Reliability-Based Voting Schemes Using Modality-Independent Features in Multi-classifier Biometric Authentication

Jonas Richiardi and Andrzej Drygajlo

Laboratory of IDIAP, Signal Processing Institute
Swiss Federal Institute of Technology Lausanne
jonas.richiardi@epfl.ch
http://scgwww.epfl.ch

Abstract. We present three new voting schemes for multi-classifier biometric authentication using a reliability model to influence the importance of each base classifier's vote. The reliability model is a meta-classifier computing the probability of a correct decision for the base classifiers. It uses two features which do not depend directly on the underlying physical signal properties, verification score and difference between user-specific and user-independent decision threshold. It is shown on two signature databases and two speaker databases that this reliability classification can systematically reduce the number of errors compared to the base classifier. Fusion experiments on the signature databases show that all three voting methods (rigged majority voting, weighted rigged majority voting, and selective rigged majority voting) perform significantly better than majority voting, and that given sufficient training data, they also perform significantly better than the best classifier in the ensemble.

1 Introduction

A voting combiner operating on the output of classifier ensembles with differing accuracies can be made more effective by supplying it with additional data to influence the importance of each base classifier's vote. A typical scheme is to weight the vote of each classifier proportionally to its accuracy, by training the weights on a development dataset. This paper is concerned with the use of other sources of information for improving voting schemes in biometric authentication.

It has previously been shown that using modality-specific, signal-level quality information can improve classifier combination [1,2]. These quality measures must be tailored to each signal to be used (for instance, image sharpness cannot be used with speech-based biometrics). In this paper, we show that other, modality-independent quality measures can be used in order to estimate the reliability of a classifier's decision, that is, the probability that the base classifier has taken a correct decision.

The estimate of reliability can be used for rejecting the sample (thus decreasing a base classifier's error rate via the reject-error tradeoff), providing a value

M. Haindl, J. Kittler, and F. Roli (Eds.): MCS 2007, LNCS 4472, pp. 377–386, 2007.

to a human layperson (useful in situations such as border control for biometric passports), or improving classifier combination (confidence information has been used to perform classifier selection [3,4,5,6] and classifier fusion [7,8]). In this paper, we propose different ways of using the reliability information in order to improve voting for classifier combination.

First, we introduce modality-independent quality measures in Section 2. We then discuss the process and limits of reliability modelling using the quality measures as features in Section 3. Section 4 proposes three voting schemes using the reliability information, and section 5 shows experimental results of reliability classification on signature and speech, and reliability-based voting for combining multiple signature classifiers. We close the paper by discussing theoretical points and further work in Section 6.

2 Modality-Independent Quality Measures

In order to predict the errors of the base classifiers in the ensemble, it is necessary to find quantities which are indicative of potential mistakes. We call these features *quality measures*. For example, in speaker recognition, a quality measure that is interesting to use is the signal-to-noise ratio (SNR), as a lower SNR tends to increase the probability of error[1] . The two quality measure we use, score and difference between user-specific and user-independent decision threshold, constitute features for the reliability classifier.

Most base classifiers can provide a continuous-valued output (measurement-level) indicating how close or far a particular sample is to a particular class, a quantity generally called score in biometrics. This can be a likelihood or posterior probability value for a probabilistic classifier, an Euclidean distance for a nearest-neighbour classifier, etc. Since the probability of classification error increases as the distance gets closer to the decision boundary between classes, this "soft" classifier output, and its distribution, constitute valuable data for error prediction, and are applicable to any biometric modality whose classifier is capable of producing measurement-level output. Estimation of classifier reliability based only on this soft classifier output is generally called *confidence estimation*.

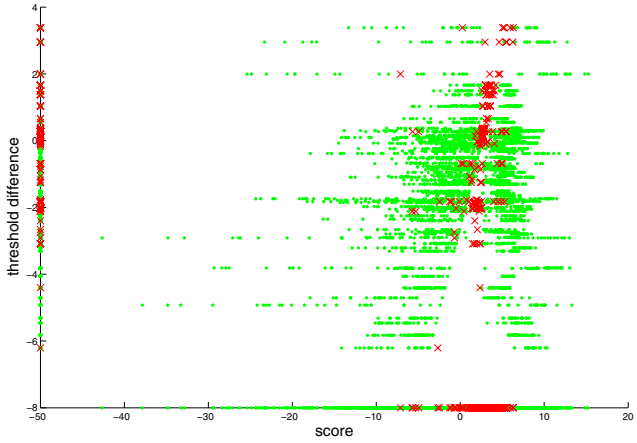
In our experience on speech and signature, however, the boundary defined by the measurement-level output distribution between correct decisions and incorrect decisions of the base classifier is complex, and it is difficult (but not impossible) to train a meta-classifier that performs with fewer errors than the base classifier whose behaviour it models[2]. This is illustrated by the projections on the horizontal axis shown in Figure 1.

Thus, we introduce a second modality-independent quality measure, that is well correlated with errors and the score: the difference between the user-specific

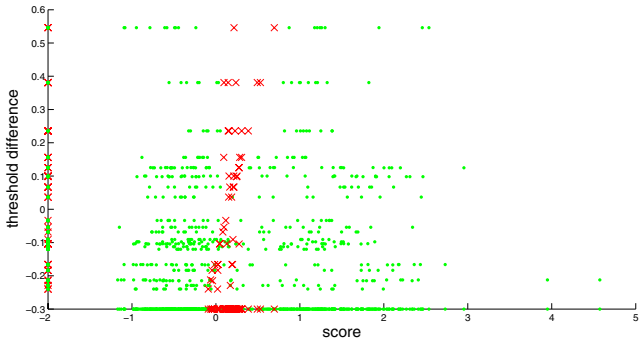
[1] However, it is generally not the case that the relationship between quality measures and base classifier errors can be modelled effectively by linear or low-order polynomial regression.

[2] This is the likely reason for the lack of improvement in fusion mentioned in [8] when using a score-based confidence model.

threshold and the user-independent threshold. In a verification system using user-independent thresholds[3], some users will be more systematically subjected to false rejects, respectively false accepts, than others. As can be seen in Figure 1, this feature makes the reliability classification task easier for both the speech and signature modality.



(a) Quality measures computed from the output of a signature base classifier using local features.



(b) Quality measures computed from the output of a speaker verification base classifier using Mel-frequency cepstral coefficients (MFCC) features.

Fig. 1. Score and threshold difference quality measures for signature verification (MCYT100) and speaker verification (BANCA, G2). Dots indicate reliable (correct) decisions, crosses indicate unreliable (erroneous) decisions of the base classifier. Each quality measure is also projected onto its axis.

[3] For instance because it has recently been deployed and there is not enough data for each user to reliably set a personalised threshold.

3 Reliability Estimation

Once the two features (score and threshold difference) are extracted, we can use nearly any classication algorithm to estimate the reliability of base decisions, with some limitations we discuss in section 3.1. In our case, we use an ensemble of decision trees, either a C4.5 pruned decision tree [9] with bagging or a random forest classifier [10]. In previous work, we have used Bayesian networks to perform reliability estimation [11]. The training data (development set) is separate from the base classifier's training data and the test data, and is generated by running the base classifiers on the development samples.

3.1 Limits of Reliability Modelling

Since we use measurement-level output of the base classifier as one of the features for modelling reliability of decisions, the reliability model is dependent on the accuracy of the base classifier. By definition a well-performing base classifiers has a lower density of soft outputs (which correspond to reliable or unreliable decisions) near the decision boundary than a base classifier with a higher error rate.

However, we can guarantee that the reliability classifier will perform better than the base classifier under certain conditions, which we will phrase in terms of confusion matrices (contingency tables). Let us define $\mathbf{B}$ as the confusion matrix of the base classifier, and $\mathbf{R}$ as the confusion matrix of the reliability classifier. The classes in $\mathbf{B}$, used by the base classifier, are *0—impostor* and *1—client*, while the classes in $\mathbf{R}$, used by the reliability model, are *0—unreliable* and *1—reliable*.

$$\mathbf{B} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \mathbf{R} = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \tag{1}$$

The two confusion matrices are linked by the fact that the reliability model has as class 0 (unreliable) the errors of the base classifier (off-diagonal elements in $\mathbf{B}$), and conversely as class 1 (reliable) the correct decisions of the base classifier (diagonal elements in $\mathbf{B}$):

$$b + c = e + f, \quad a + d = g + h \tag{2}$$

The condition for the reliability model to be able to improve on the output of the base classifier is that the reliability model must make less errors than the base classifier, meaning that the sum of the number of base errors considered reliable and the number of base correct decisions considered unreliable must be less than the sum of the base errors. Equivalently, the accuracy of the reliability model must be higher than that of the base classifier. This formulation can be written as in Equation (3) and simplified by using Equations (2) to obtain Equation (5).

$$\frac{e+h}{(e+f)+(g+h)} > \frac{a+d}{(a+d)+(b+c)} \tag{3}$$

$$\frac{e+h}{(e+f)+(g+h)} > \frac{g+h}{(g+h)+(e+f)} \tag{4}$$

$$e > g \tag{5}$$

Any reliability model whose confusion matrix satisfies the condition expressed in Equation (5) is guaranteed to have less errors than the base classifier it models, and to be useful in reducing base classifier error rates, even if the base classifier performs below chance. If, in addition to reducing base errors, we want the reliability model to perform above chance, we can add the condition

$$e + h > f + g \tag{6}$$

4 Using Reliability in Voting Combiners

While majority voting is an appealing combining scheme, its optimality depends on several assumptions[4], of which we will mention chiefly the fact that it assumes comparable expertise of the ensemble base classifiers. In biometric applications it is often not the case, especially when combining several modalities, with sometimes one or more orders of magnitude of difference between the error rates of the base classifiers. Therefore, we propose three schemes that use classifier-specific reliability models as an input to a controller driving the voting process to improve on majority voting.

4.1 Rigged Majority Voting

The first scheme we propose, rigged majority voting (RMV), uses the base classifier's reliability model to estimate, on an instance-by-instance basis, when its decision is likely to be unreliable. In such cases, the voting controller will rig the vote by inverting it (the role of prior probabilities in the inversion process is discussed in [12]). Denoting the base classifier decision by a binary variable CID (0 for impostors, 1 for clients), the reliability classification by a binary variable DR (0 for unreliable, 1 for reliable), and the rigged decision by RD, the voting controller implements the negative exclusive-or function: $RD = \overline{CID \oplus DR}$ This method works instance-by-instance, by estimating for each case the reliability of the decision.

If the reliability models satisfy Eq. (5), and assuming the correlation between the rigged votes is the same as the correlation between the votes of the base classifiers, this scheme guarantees a better lower and upper bounds on the achievable fused accuracy than simple majority voting on the base classifiers, because the rigged decisions will have higher individual accuracies. This result can be proved using the method in [13].

[4] Such as independence of ensemble members.

However, in the case of base classifiers with very different error rates (say, an order of magnitude), this scheme does not guarantee that we can outperform the best base classifier. We therefore introduce a variation on the voting controller by weighting the contributions of individual classifiers.

4.2 Weighted Rigged Majority Voting

The second scheme we introduce, weighted rigged majority voting (WRMV) is also based on rigged votes, which is an instance-specific method, but the rigged votes are subsequently weighted by a factor proportional to the accuracy of that classifier's reliability model. Thus, we also take into account the overall performance of the base classifier on a development set.

Even though the classifiers violate the independence assumption, and the weights may therefore be suboptimal [14, p.124], we set the classifier-specific weights w_n to

$$\sum_{n=1}^{N} w_n = 1, \;\; w_n \propto \frac{acc_n}{1 - acc_n}, \tag{7}$$

where the accuracy of each reliability model acc_n is computed according to the confusion matrix $\mathbf{R}$ in Eq. (1).

The difference with standard practice for weighted majority voting is that the accuracy used in weighting is not that of the base classifier, but is replaced by the accuracy of the reliability model, which is higher. Thus, the weights are dependent on the effectiveness of the reliability model. However, since the accuracies of the reliability models may follow the same ordering as the accuracies of the base models, the results may not always differ significantly.

The majority threshold is changed from $\tau \geq \lfloor N/2 \rfloor + 1$ for unweighted majority voting to $\tau > \sum_{N_{worst}} w_n$. Thus, the vote of the worst N classifiers N_{worst} in the ensemble is insufficient to win the vote, and if reliabilities are unbalanced the opinion of the most reliable classifiers will count much more. N_{worst} can be chosen as $\lfloor N/2 \rfloor + 1$.

4.3 Selective Rigged Majority Voting

The selective rigged majority voting scheme (SRMV) operates on the same principle as the confidence gating method used in [3], the reliability-based decision table in [1], and the arbitration scheme of [4]: the classifier with the highest confidence gets to label the sample. The difference in our case is that we are operating on decisions that have been rigged by the voting controller before the selection.

Under some conditions (e.g. three classifiers, one of which clearly dominates for most patterns), selective voting can give results very close to weighted rigged majority voting. This is because the weights assigned to the members of the ensemble are proportional to the error rate of their associated reliability classifier.

5 Experiments

In these experiments, we first test the accuracy of the reliability model of each classifier for predicting errors (Section 5.2). Then, in Section 5.3 we apply reliability models to voting on a signature verification task.

5.1 Databases and Base Classifiers

For the signature modality, we use the 100-users MCYT-100 database [15] and the 40-users training set of the SVC2004 database [16]. For the speech modality we use the 52-users, English part of the BANCA database [17] and the 295-users XM2VTS database [18].

The base classifiers for signature are a Gaussian mixture model (GMM) using 15 local features [19] (abbreviated *LGMM*), a GMM using 12 global features (abbreviated *GGMM*), and a multi-layer perceptron (MLP) using the same 12 global features (abbreviated *GMLP*). Both the GMMs and the MLP are learned from 5 signatures, and the MLP is learned using discriminative training.

The base classifier for speaker verification is a GMM based on the Alize toolkit [20] (abbreviated *AGMM*), trained following each speech database's specific protocol (P for BANCA, configuration I for XM2VTS).

5.2 Reliability Prediction with Modality-Independent Quality Measures

The experiments are performed using 10-fold cross-validation and data from all users. Essentially, we want to verify whether we can learn a reliability model that will make less mistakes than the underlying base classifiers. If it is the case, then the reliability model can be used to enhance the performance of the base classifier.

Several types of classifiers were tested for reliability modelling, and the two most promising ones were: bagging of C4.5 trees (abbreviated *BC45*), and random forest classifiers (abbreviated *RF*). For space reasons we will report here only the best performing of the two. The results are reported in Table 1.

5.3 Voting Schemes with Reliability

We compare two baseline combiners, majority voting (abbreviated MV) and weighted majority voting (WMV), to three reliability-based voting combiners: rigged majority voting (RMV), weighted rigged majority voting (WRMV), and selective rigged majority voting (SRMV). The base classifiers are those presented above, with the decision thresholds computed a posteriori.

Table 2 presents the results of the tests on the SVC 2004 signature database. In addition, we performed the McNemar hypothesis test to assess whether the combiners presented are significantly different ($p = 0.05$). Despite the encouraging results, the small size of the dataset (40 users, 1400 cases available for fusion tests) means that the only significant difference (in the majority of the

Table 1. 10-fold cross-validation results of reliability prediction. *DB* indicates the database: S for SVC2004, M for MCYT100, B(G1/2) for BANCA G1 or G2, X(E/T) for XM2VTS Eval or Test set. *Classifier* refers to the type of base classifier used. *Rel Classifier* refers to the type of the reliability classifier used. Err is the error rate (in percent) of the base classifier. Err_r is the error rate (in percent) of the associated reliability model. *Decrease* shows the relative reduction in error rate that can be obtained by using the reliability model along with the base classifier. For BANCA G1, an AdaBoosted ensemble of C4.5 trees brings about a 21.4% relative improvement in the error rate.

DB	Classifier	Rel Classifier	Err [%]	Err_r [%]	Decrease [%]
S	LGMM	RF	8.5	4.0	53.0
S	GGMM	BC45	22.0	16.6	24.0
S	GMLP	BC45	23.8	19.8	17.0
M	LGMM	BC45	3.3	1.8	46.7
M	GGMM	BC45	19.0	12.4	34.7
M	GMLP	BC45	22.6	16.8	26.0
X(E)	AGMM	BC45	1.0	0.8	23.5
X(T)	AGMM	BC45	0.3	0.2	33.0
B(G1)	AGMM	RF=BC45	7.7	7.7	0
B(G2)	AGMM	RF	8.4	4.8	44.0

Table 2. 10-fold cross-validation results of reliability-based decision fusion on the SVC2004 signature database (denoted 'S') and the MCYT100 signature database (denoted 'M'). Baseline best is the best classifier in the ensemble. The standard deviation over the 10 folds is given along with the error rates. FAR is the false accept rate (impostor accepted as a client), FRR the false reject rate (client rejected as an impostor), and $HTER$ is the half total error-rate, $HTER = \frac{FAR+FRR}{2}$.

DB	Scheme	FAR [%]	FRR [%]	$HTER$ [%]
S	Baseline best	8.6 ± 3.6	8.5 ± 3.0	8.6 ± 2.1
S	MV	10.3 ± 3.0	12.9 ± 3.9	11.6 ± 2.2
S	WMV	6.1 ± 3.9	15.9 ± 4.6	11.1 ± 2.0
S	RMV	4.9 ± 2.2	11.2 ± 4.9	8.0 ± 2.2
S	WRMV	2.2 ± 3.0	9.2 ± 5.2	5.7 ± 2.5
S	SRMV	3.3 ± 2.5	6.1 ± 3.4	4.7 ± 1.6
M	Baseline best	3.4 ± 1.1	3.3 ± 0.9	3.3 ± 0.8
M	MV	7.8 ± 1.2	9.0 ± 2.8	8.4 ± 1.2
M	WMV	3.4 ± 1.1	3.3 ± 0.9	3.3 ± 0.8
M	RMV	3.7 ± 1.0	5.0 ± 1.6	4.3 ± 0.8
M	WRMV	1.3 ± 0.8	2.3 ± 0.9	1.8 ± 0.5
M	SRMV	1.5 ± 0.8	2.4 ± 0.1	2.0 ± 0.4

cross-validation runs) is between the MV and SRMV combining schemes. Additionally, WMV and WRMV as well as WMV and SV are significantly different in 50% of the cross-validation folds. Note that using MV or WMV on this ensemble would actually degrade the performance.

Thus, we ran the same experiment on MCYT-100, a larger database comprising 4500 cases. The results are shown in Table 2. On this dataset, all three reliability-based schemes significantly outperform MV and WMV, and WRMV and SRMV both significantly outperform the best base classifier.This underlines the importance of properly assigning weights in imbalanced ensembles. As can be seen from the results for WMV, however, this is not always sufficient, and a finer modelling of the underlying classifier's behaviour can bring enhanced performance.

6 Conclusion

We have presented a new model for classifier reliability, based on features that can be applied independently of the underlying modality. We have used the new reliability model in three new decision-level fusion methods that take into account the overall reliability of individual classifiers on a development set, the instance-by-instance reliability of each classifier's decision, or both.

The rigged voting scheme improves over baseline methods by lowering the bias of the base classifiers. However, the current approach makes no guarantee about the remaining correlation between the rigged votes of the base classifiers, an important factor in voting-based schemes. It is likely that the results would be better with less correlation between base classifiers, as would be the case for majority voting in multimodal verification.

Also, to more clearly show the difference between the WRMV and the SRMV method, it would be interesting to perform experiments with more than 3 classifiers, and with more evenly matched classifiers.

Acknowledgements

We thank J. Ortega-Garcia and J. Fierrez-Aguilar for the provision of the MCYT-100 signature sub-corpus. The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in [18].

References

1. Kryszczuk, K., Richiardi, J., Prodanov, P., Drygajlo, A.: Error handling in multimodal biometric systems using reliability measures. In: Proc. 12th European Conference on Signal Processing (EUSIPCO), Antalya, Turkey (2005)
2. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Bigun, J.: Discriminative multimodal biometric authentication based on quality measures. Pattern Recognition **38** (2005) 777–779
3. Sadeghi, M.T., Kittler, J.: Confidence based gating of multiple face authentication experts. In: Proc. Joint IAPR Int. Workshops, Structural, Syntactic, and Statistical Pattern Recognition 2006. Volume 4109/2006. (2006) 667–676

4. Ortega, J., Koppel, M., Argamon, S.: Arbitrating among competing classifiers using learned referees. Knowledge and Information Systems **V3** (2001) 470–490
5. Alpaydin, E., Kaynak, C.: Cascading classifiers. Kybernetika **34** (1998) 369–374
6. Koppel, M., Engelson, S.P.: Integrating multiple classifiers by finding their areas of expertise. In: Working Notes of the Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms. (1996) held in conjunction with the 13th Nat. Conf. on Artificial Intelligence (AAAI-96).
7. Dutra, T., Canuto, A.M.P., de Souto, M.C.P.: Using weighted combination-based methods in ensembles with different levels of diversity. In: Proc. 13th Int. Conf. on Neural Information Processing (ICONIP 2006), Hong Kong, China (2006) 708–717
8. Bengio, S., Marcel, C., Marcel, S., Mariéthoz, J.: Confidence measures for multimodal identity verification. Information Fusion **3** (2002) 267–276
9. Quinlan, J.: Induction of decision trees. Machine Learning **V1** (1986) 81–106
10. Breiman, L.: Random forests. Machine Learning **45** (2001) 5–32
11. Richiardi, J., Prodanov, P., Drygajlo, A.: Speaker verification with confidence and reliability measures. In: Proc. 2006 IEEE International Conference on Speech, Acoustics and Signal Processing, Toulouse, France (2006)
12. Kryszczuk, K., Drygajlo, A.: Reliability measures and error prediction in biometric identity verification. Journal of Signal Processing (2006) (submitted).
13. Matan, O.: On voting ensembles of classifiers (extended abstract). In: Working Notes of the Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, Portland, USA (1996) held in conjunction with the 13th Nat. Conf. on Artificial Intelligence (AAAI-96).
14. Kuncheva, L.I.: Combining Pattern Classifiers. Wiley and sons (2004)
15. J. Ortega-Garcia et al.: MCYT baseline corpus: a bimodal biometric database. IEE Proc. Vision, Image and Signal Processing **150** (2003) 395–401
16. Yeung, D.Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., Rigoll, G.: SVC2004: First international signature verification competition. In: Proceedings 2004 Biometric Authentication: First International Conference, (ICBA 2004), Hong Kong, China (2004) 16–22
17. Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.P.: The BANCA database and evaluation protocol. In Kittler, J., Nixon, M., eds.: Proceedings of 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA). Volume LNCS 2688. (2003) 625–638
18. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended M2VTS database. In: Proceedings of 2nd International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA). (1999) 72–77
19. Richiardi, J., Drygajlo, A.: Gaussian mixture models for on-line signature verification. In: Proc. ACM SIGMM Multimedia, Workshop on Biometrics methods and applications (WBMA), Berkeley, USA (2003) 115–122
20. Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, USA (2005) 737–740