Assignment 1: Evaluating performance of biometric systems

everaert Karel

March 2021

1 Introduction

The goal of this assignment is to compare two biometrics systems in both verification and identification mode. The first system makes use of the left index as biometric identifier, while the second system makes use of the right index. In the rest of this work system1 will be used to refer to the system that uses the left index, while system2 refers to the system that makes use of the right index. All the calculations of the values and graphs that can be found in this work can be found in the notebook that comes with it.

2 Validation of verification system

2.1 Evaluation using FMR, FRR, ROC and Precision/Recall curves

2.1.1 Genuine and impostor score distributions

By making use of the genuine and imposter scores that where provided with the assignment it is possible to plot the imposter p(s |I) and genuine distribution p(s |G). Figure 1 and Figure 2 display the genuine versus imposter distributions for respectively system1 and system2. The 2 graphs are very similar too each other. It appears that both systems give imposters very similar scores. This means that when an appropriate threshold is picked e.g. at 0.2 that most imposters will be detected by the systems in verification mode. The genuine distribution in both systems is less centered around a specific score and overlaps a lot with the imposter distribution. Finding a good threshold will be trying to balance the number of genuine users that are denied by the system when the threshold is too high with the number of imposters that are accepted by the system when the threshold is too low.

Figure 1: Genuine vs Imposter distribution for left index

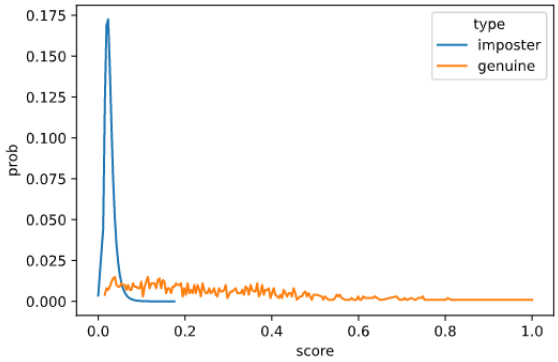
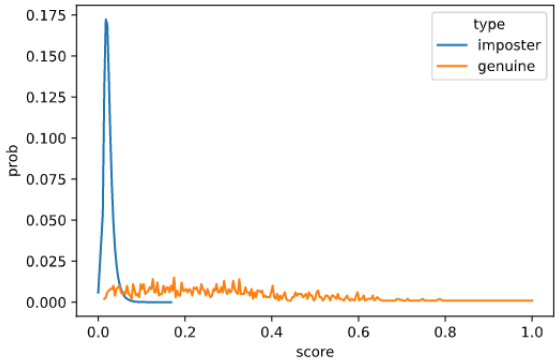


Figure 2: Genuine vs Imposter distribution for right index



2.1.2 FMR, FRR and Receiver Operating Characteristic (ROC) curve

In order to calculate the FPR and TPR, there needs to be defined a value for the threshold. Once a threshold is defined the number of imposters with a score greater then or equal to the threshold are counted in order to calculate the FPR. While the number of genuine users with a score greater then or equal to the threshold are counted to calculate the TPR. In system1 for a threshold value of 0.05 the FPR is 0.037 and the TPR is 0.903. This means that with a threshold set to 0.05 the system would identify 0.037 of imposters as genuine users and 0.903 of genuine users as such. In system2 for a threshold value of 0.05 the FPR 0.027 and the TPR is 0.925. Based on this values we can conclude that for a threshold value of 0.05 system2 works better than system1.

Of course taking 0.05 as threshold value is quite arbitrary therefore, in order to get a better idea of the impact of the threshold, the FAR and the FRR are plotted as a function of the decision threshold. The FRR, or false rejection rate, is the portion of genuine users that where labelled as imposters by the system

and the FAR, or false acceptance rate, is the portion of imposters that were labelled as genuine users by the system. This basically boils down to counting the number of imposters with a score greater than or equal to the threshold to get the FAR and the genuine users with a score smaller than the threshold for the FRR. The threshold value ranges from 0.01 to 1.0 with steps of 0.01 and for each of this threshold values the FAR and FRR are calculated. When all this values are calculated the FAR and FRR are graphed over the different threshold values. The plot for system1 and system2 can respectively be found in Figure 3 and Figure 4. The differences between both plots are minimal. It stands out that for both systems the FAR values drop a lot faster than the value for FRR. This means that the threshold value will not have to be too high to prevent imposters of getting access to the system. This lays in line with what was seen in Figure 1 and Figure 2, due to the fact that all scores for imposters seem to lay very close together it will be easier to stop imposters from being seen as genuine users.

Figure 3: FAR and FRR plotted over threshold value for left index

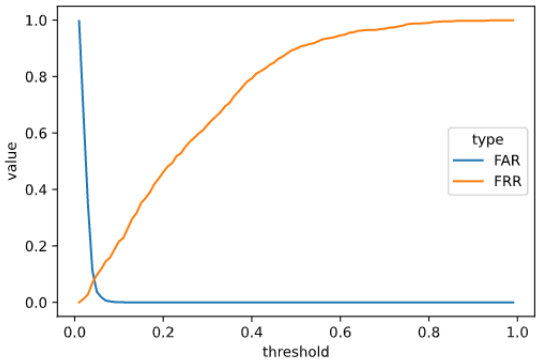
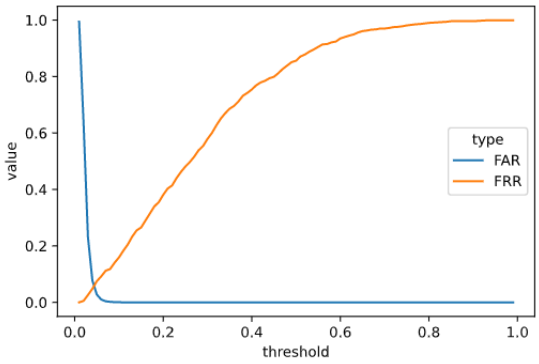
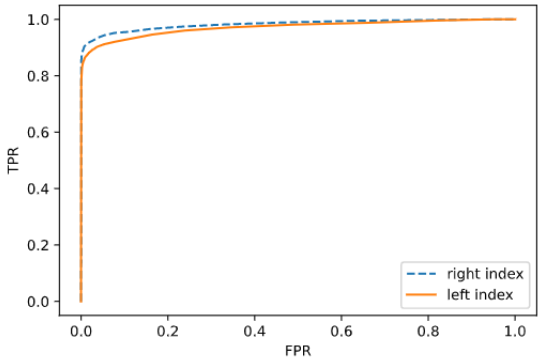


Figure 4: FAR and FRR plotted over threshold value for right index



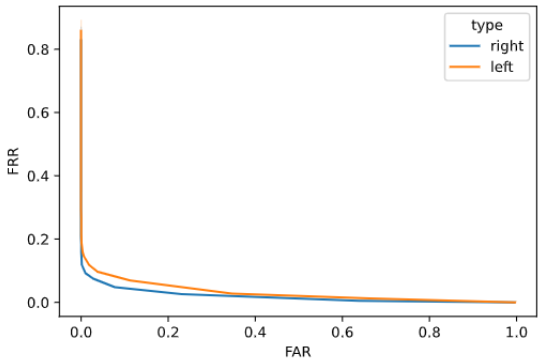
Another interesting way in how to study a biometric system is by plotting the ROC curve. In a ROC curve the TPR is plotted against the FPR for different values of the threshold. The ROC curve for system1 and system2 can both be found in Figure 5. Once again it becomes clear that system2 will outperform system1, since for for a lot of points in the graph system2 has a higher TPR for the same FPR. The most important aspect of the ROC curve is its flatness. The flatness reveals to which extend the FPR can be reduced without much cost to the TPR.

Figure 5: ROC curve of left index and right index based system



When plotting the DET curve, FRR against the FAR for different threshold values, it once again seems that system2 is superior. The DET curves of the two systems can both be found in Figure 6.

Figure 6: DET curve of left index and right index based system



2.1.3 F1 and accuracy as metrics

Firstly F1 and accuracy are plotted in function of the decision thresholds on the similarity score for the two systems. The graphs for system1 and system2 can be

found respectively in Figure 7 and Figure 8. It is clear that the accuracy isn't a good metric to describe the quality of our systems. There are way more imposter scores then that there are genuine scores. Due to this imbalance between the imposter class and genuine class the accuracy is very high if the threshold is set to 1. This makes sense since if the system denies access to all users it will classify all the imposters correctly. However this isn't how the systems should function. The F1 score accounts for this class imbalance and is there a better way to study the quality of the systems. When comparing the F1 graph for system1 with the F1 graph for system2 it once again seems that the peak is higher for system2.

Figure 7: F1 and accuracy plotted over the threshold value for the system that uses the left index

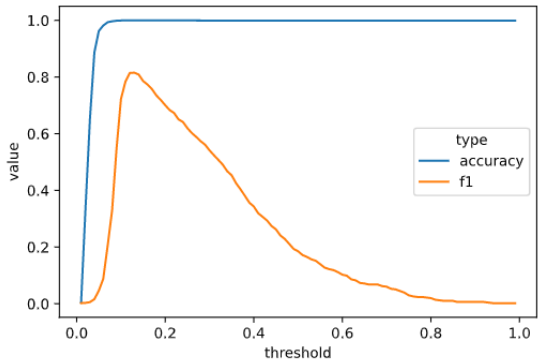
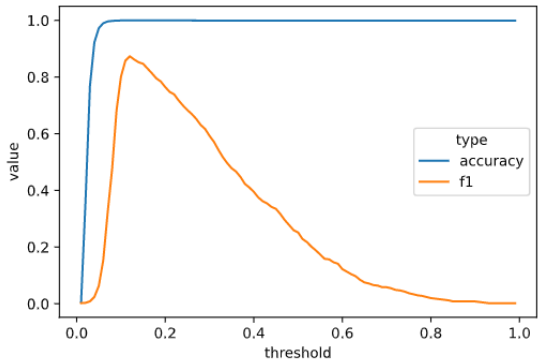


Figure 8: F1 and accuracy plotted over the threshold value for the system that uses the right index



For system1 the threshold for which F1 is maximal is 0.13 and the corresponding value of that maximal F1 score is 0.815. For this system there are multiple values for which the accuracy reaches the maximal accuracy of 0.999

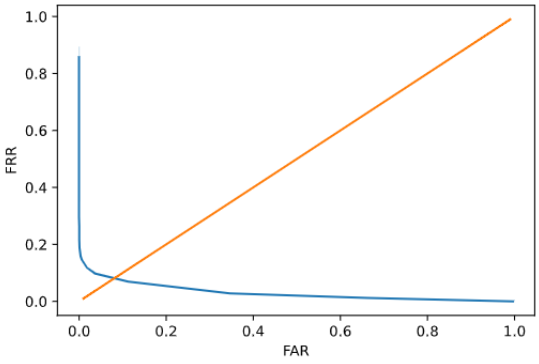
, but the first threshold value for which this happens is 0.13 as well. For system2 the maximal value for the F1 score and accuracy both occur for the same threshold value 0.12. The maximal F1 score for this system is 0.873 and the maximal accuracy is 0.999. As seen before when comparing Figure 7 and Figure 8, the maximal F1 value of system2 is indeed higher than the maximal F1 value for system1.

2.1.4 AUC and EER as summary measures

The AUC, or area under the ROC curve, is typically used to express to overall performance of the system. The higher the AUC value of a system the better this system will perform. However since the AUC value is a summary value one should always inspect the ROC curve as well since this allows for a more complete inspection of the system since it is possible for a lower AUC classifier to outperform a higher AUC classifier in a specific region. The AUC value of system1 turns out to be 0.971, while the AUC value of system2 is 0.983. This results once again are in line with the results discussed previously in this assignment.

For system1 the EER is 0.079, while the EER of the system that uses the right index is 0.055. The EER shows the point where the FAR is equal to the FRR. The lower the EER, the higher the accuracy of the system. The plots with the visualisation of the EER on the FAR-FRR curves can be found in Figure 9 and Figure 10.

Figure 9: EER plotted on FAR-FRR curve for left index system

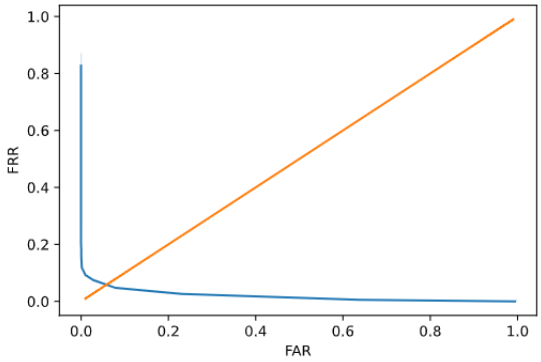


The decision threshold for which the sum of FRR and FAR is minimal turns out to be 0.05 for system1 and 0.06 for system2.

2.1.5 Evaluation using Precision and Recall

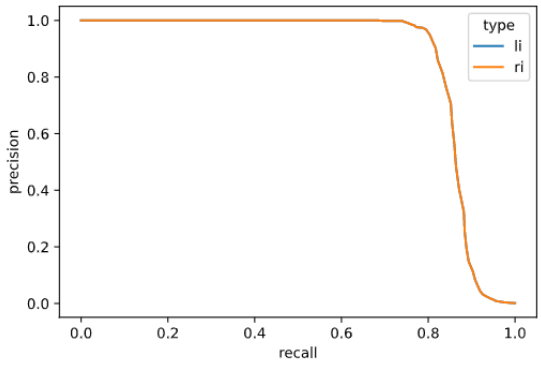
When plotting the precision-recall curve of both systems it seems that these curves completely overlap with each other. The curves can be found in Figure 11.

Figure 10: EER plotted on FAR-FRR curve for right index system



The precision of a system is the proportion of genuine users that gets identified as such. The recall of a system is the probability of correctly identifying a genuine user. The recall only depends on the genuine user class and is therefore class prior independent. The use of precision-recall curves allows to visualise the trade-off between false positives and false negatives. A system with low precision most likely suffers from a high number of false positives while a system with low recall most likely suffers from a system with a high number of false negatives.

Figure 11: Precision-recall curves for both systems



The average precision score for the left index system is 0.799 while the average precision score for the right index system is 0.860.

3 Validation of identification system

When comparing the CMC curves, displayed in Figure 12, to each other it once again is clear that system2 is the better choice for an identification system. At any choice of rank the system will recognize more users than system1. The

Rank-1 Recognition Rate for system1 is 0.832, while for it is 0.893 for system2.

Figure 12: CMC curves for the two systems