On Using Error Bounds to Optimize Cost-Sensitive Multimodal Biometric Authentication

Norman Poh and Josef Kittler
CVSSP, University of Surrey, Guildford, GU2 7XH Surrey, U.K.
normanpoh@ieee.org, j.kittler@surrey.ac.uk

Abstract

While using more biometric traits in multimodal biometric fusion can effectively increase the system robustness, often, the cost associated to adding additional systems is not considered. In this paper, we propose an algorithm that can efficiently bound the biometric system error. This helps not only to speed up the search for the optimal system configuration by an order of magnitude but also unexpectedly to enhance the robustness to population mismatch. This suggests that bounding the error of biometric system from above can possibly be better than directly estimating it from the data. The latter strategy can be susceptible to spurious biometric samples and the particular choice of users. The efficiency of the proposal is achieved thanks to the use of Chernoff bound in estimating the authentication error. Unfortunately, such a bound assumes that the match scores are normally distributed, which is not necessarily the correct distribution model. We propose to transform simultaneously the class conditional match scores (genuine user or imposter scores) into ones that are more conforming to normal distributions using a modified criterion of the Box-Cox transform.

1 Introduction

While using multiple biometric systems can increase the system performance, each additional biometric device and/or matching algorithm will certainly add to the hardware cost and its maintenance. Furthermore, there are also intangible costs such as increased acquisition and processing time, hence resulting in a lower throughput, i.e., the number of verification requests that can be performed within a specified period. Therefore, pragmatically, one should consider a trade-off between performance and the associated cost. The goal is therefore to maximize the performance (or minimize recognition error) subject to maximum permissible costs. This can be considered a constrained optimization problem because one cannot minimize both the error and the cost simultaneously.

An optimization procedure applicable to the aforementioned trade-off problem requires repeated evaluation of the fusion performance of a selected subset of biometric systems. This problem can be cast as one of feature selection. A direct approach to this problem is to to perform empirical evaluation of each candidate combined subsystem, e.g., [4]. As an alternative, we propose to estimate the performance by modeling the distribution of the match scores. The classification error of the systems or its bound is then estimated analytically from the model. Compared to the empirical approach, the model-based approach has two advantages. First, the complexity of using error bound is significantly lower. As in any feature selection problem, with $N$ system outputs, the complexity of the search space is in the order of $2^N$. The empirical approach will have a complexity of $O(2^N \times L)$ where $L$ is the number of training (and validation) data. The model-based approach, on the other hand, has a complexity of $O(2^N + L)$. The addition between the two complexity term is due to the fact that all samples are used once and for all to estimate the model parameters, and they are not used when evaluating the criterion; only the model parameters are. Second, the empirically estimated performance is very sensitive to the choice of biometric samples and also the choice of population of users. While the model based approach does not completely remove such a dependency, it certainly provides some degree of robustness. This conjecture is well supported by our experiments.

Our approach can be seen as a refinement to [3] which predicts the performance of a linear classifier given the class conditional Gaussian assumption. However, instead of using a linear classifier, the weights of which have to be found by optimizing a criterion, we propose directly to estimate the error, using a well known result called the Chernoff bound [2]. We modified this bound slightly in order to bound the authentication error from above.

A possible weakness of the model-based approach is that when the assumption on the match score distribution is violated, the estimated performance may be inaccurate. We therefore propose to improve the estimation by first transforming the match scores such that both class conditional match scores conform better to the normal distribution. This is done by using the Box-Cox transform [1] with a modified optimization criterion.

This paper is organized as follows: Section 2 describes the model-based approach. We will first address how the class conditional match scores can be made more conforming to the normal distribution using the Box-Cox transform. Then, we will discuss the Chernoff bound, present the proposed algorithm and analyze its complexity. The database, reference systems and experimental protocols are presented in Section 3. This is followed by experimental results in Section 4 and conclusions in Section 5.
2 Biometric Error Bound

2.1 The Box-Cox Transform

Let $y$ be the variable describing a match score. We also introduce $y|k$ to be a variable representing the class-conditional match scores, and $k \in \{C, I\}$ can be either genuine user (client) or impostor.

The Box-Cox transform [1] is given by

$$\tau(y, \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

The optimal value of $\lambda$ can be found by optimizing a criterion that tests the deviation of $y$ from a normal distribution. This can be done using the Kolmogorov-Smirnov test, defined as the maximum difference between two cumulative density functions (cdfs), one being the empirical cdf of $y$, i.e., $F(y)$ and the other being the cdf of $y$ drawn from a standard normal distribution, i.e., $G(y)$:

$$\text{KS}(y) = \max_y |F(y) - G(y)|$$

The larger the KS value, the further $y$ deviates from a normal distribution.

For our application, we have to find the value of $\lambda$ which jointly transforms both class conditional distributions $p(y|k)$ to normal distributions, which is required in Section 2.2. It is therefore necessary to modify the optimization criterion for $\lambda$.

One way to achieve this is to optimize $\lambda$ such that the average KS value for both the genuine user and impostor match scores is minimum, i.e.:

$$\lambda_* = \arg \min_{\lambda \in \mathbb{R}} \left\{ \frac{1}{2} \sum_{k \in \{C, I\}} \text{KS}(\tau(y|k, \lambda)) \right\}.$$  

Once the optimal value of $\lambda_*$ is found, we can apply the Box-Cox transform on the unconditional scores $y$. Note that (2) does not guarantee that the conditional score $y|k$ is normal but only guarantees that the average KS value of the conditional scores is minimal.

An example of optimization of $\lambda$ for match scores of three biometric modalities are shown in Figure 1. The match score data set will be described in Section 3.

2.2 Chernoff Bound

In this Section, we shall treat the multivariate match score $y = [y_1, \ldots, y_N]^T$ where $y_i$ is the output of the $i$-th biometric system. We assume that the class conditional distributions $p(y|k)$ for all $k$’s are normal. If they are not, the Box-Cox transform described in the previous section will be used to pre-process each $y_i$ independently. For the sake of notational economy, we shall not introduce a different notation for the Box-Cox transformed $y$.

The aim of this section is to provide an upper bound of the minimum Bayes error given $p(y|k)$. This problem is well studied for a two class problem when $p(y|k)$ is a multivariate normal distribution for each $k$ [2], with mean $\mu_k$ and covariance $\Sigma_k$. By adapting the Chernoff measure of class separability, we can bound the classification error from above by the so called Chernoff bound. We shall use this bound to further derive an upper bound for Half Total Error Rate (HTER) as a special case. HTER is defined as the average of False Acceptance Rate (FAR) and False Rejection Rate (FRR), both of which measure how probable a system accepts an impostor claim and rejects a genuine claim, respectively.

In a two class problem, the probability of making an error given observation $y$ is:

$$P(\text{error}|y) = \begin{cases} P(I|y) & \text{if decision is accept} \\ P(C|y) & \text{if decision is reject} \end{cases} = \min[P(I|y), P(C|y)].$$

(3)

Note that this is the minimal possible error, or minimal Bayes error. According to [2, Chap. 2], this quantity can be bounded as follows:

$$P(\text{error}) \leq P^\beta(1)P^{1-\beta}(C) \min_\beta \{\exp(-k(\beta))\}.$$  

where

$$k(\beta) = \frac{\beta(1-\beta)}{2} (\mu_C - \mu_I)^T \Sigma^*_C(\beta)^{-1} (\mu_C - \mu_I)$$

$$+ \frac{1}{2} \log \left| \frac{\Sigma^*_C(\beta)}{\beta \Sigma^*_I + (1-\beta) \Sigma^*_C} \right|.$$  

(4)

The advantage of introducing an upper bound via $\beta$ is that the search is not dependent on the $N$ dimensional spaces of $y$ but on a single dimension spanned by $\beta$. The search of $\beta$ is straightforward since $\beta$ is bounded: $\beta \in [0,1]$.

A special case of error bound, called the Bhattacharyya bound [2] is given by $\beta = 0.5$. This parameter value is practical because it does not involve any numerical search but may result in a looser estimate of the error bound.

Finally, one can calculate HTER from $P(\text{error}|\beta)$ by setting $P(C) = P(I) = 0.5$. In this case, the following inequality for HTER is true:

$$\text{HTER}_{\min} = \frac{1}{2} \int \min [p(y|I), p(y|C)] dy$$

(5)

$$\leq \frac{1}{2} \min_\beta \exp(-k(\beta))$$

In theory, the following inequality holds:

$$\text{HTER}_{\min} \leq \frac{1}{2} \min_\beta \{\exp(-k(\beta))\} \leq \frac{1}{2} \exp(-k(0.5))$$

(6)

A related measure, called Equal Error Rate (EER), further requires that $p(y|I) = p(y|C)$ in addition to satisfying the minimization
constraint in (5). Since $\text{HTER}_{\text{min}}$ is the minimum possible value of HTER by definition, with the additional constraint, the following constraint will hold:

$$\text{HTER}_{\text{min}} \leq \text{EER}$$

In practice, we find that both these errors are very similar.

### 2.3 Cost-Sensitive System Subset Selection

The proposed algorithm can be summarized as follows:

1. For each system $i = 1, \ldots, N$:
   - Calculate the optimal $\lambda$ for $y_i$ using (2)
   - Transform $y_i$ using (1):
     $$y_i^{\text{norm}} = \tau(y_i, \lambda_*)$$

2. Let $y = [y_1^{\text{norm}}, \ldots, y_N^{\text{norm}}]'$ where $y_i$'s is the transformed match score

3. Compute $p(y|k) = N(y|\mu_k, \Sigma_k)$

4. For each combination $y_c \in \mathcal{P}\{\{y_i|\}$) — θ (indexed by c):
   - Calculate criterion
     $$\frac{1}{2} \min_{\beta} \exp(-k(\beta|\mu_k, \Sigma_k, \forall k))$$

5. Output: arg $\text{sort}_c \{\text{criterion}(c)\}$

In Step 4, $\mathcal{P}(A)$ denotes the power set of $A$, and it returns all the possible combinations of elements in the set $A$. If there are $N$ elements in $A$, then there are $2^N$ possible combinations, made up of $\mathcal{N}C_0 + \mathcal{N}C_1 + \ldots + \mathcal{N}C_N$, where $\mathcal{N}C_k = n!/(n-k)!k!$ denotes choosing $k$ out of $N$. Note that the first combination, i.e., choosing none out of $N$ is not possible. For this reason, the combinations considered here should not have empty set, i.e., $\mathcal{P}(A) - \emptyset$ and this subset contains $2^N - 1$ combinations, indexed by $c = 1, \ldots, 2^N - 1$.

For the $c$-th combination, we also introduce the parameters $\mu_k^c$ and $\Sigma_k^c$ to denote the distribution

$$p(y_c|\mu_k^c, \Sigma_k^c) \equiv \int_{y_c} p(y_c, y_c|\mu_k^c, \Sigma_k^c)dy_c$$ (7)

where $y_c$ denotes the elements not in $y_c$. The property in (7) makes the normal distribution particular efficient for feature selection since integration over the unobserved variable $y_c$ is carried out by simply manipulating the mean vectors and covariance matrices.

Since each combination $c$ may have a different cost, it is appropriate to define a cost function for each $c$. Let this cost function be $\text{cost}(c)$. For example, adding a biometric trait will incur an additional cost. The optimization goal can then be defined as one that maximizes the performance for a pre-specific cost $\text{cost}_0$, i.e.,

$$\min_{\{c|c \in C, \text{cost}(c) = \text{cost}_0\}} \text{criterion}(c)$$

We expect that for a sequence of increasing costs $\text{cost}_0 < \text{cost}_1 < \text{cost}_\gamma$, we have the corresponding sequence of decreasing error: $\text{criterion}_0 > \text{criterion}_1 > \text{criterion}_\gamma$.

### 3 Database, Reference Systems and Experimental Protocols

The data used in our evaluation scheme is taken from the Biosecure database. Biosecure is a European project whose aim is to integrate multi-disciplinary research efforts in biometric-based identity authentication.

For the purpose of our experiments, we used the subset of desktop scenario, which further contains a subset of still face, 6 fingers and iris modalities, denoted by fa1, ft1–6 and ir1, respectively. The development set contains 56 users whereas the evaluation set contains 156 users. In both cases, different sets of impostors are used.

Note that for the purpose of performance assessment, the main objective of this paper, the data set and experimental protocols are not the primary concern; any database could have been used. The only requirement is that a wide variety of biometric modalities are used in order to illustrate the generality of our approach.

It is important to note that there are two score data sets: development and the evaluation sets. Note that being baselines, no effort was made to optimize the system performance; the only requirement is that all systems must output match scores, or otherwise a dummy value of “-999” is produced. For the purpose of calculating the user-specific parameters, samples with these dummy values are simply removed since they are, in essence, samples with missing values.

Assigning a cost to a channel of data is a very subjective issue. In this study, we adopt the following rules of thumb:

- If a device is used at least once, a fusion algorithm will be charged a unit cost. While in reality, different devices may have different costs, we opted for a unit cost because it is simply difficult to arrive at a consensus otherwise.
- The subsequent use of the same device will be charged 0.3 of a unit in view of the fact that the same hardware is being reused.

For example, if two fingers are used, a cost of 1.3 will be charged for an access. However, if a finger and a still face are used, the cost will be 2.

### 4 Experiments

We first test the effectiveness of the Box-Cox transform. We used the match score of the development set to find the optimal $\lambda$ and then applied the transformation with the optimal parameter on the evaluation set. Using a data set for optimization and another for testing ensures that our assessment is unbiased with respect to the training data. We then calculated the KS statistics for the match scores before and after the Box-Cox transform. For all 8 systems, we found that the KS statistics are in general reduced after the transformation. For instance, for fa1, the KS statistics of the imposters is reduced from 0.0778 to 0.0428; and that of the genuine users from 0.1548 to 0.1282. See also Figure 1.

When the $p$-value of each class conditional score sets are compared to their corresponding critical values (at 95% level of confidence), we found that most data sets still were far from being Gaussian. Nevertheless, we proceeded to calculating the error bounds.

1\text{http://www.biosecure.info/}
Table 1. Pairwise correlation of error measures

<table>
<thead>
<tr>
<th>data</th>
<th>Criteria</th>
<th>Chern.</th>
<th>Bhat.</th>
<th>GMM</th>
<th>QDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>eva</td>
<td>Chern.</td>
<td>0.971</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Bhat</td>
<td>—</td>
<td>0.973</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>GMM</td>
<td>0.9667</td>
<td>0.978</td>
<td>0.951</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>QDA</td>
<td>0.9644</td>
<td>0.977</td>
<td>—</td>
<td>0.880</td>
</tr>
</tbody>
</table>

In Section 2.3, we conjectured that with increasing cost (hence acquiring more biometric traits), the error will gradually decrease. Using the exhaustive search space defined by our 8 component systems, we obtain a total of 255 fusion systems and can thus effectively trace this curve.

We used two theoretical error criteria: Chernoff and Bhattacharyya and two empirical error measures of EER based on Bayesian classifiers: Gaussian Mixture Model as a density estimator (denoted as GMM) and a Quadratic Discriminant Analysis classifier (QDA). These four measures are then calculated on both the development and evaluation data sets, which consist of different genuine and impostor population of users. For the theoretical error criteria, one only needs to fit the match score distributions using the respective (development or evaluation) data set and evaluate the error. However, for the empirical error of the development set (which applies to the GMM and QDA classifiers), a two-fold cross-validation procedure is employed. The average error of the two folds is used as an indication of the error on the development set. The trained classifier on the entire development set is then used to assess the empirical error of the evaluation set.

A table of pairwise correlations among these four error measures/criteria, conditioned on the development and evaluation sets, is shown in Table 1. Each entry of correlation has been measured on the entire fusion search space (a total of 255 fusion candidates).

Note that not all entries in the table are of interest; only the diagonal and lower-left quadrant are. The diagonal indicates how robust a given error measure/criterion in the presence of population mismatch. As can be seen, the Bhattacharyya criterion is most resilient to population mismatch. The lower left quadrant demonstrates the usefulness of the two theoretical error criteria in predicting the generalization performance (on unseen data) merely from the development data. Again, the Bhattacharyya criterion is able to predict the generalization performance of our two classifiers more accurately than their own averaged cross-validated errors (obtained from the development set).

Finally, to conclude the experiment, we assess how well a criterion/error measure does in order to select the best system to combine for a given cost. For our problem, the cost ranges from 1 (using a single system) to 4.5 (using all 8 systems). Although many fusion classifiers can be used, as examples, we consider only the GMM and QDA classifiers shown in Table 1 (since the results are readily available). We plot here a “rank-one” cost-sensitive performance curve (performance versus cost). Since the goal here is to achieve minimum generalization error with minimum cost, a curve towards the lower left corner is the desirable. This curve is called rank-one because only the generalization of the top recommended system (according to a criterion assessed on the development set) is shown here. A “rank-two” cost-sensitive curve would be the minimum of the generalization errors of the top two candidates found on the development set. With enough rank order, a performance curve will lead to the oracle one (the ideal curve with error estimated on the test set). The rank-one curve found with the Bhattacharyya is satisfactory, its rank-three curve leads to exactly the same one as the oracle for QDA and rank-five curve for GMM. Comparatively, using the averaged cross-validated empirical error, a rank-six curve is needed to achieve that of the oracle for QDA and more than rank ten is needed for GMM.

This suggests that bounding the error using the Bhattacharyya criterion is advantageous over direct estimation of empirical error (by cross-validation) under population mismatch.

5 Conclusions

In this paper, we propose an algorithm to estimate the upper bound of EER. It is based on the Chernoff and the Bhattacharyya bounds [2]. Although class conditional Gaussian assumption is inherent in these bounds, we proposed to use the Box-Cox transform to simultaneously transform both the genuine and impostor match scores to conform better to the Gaussian assumption. Although the quality of prediction is very encouraging, there is still room for improvement. First, a better approach based on user-specific score distributions can be devised in order to simultaneously transform both client and impostor scores to exhibit normal distribution more effectively. Second, instead of calculating the minimum Bayes error, one can also assess the system performance in terms of minimum risk.

References