Biometrics System Concepts - Assignment 1

Deniz Soysal - r0875700

March 25, 2022

1 Introduction

In this assignment, we will evaluate the performance of fingerprint recognition as a biometric system, by using metrics and plots. We will compare two fingers : the left and the right index.

1.1 Dataset

The dataset used is available at https://www.nist.gov/itl/iad/image-group/nist-biometric-scores-set-bssr1. The data used is similarity scores between the enrolled user fingerprint and the same user or an impostor. We have 1000 identities : The dataset is then a matrix of size 1000 x 1000.

2 Validation of verification system

In this part of the rapport, we will discuss the performance of the biometric system in the verification scenario : We have a test sample that claims an identity I. We then compute the similarity score between the features of the test sample and the features of the claimed identity (stored in the database). If the score is higher than a selected threshold, we classify the sample as "genuine". Otherwise, we classify it as an impostor. Therefore, the verification scenario can be seen as a binary classification problem.

2.1 Question 1 : Score distributions

The first metric that we will consider is the distribution of the score. In an ideal world, we want to have well-separated impostor and genuine distributions. Indeed, if the distributions are completely separated, we can just select a threshold between the 2 distributions and thus never make a mistake. But in real life applications, these 2 distributions are often overlapping, as presented in Figure 1. This lead to False Positives and False Negatives. Depending on the threshold we choose, we will change the behaviour of the system :

- By increasing the threshold, we will decrease the False Positive Rate but increase the False Negative Rate : This will lead to a more secure system (more difficult to accept a sample as genuine) but less convenient (more genuine samples will be rejected).
- By decreasing the threshold, we will decrease the False Negative Rate but increase the False Positive Rate : This will lead to a more convenient system (we will reject less genuine samples) but less secure (more impostor samples will be accepted).

So, we can see that selecting the threshold depends on the type of applicaiton

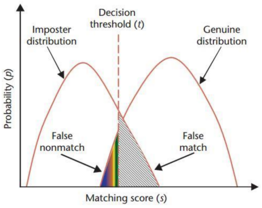

Figure 1: Example of genuine and impostor distributions

If we plot the genuine and impostor score distributions for our dataset, at first only the impostor score is visible : indeed, our dataset is very imbalanced, we have 99.9% of negative samples and only 0.1% positive samples. So, we need normalization to be able to see something. After normalization, we obtain the plot presented in Figure 2.

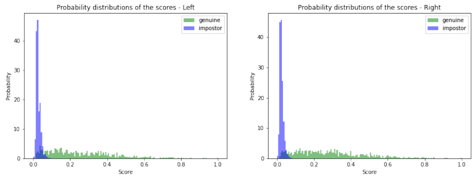

Figure 2: Probability distributions of the scores in our dataset

We can observe in this figure that we have an overlapping between the genuine and the impostor distributions for both of the fingers.

2.2 Question 2 : ROC Curves

In this question, we will compute and interpret the False Acceptance Rate (FAR), the False Rejection Rate (FRR), the ROC curve and the DET curve. To give a brief explanation of these concepts :

- False Acceptance Rate (FAR): Also known as False Match Rate or False Positive Rate, it is the rate of failing to identify an impostor. So, it is the proportion of impostor scores that are greater or equal than the threshold.
- False Rejection Rate (FRR): : Also known as False Non-Match Rate or False Negative Rate, it is the rate of failing to recognize a genuine person. So, it is the proportion of genuine scores that are less than the threshold.
- ROC curve : It is a graphical plot of the True Acceptance Rate (TAR) against the False Acceptance Rate (FAR). We are interested in having a ROC curve with a steep slope, because this will mean that we can have a high TAR with a low FAR.
- DET curve : It is a graphical plot of the False Rejection rate (FRR) against the False Acceptance Rate (FAR) : this graph shows how much the FRR increase when we decrease the FAR and vice-versa, and thus help to select an operating point for the threshold.

<u>FAR - FRR vs threshold</u>

As we can see in the Figure 3, the FPR and FRR rates are pretty similar for the 2 fingers : We see that for a threshold lower than 0.05, we start to have a large increase of the FPR and a large decrease of the FRR.

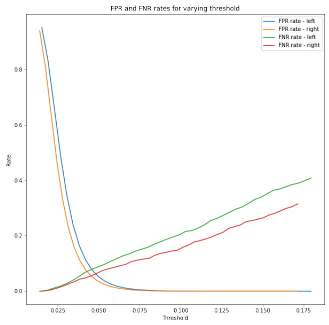

Figure 3: FAR and FRR for varying thresholds

The threshold to select depends on the application :

- If we want to use the fingerprints to unlock the phone for example, having a lower FRR is more suitable : indeed, we want a convenient system, and can "sacrifice" a little bit of security. So, we can select a quite low threshold.
- If we want to use the fingerprints for security applications (forensic, ...), we want the lowest possible FAR ! The threshold selected will be higher compared to the precedent case.

ROC

The ROC curves for the 2 fingers are presented in the Figure 4. A good ROC curve is one that goes up quickly. By looking at the shape, we can observe that the curves in our cases are going up very quickly. The right finger ROC curve has a slightly better shape. This is also observed when computing the ROC AUC in section 2.4. However, we have to keep in mind that the ROC is only based on the FAR (False Acceptance Rates) and the TAR (True Acceptance Rates). In our case, the dataset is very unbalanced, thus ROC might not be really suited.

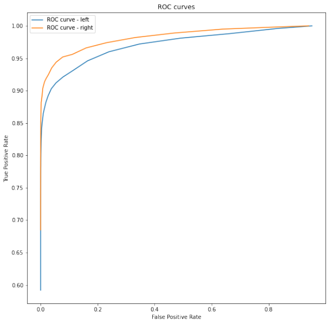

Figure 4: ROC curves for the 2 fingers

DET

The DET curves for the 2 fingers are presented in Figure 5. We can observe that the slopes for 2 fingers are quite the same, but the absolute value of the FRR is higher for the left index : this is most probably due to the measures itself.

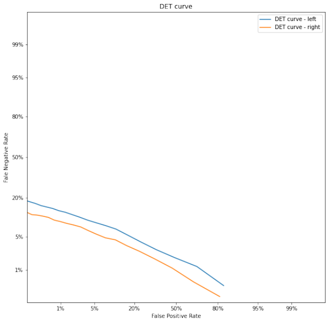

Figure 5: DET curves for the 2 fingers

2.3 Question 3 : Classification Metrics

The F1-score and accuracy curves for the 2 fingers in our dataset are presented in the Figure 6. Let's discuss each of these metrics.

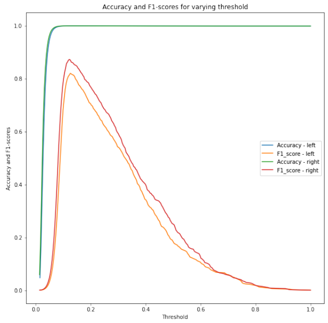

Figure 6: F1-scores and accuracy for the two fingers

F1-score

F1-score is a metric that combine precision and recall. Precision compute the ratio of correct positive predictions over all positive predictions ; Recall compute the ratio of correct positive prediction over all positive samples in the dataset. F1-score is a harmonic mean of precision and recall : both of them needs to be high to have a high F1-Score. This metric is widely used for datasets with unbalanced classes.

At Table 1, we can observe the points with the highest F1-score. One might want to use these points as operating points : indeed, higher the F1-score the better. But note that in our case, F1-Score might not be really suited, because it gives equal weight to precision and recall.

Finger	Max F1-score	Corresponding threshold
Left	0.821	0.126
Right	0.873	0.122

Table 1: Maximum F1-score and correspond thresholds for the 2 fingers

Depending on the application where we will use the fingerprint, the cost of a False Positive (given by Precision) and the cost of not recognizing a True Positive (given by Recall) can be largely different. Using a F1-Score with different weight can in such cases be a better idea.

Accuracy

Accuracy is simply the ratio between correct predictions and all predictions. In biometric systems, using accuracy as a metric is not advised : indeed, it does not take into account the imbalance of the dataset. In our case, we have a lot of impostor samples compared to genuine samples (99.9 % of the samples are impostors !). Therefore, a model that simply predict "impostor" will have an accuracy of nearly 100%. This explains why in Figure 6, Accuracy has a value of 100% after a certain threshold, and stays at 100% as far as the threshold grows : indeed, for high thresholds, we will always predict "impostor" because no sample will have a higher score than the threshold.

Finger	Max accuracy	Corresponding threshold
Left	1.0	0.134
Right	1.0	0.122

Table 2: Maximum accuracies and correspond thresholds for the 2 fingers

2.4 Question 4 : AUC, EER and alternatives

ROC AUC

The ROC AUC under the ROC curve . We want to have the highest ROC AUC possible : indeed, a higher ROC AUC means a ROC curve with a steeper slope, so we can have a high TAR with a low FAR. But note that sometimes, we might want to select a model with a lower AUC if the model is more performing in our desired operating region as we have seen in the lectures.

In our dataset, the results are coherent with the shapes of the ROC curves presented in Figure 4 : The right finger has a higher ROC AUC. But again, the ROC is not a metric suited for unbalanced datasets.

Finger	ROC AUC
Left	0.971
Right	0.983

Table 3: ROC AUC of the 2 fingers

EER

The Equal Error Rate (EER) is the point on the ROC curve where FAR = FRR. A lower EER means a more performing system at the particular point where FAR = FRR. But is it really a good operating point ? Again, it depends on the application : We might want to have a really secure system (so selecting a high threshold to have a low FAR), or a more convenient system (so selecting a low theshold to have a low FRR). In practice, EER is often used as operating point where the cost of a False Acceptance and the cost of a False Rejection is similar. In the case of our dataset, we have :

Finger	FAR = FRR	Threshold
Left	0.0078	0.046
Right	0.055	0.045

Table 4: EER of the 2 fingers

Threshold for which FAR + FRR is minimal

Finger	FAR + FRR	Threshold
Left	0.134	0.057
Right	0.1	0.059

Table 5: Thresholds for which FAR + FRR is minimal

Choosing an operating point where the sum of FAR and FRR might seem appealing. But if we look at the ROC curve in figure 4, we see that when we lower the FAR rate, we also lower the TAR ! Indeed, we will have a more secure system compared to higher FAR, but we will also reject more genuine persons.

Other strategies

All the metrics we have discussed depends on the threshold. We might want to consider a metric independent of the threshold : d-prime value can be considered, which is a metric that measures the separation between the means of the genuine and the impostor probability distributions, by using standard deviations.

2.5 Question 5 : Precision-Recall curves and related summary measures

Precision-Recall

As we have seen in section 2.3, Precision compute the ratio of correct positive predictions over all positive predictions ; Recall compute the ratio of correct positive prediction over all positive samples in the dataset. Precision-Recall curves plot the Precision against the Recall. In our case (see Figure 7), the Precision-Recall curve has a particular shape : we have 100% of Precision for a Recall between 0% and 60% and then the Precision goes down very quickly. We can interpret this behaviour in the following way :

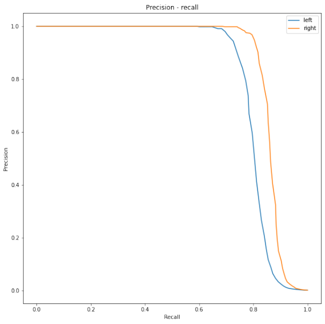

Figure 7: Precision-Recall curves for the two fingers

- For a Recall between 0% and 60%, the model has a very high threshold and predict the majority of samples as "impostor". Therefore, the False Positive rate is very small, and we have a Precision of nearly 100%.
- But if we want to increase the Recall to be able to identify more Positive samples, we will lower the threshold. By doing that, we will also classify some impostor as genuine : the Precision will go down.
- In our case, the right finger has a "better shape" : indeed, we can select an operating point with a higher Recall with the same Precision than the left finger.

Precision-Recall curves should be used when there is class imbalance. However, we must stress that Precision and Recall both ignore the False Negative rate ! Therefore, Precision and Recall can deal with imbalance, yes, but **only** in the cases where we have much more negative samples than positive samples (which is often the case in practical applications).

Precision-Recall : AUC

The AUC values are coherent with the shapes of the PR-curves presented in Figure 7 : The right finger has a higher PR - AUC.

Finger	PR - AUC
Left	0.803
Right	0.863

Table 6: PR - AUC of the 2 fingers

Average Precision Scores

Finger	Average Precision
Left	0.799
Right	0.860

Table 7: Average Precision of the 2 fingers

We can observe that the right finger has a higher average Precision. This can mean that the impostor samples of the right finger have a lower score than those of the left finger.

3 Validation of identification system

Until now, we have considered the verification scenario : we try to identify if the test sample is an "impostor" or a "genuine person". This can in fact be seen as a binary classification problem. Let's now try to validate the performance in the identification scenario : the idea is to identify the identify of the test sample. Based on a database of "N" known person, we try to identify the test sample by matching its features to the features of the database. The identification scenario can be seen as a "N+1" classes classification problem (the N+1$^{\text{th}}$ class being the class not known in the database, when the test sample is not similar enough to any known identity).

3.1 Question 6 : CMC curves

The CMC curve plots the probability that a correct identification is returned with the top "t" ranked matching scores. So, for example, if we have a probability of 85% at a rank of 5, this means that for 85% of the cases, we will find the correct identify in the top 5 predictions. To compute the CMC curve in our dataset, we use two for loops on the similarity matrix : for each identity "i", we first sort the other identities based on the similarity scores. Then, for an increasing rank "j", we look if the identity "i" is in the "j" highest similarity scores. If so, the identification is successful at this rank. If not, it is unsuccessful.

The CMC-curves for our dataset are plotted in Figure 8, with the Rank-1 recognition rate in table 8

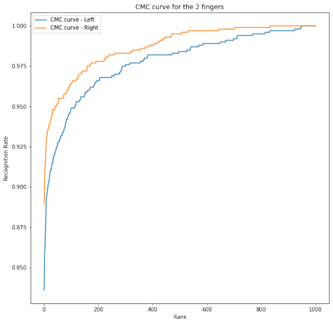

Figure 8: CMC curves for the two fingers

Finger	Rank 1 Recognition Rate
Left	0.836
Right	0.890

Table 8: Rank 1 Recognition Rate

Note that this rate has to be interpret very carefully. Do not forget that we are in an identification scenario : We have 1000 identities in the database, so we will match the test sample with 1000 other identities at each identification !

4 Question 7 : Evaluate different biometric systems

Trough the report and the notebook, we have compared the 2 fingers : it seems like the performance of the right finger is higher compared to the left finger.