 is drawn from the hypothetical distribution.
- Repeat step 3 and 4 N steps (N needs to be large) resulting in N test statistics t_1, t_2, ..., t_N. The *p*-value of the bootstrap test is defined to be $\sum_i^N I(t_i \geq t_{calc})$ / N, where I(. . .) = 1 if $t_i \geq t_{calc}$; 0 otherwise.
- A large *p*-value indicates evidence in favor of the null hypothesis.

Since all of the *p*-values are large, we can conclude that the distribution models are consistent with the observed data. We now turn to using this simulated data to evaluate a Random Forest classification scheme.

One area where the use of synthetic data is appropriate is the integrating of information from multiple modalities. The Random Forest algorithm, developed by Leo Breiman in 2001, has already been introduced in some fields due

to its encouraging predictive accuracy. However, the application of this algorithm in biometrics is still relatively new. Here, we apply the Random Forests (RF) algorithm on match score data simulated from the models described above. The purpose is to evaluate the stability of this algorithm and to see if the classifier continues to perform in the same way with different data sets.

RF is an ensemble method in the sense that instead of growing a single classification tree, we could build hundreds to thousands of trees. An improved classifier is obtained by integrating tree models in the forest. Each single tree is grown as follows [19, 21]:

- Take a bootstrap sample from the original data and the root node of the tree contains this sample instead of the original data.
- At each node of the tree, except for the terminal nodes, randomly select a subset of the variables, then the locally optimal split is based on only this feature subset. Grow the tree as large as possible with no pruning.

Every tree in a forest of N trees possesses certain classification rule. Given a new case with matching score vector $\boldsymbol{x} = \{x_1, x_2, \ldots, x_P\}$, we begin with tree 1 in the forest. The search starts from the root, the splitting rule is checked and the case is sent to one of the children node according to the rule. This is repeated until the terminal node is reached and the class label attached to the terminal node is assigned to this case. Thus, tree 1 has made its decision. Then we go to tree 2, do the same thing and find the class label for this case. Upon tried every single tree in the forest, we have N votes. The trees raise their own opinion, fight with each other and the majority wins.

The procedures for growing a single tree outlined above randomize inputs in model building. Therefore, the trees will have different looks. And the low correlation lowers the classification error rate of RF [20]. For each bootstrapped sample, about one-third of the cases are not used in the tree construction. These left-out cases are called out-of-bag (OOB) cases which are playing an important role in performance assessment.

In a random forest, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error [21]. The OOB cases can be used as a testing set. For the j^{th} tree in the forest, put the OOB cases down the tree to get a test set classification and repeat this to all the trees. Let the final test set classification of the forest be the class having the most votes. Comparing the classification with the class label present in the data gives an estimate of the test set error rate. An unbiased estimate of the misclassification rate is thus obtained automatically and internally during the run.

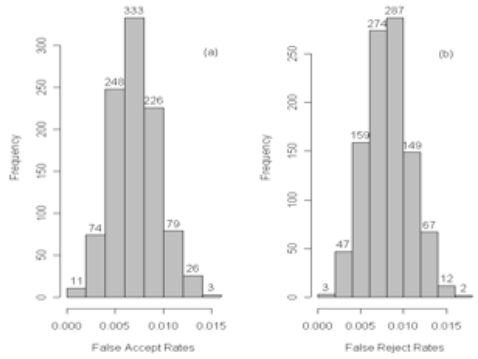

Figure 3. Histogram of the testing rates by applying random forests algorithm with frequencies labelled put on the top of histogram. (a). false accepts testing rate (b). false reject testing rates

1000 samples of genuine and imposter cases are created and the matching scores from each modality are random numbers generated from corresponding fitted distributions. With each simulated data, we apply random forests algorithm to obtain the misclassification error rates. The histogram of false accept rates (FAR) and false reject rates (FRR) from these 1000 simulated sets are shown in Figure 4. The FARs range from 0.001 to 0.016 with more than 80% of the rates less than 0.01. The FRRs range from 0.001 to 0.017 with at least 60% less then 0.01 meaning that RF classifier is able to correctly identify an genuine user with 99% confidence.

5. Conclusions

Simulated data, appropriately used, can be a useful research tool. It is important to know when their usage is appropriate. Here we have laid out three criteria to be used when considering artificial data. Models for data synthesis should be flexible, parsimonious and consistent with available data. We want to choose as simple a data generation models as possible while also being flexible enough to fit a range of data. When observed data is available, it is necessary to ensure that the population models are consistent with this data. Several methods are available for ensuring that a quality fit is made. Further we have illustrated the utility of such data having first ensured that our models for data generation met the criteria given here. This was done for a Random Forest classifier.

References

[1] Andy Adler. *Images can be Regenerated from Quantized Biometric Match Score Data* Canadian Conference on Electrical and Computer Engineering (CCECE), 469-472 Niagara Falls, Canada, May 2004.

[2] Svetlana N. Yanushkevich, Adrian Stoica, Sargur N. Srihari, Vlad P. Shmerko and Marina L. Gavrilova: *Simulation of Biometric Information: The New Generation of Biometric System*, available at: http://btlab.enel.ucalgary.ca/docs/pdf/BT04.pdf.

[3] D. Maltoni, D. Maio, A. K. Jain and S. Prabhakar: *Handbook of Fingerprint Recognition*, Springer, 2003

[4] Jiali Cui, Yunhong Wang, Junzhou Huang, Tieniu Tan, Zhenan Sun, and Li Ma: *An Iris Image Synthesis Method based on PCA and Super-resolution*, International Conference on Pattern Recognition, 2004.

[5] Umut Uludag and A.K. Jain: *Attacks o Biometric Systems: A Case Study in Fingerprints*, available at: http://www.biometricscatalog.org/documents/EI5306-62-manuscript-5.pdf.

[6] B. Chalmond: *Modeling and Inverse Problems in Image Analysis*, Springer, Applied Mathematical Sciences, Vol. 155, 2003.

[7] Michael E. Schuckers: *Estimation and Sample Size Calculations for Matching Performance of Biometric Authentication*, submitted to Pattern Recognition, May 2005.

[8] http://searchsecurity.techtarget.com/sDefinition

[9] Anil K. Jain, Arun Ross and Salil Prabhakar: *An Introduction to Biometric Recognition*, in IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics, Vol. 14, No. 1, January 2004.

[10] Nicholas M. Orlans, Douglas J. Buettner and Joe Marques: *Survey of Synthetic Biometrics*, available at: http://www.mitretek.org/biometricslides/BID_Announcement_110804.ppt

[11] Lawrence M. Wein and Manas Baveja: *Using Fingerprint Image Quality to Improve the Identification Performance of the U.S. Visitor and Immigrant Status Indicator Technology Program*, in Proceedings of the National Academy of Sciences, Vol. 102, No. 21, May 2005.

[12] Arun Ross and Anil Jain: *Information Fusion in Biometrics*, Pattern Recognition Letters 24, 2115-2125, 2003

[13] Ruud M. Bolle, Sharath Pankanti and Nalini K. Ratha: *Evaluation Techniques for Biometrics-Based Authentication Systems (FRR)*, in Proceedings of International Conference on Pattern Recognition (ICPR'00)-Vol. 2, Barcelona, Spain, 2000

[14] Chakravart, Laha, and Roy: *Handbook of Methods of Applied Statistics*, Vol. I, John Wiley, pp. 392-394, 1967

[15] George W. Snedecor and William G. Cochran: *Statistical Methods*, 8th Edition, Iowa State University Press, 1989

[16] Jack P.C. Kleijnen: *Validation of Models: Statistical Techniques and Data Availability*, in the Proceedings of the 1999 Winter Simulation Conference, P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds

[17] James E. Gentle: *Random Number Generation and Monte Carlo Methods*, 2nd Edition, Springer-Verlag, 2003

[18] Jun S. Liu: *Monte Carlo Strategies in Scientific Computing*, Springer, New York 2001

[19] Leo Breiman: *Random Forests*, Machine Learning, Vol. 45, 5-32, 2001

[20] Leo Breiman: *Wald Lecture II, Looking Inside the Black Box*, available at: http://www.stat.berkeley.edu/users/breiman

[21] Leo Breiman and Adele Cutler: *Random Forests: Classification/Clustering*, available at: http://www.stat.berkeley.edu/users/breiman/RandomForests, 2004

[22] Mike Brandl and Ray G. Van Ausdal: *Generating Simulated data*, available at: http://

[23] Sheldon M. Ross, *Simulation* 3rd Edition, Academic Press, January, 2002

[24] Stephens, M. A. (1974). *EDF Statistics for Goodness of Fit and Some Comparisons*, Journal of the American Statistical Association, Vol. 69, 730-737, 1974

[25] *Introduction to Monte Carlo Algorithm*, available at http://www.cse.msu.edu/ñandakum/nrg/Tms/tms2/intro.htm

[26] Nicholas M. Orlans and Douglas J. Buettner and Joe Marques: *A Survey of Synthetic Biometrics: Capabilities and Benefits* , Proceedings of the International Conference on Artificial Intelligence (IC-AI'04), CSREA Press, Vol. I, 499-505, 2004.

[27] R. Cappelli and D. Maio and D. Maltoni and L. Nanni, *A two-stage fingerprint classification system* , Proceedings ACM SIGMM Multimedia Biometrics Methods and Applications Workshop (WBMA03),95-99, 2003

[28] Arun Ross and J. Shah, and Anil K. Jain, *Towards Reconstructing Fingerprints from Minutiae Points*, Proceedings of SPIE Conference on Biometric Technology for Human Identification II, (Orlando, USA), 68-80, 2005.

[29] Michael E. Schuckers, *Estimation and sample size calculations for correlated binary error rates of biometric identification rates*,Proceedings of the American Statistical Association: Biometrics Section [CD-ROM], American Statistical Association, 2003.

[30] , Michael E. Schuckers and Anne Hawley and Katie Livingstone and Nona Mramba and Collen J. Knickerbocker, *A comparison of statistical methods for evaluating matching performance of a biometric identification device- a preliminary report*, Proceedings of SPIE Conference on Biometric Technology for Human Identification, (Orlando, USA), 144-155, 2004.

[31] John Daugman, *The importance of being random: Statistical principles of iris recognition*, Pattern Recognition, 36(2), 279-291, 2003.

[32] Dennis D. Wackerly and William Mendenhall III and Richard L. Scheaffer, *Mathematical Statistics with Applications*, Duxbury, 6^{th} edition, 2002.