Fixed and Trained Combiners for Fusion of Imbalanced Pattern Classifiers

Fabio Roli, Giorgio Fumera
Dept. of Electrical and Electronic Engineering
University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{roli, fumera}@diee.unica.it

Josef Kittler
Centre for Vision, Speech and Signal Processing
School of Electronics, Computing and Mathematics
University of Surrey
Guildford GU2 7XH, U.K.
J.Kittler@eim.surrey.ac.uk

Abstract - *In the past decade, several rules for fusion of pattern classifiers' outputs have been proposed. Although imbalanced classifiers, that is, classifiers exhibiting very different accuracy, are used in many practical applications (e.g., multimodal biometrics for personal identity verification), the conditions of classifiers' imbalance under which a given rule can significantly outperform another one are not completely clear. In this paper, we experimentally compare various fixed and trained rules for fusion of imbalanced classifiers. The experiments are guided by the results of a previous theoretical analysis of the authors. Linear, order statistics-based, and trained combiners are compared by experiments on remote-sensing image data and on the X2M2VTS multimodal biometrics data base.*

Keywords: Pattern Classification, Classifier Fusion, Multiple Classifier Systems, Imbalanced Classifiers.

1 Introduction

One of the problems that are to be faced in designing a multiple classifier system is the choice of a suitable fusion rule for the problem at hand [4]. So far few theoretical works investigated the conditions under which specific fusion rules can work well, and a unifying framework for comparing rules of different complexity is clearly beyond the state of the art [4]. In particular, in many applications, one has to deal with classifiers exhibiting very different accuracy, or different pair-wise correlation ("unbalanced" classifiers [8]). As an example, multi-modal biometrics systems for personal identity verification are made up of classifiers processing different information sources (e.g., speech and face data), whose performance is usually very imbalanced. Since different biometrics modalities measure complementary information, their fusion can lead to a more robust and better performing system [1,2,3]. Although imbalanced classifiers are common in many other applications, the conditions of classifiers' imbalance under which a given rule can significantly outperform another one are not completely clear [8]. Therefore, it is of great interest to investigate the performance of different rules for fusion of imbalanced classifiers.

In the past decade, several rules for fusion of classifiers outputs have been proposed in the literature [4]. For the purposes of our discussion, the different combining rules can be subdivided into two main categories: fixed and trained rules. Fixed rules, like the majority voting and the simple averaging, do not require any training. Trained rules, like the weighted averaging and the Behaviour Knowledge Space rule [5,6], require a learning phase. On the basis of experimental and theoretical results, researchers agree that fixed rules usually perform well for classifiers exhibiting similar accuracy, and zero or similar negative correlation among their outputs (balanced classifiers [8,10]). Trained rules are instead claimed to outperform fixed rules for imbalanced classifiers. However, the conditions of classifiers' imbalance under which trained rules can significantly outperform the fixed ones are not completely clear. This problem is of great practical interest, since it is known that the performance of trained rules strongly depends on the quality and size of the training set. This means that it is not guaranteed that the theoretical advantage of trained rules can be achieved in practice. For example, the theoretical advantages of asymptotically optimal trained rules (e.g., the Behavior Knowledge Space rule) are cancelled in the case of small data sets.

In this paper, we report an experimental comparison between well-known fixed and trained rules (Section 2). In particular, we focused on the majority voting rule, simple average and order statistics combiners as fixed rules, and weighted average and Behaviour Knowledge Space as trained rules. On the basis of the theoretical results reported in [7,8], our experiments were aimed to investigate the behaviour of the above fusion rules for imbalanced classifiers. To this end, we considered two pattern recognition applications: classification of remote-sensing images, and personal identity verification with multimodal biometrics data. Results are reported in Section 3. Conclusions are drawn in Section 4.

2 Fixed and trained combiners

In this Section, we describe the fusion rules which have been used for our experimental comparison. In particular, we focus on behaviour of each rule when imbalanced classifiers are used.

In Sect. 2.1, we describe two rules based on linearly combining the outputs of individual classifiers, namely, simple and weighted average. In Sect. 2.2, we describe fixed rules based on order statistics operators. In Sect. 2.3, a trained rule, the Behaviour Knowledge Space, is briefly described. For our experimental comparison, we also considered the majority voting rule. Despite its simplicity, this rule proved to be effective in several applications.

In the following, we consider classifiers whose outputs are approximations of the class posterior probabilities. The a posteriori probability of the *i*-th class provided by the *k*-th classifier for a given input pattern **x** is denoted as $\hat{p}_i^k(\mathbf{x})$.

2.1 Linear combiners

One of the simplest rules for fusing classifiers with continuous outputs is the linear combination. For an ensemble of *N* classifiers, the a posteriori probabilities computed by the linear combiner can be denoted as:

$$\hat{p}_i^{ave}(\mathbf{x}) = \sum_{k=1}^{N} w_k \hat{p}_i^k(\mathbf{x}) \ ,$$

where the w_k's are the coefficients of the linear combination. If all the coefficients are equal (i.e., w_k = 1/*N*), we obtain a fixed combining rule (simple averaging), otherwise we have a trained rule (weighted averaging). Simple averaging is widely used for its simplicity and effectiveness, demonstrated in several experimental studies. However, it can suffer from individual classifiers whose performance is significantly different. Weighted averaging can handle imbalanced classifiers more effectively, but it requires a training phase (weights estimation).

A theoretical framework for the analysis of the simple average combining rule was developed by Tumer and Ghosh [9,10]. Roli and Fumera extended this framework to weighted averaging, in order to provide an analytical comparison between these two rules [7,8]. In the following, we summarise the main results described in [7,8].

For a one-dimensional feature vector *x*, the outputs of an individual classifier can be denoted as:

$$\hat{p}_i(x) = p_i(x) + \varepsilon_i(x) \ ,$$

where $p_i(x)$ is the "true" posterior probability of the *i*-th class, and $\varepsilon_i(x)$ is the estimation error. The analysis of classifier performance can be focused around the class boundaries, under the hypothesis that the boundaries provided from the approximated a posteriori probabilities are close to the optimal Bayesian boundaries [9,10]. Assuming that the estimation errors $\varepsilon_i(x)$ on different classes are i.i.d. variables with zero mean and variance σ_ε^2, Tumer and Ghosh showed that the expected value of the added error (i.e., the error above the Bayes one), E_{add}, can be expressed as:

$$E_{add} = \frac{\sigma_\varepsilon^2}{s} \ ,$$

where *s* is a constant term depending on the values of the probability density functions in the optimal decision boundary. Consider now the weighted average of the outputs of an ensemble of *N* classifiers, with normalised weights w_k:

$$\sum_{k=1}^{N} w_k = 1, \quad w_k \geq 0 \quad k = 1, \ldots, N \ . \qquad (1)$$

The outputs of the weighted averaging combiner can be expressed as [7,8]:

$$\hat{p}_i^{ave}(x) = \sum_{k=1}^{N} w_k \hat{p}_i^k(x) = p_i(x) + \sum_{k=1}^{N} w_k \varepsilon_i^k(x) = \\ = p_i(x) + \bar{\varepsilon}_i(x) \, ,$$

where $\bar{\varepsilon}_i(x)$ denotes the estimation error of the combiner. We assume that, for any individual classifier, the estimation errors $\varepsilon_i^k(x)$ on different classes are i.i.d. variables with zero mean and variance $\sigma_{\varepsilon^k}^2$. We also assume that the errors $\varepsilon_i^m(x)$ and $\varepsilon_i^n(x)$ of different classifiers are correlated on the same class, with correlation coefficient ρ^{mn}, while they are uncorrelated on different classes [7,8,10]. Under these assumptions, we showed that the expected value, E_{add}^{ave}, of the added error of the weighted averaging combiner can be expressed as [7,8]:

$$E_{add}^{ave} = \sum_{k=1}^{N} E_{add}^k w_k^2 \sum_{m=1}^{N} \sum_{n \neq m} 2\rho^{mn} \sqrt{E_{add}^m E_{add}^n} \, w_m w_n \, . \qquad (2)$$

Let us now first analyse the case of uncorrelated estimation errors (i.e., ρ^{mn} = 0 for any $m \neq n$). In this case, Eq. (2) reduces to:

$$E_{add}^{ave} = \sum_{k=1}^{N} E_{add}^k w_k^2 \ .$$

Taking into account Eq. (1), it turns out that the weights that minimise E_{add}^{ave} are:

$$w_k = \left(\sum_{m=1}^{N} \frac{1}{E_{add}^m} \right)^{-1} \frac{1}{E_{add}^k} \ .$$

Such optimal weights are inversely proportional to the expected added errors of the individual classifiers. For equal values of the expected added error, the optimal weights are w_k=1/*N*, that is, simple averaging is the optimal combining rule in the case of classifiers with equal performances ("balanced" classifiers). In such case, it can be shown that $E_{add}^{ave} = E_{add}^k / N$, that is, simple

averaging reduces the expected added error of individual classifiers by a factor N [9,10].

Consider now the case of correlated estimation errors (Eq. (2)). In this case, it is not easy to derive a general analytical expression for the optimal weights. However, it turns out from Eq. (2) that the optimal weights are $w_k=1/N$ if all the classifiers exhibit both equal performances and equal correlation coefficients. Otherwise, different weights are needed to minimise the expected added error E_{add}^{ave} of the combiner. It is worth noting that simple averaging is not the optimal rule if individual classifiers have the same accuracy but different pair-wise correlation. Weights are necessary also for compensating differences in correlation [7,8].

Using the above theoretical model, we quantitatively evaluated the theoretical performance improvement achievable by weighted averaging over simple averaging for imbalanced classifiers [7,8]. The main result of this analysis was that weighted averaging can significantly outperform simple averaging only for ensembles exhibiting high values of the range of classifiers error rates (i.e., the difference between the error rate of the worst and the best individual classifier). However, our analysis also pointed out that the performance improvement is not related only to the range of classifier error rates, but also to the "scattering" of the error rates. Moreover, the differences in correlation also play an important role in determining the performance improvement achievable by weighted averaging.

2.2 Order Statisics combiners

Several fixed combining rules can be defined by using order statistics (OS) operators. Consider the outputs of the N individual classifiers, for any class i, ordered as follows:

$$\hat{p}_i^{1:N}(\mathbf{x}) \le \hat{p}_i^{2:N}(\mathbf{x}) \le \ldots \le \hat{p}_i^{N:N}(\mathbf{x}) \ .$$

The well-known *max*, *med* and *min* combiners are defined as:

$$\hat{p}_i^{\max}(\mathbf{x}) = \hat{p}_i^{N:N}(\mathbf{x}), \quad \hat{p}_i^{\min}(\mathbf{x}) = \hat{p}_i^{1:N}(\mathbf{x}),$$

$$\hat{p}_i^{\text{med}}(\mathbf{x}) = \begin{cases} \dfrac{\hat{p}_i^{\frac{N}{2}:N}(\mathbf{x}) + \hat{p}_i^{\frac{N+1}{2}:N}(\mathbf{x})}{2} & \text{if } N \text{ is even}, \\ \hat{p}_i^{\frac{N+1}{2}:N}(\mathbf{x}) & \text{if } N \text{ is odd}. \end{cases}$$

Even if combining rules based on OS operators are fixed rules, Tumer and Ghosh pointed out that they provide more flexibility than simple averaging [10]. Moreover, Tumer and Ghosh argued that OS combiners could be an effective alternative to weighted averaging for imbalanced classifiers, especially for cases in which it can be difficult to obtain good estimates of the optimal weights.

2.3 Behaviour Knowledge Space combiner

The Behaviour Knowledge Space is a trained fusion rule. The final decision is based on the amount of support received for each class jointly from all the individual classifiers [5,6]. Let us denote the decision of the k-th classifier for the input pattern $\mathbf{x}$ as $\delta_k(\mathbf{x})$. For a c-class problem, $\delta_k(\mathbf{x})$ can take on values 1, …, c. The combination of the decision outputs of the N classifiers, $\delta_k(\mathbf{x})$, k = 1, …, N, defines a point in a c-dimensional discrete space, which is called Behaviour Knowledge Space (BKS). Each point of the BKS can be considered as indexing an entry of a look-up table. Therefore, the whole BKS can be regarded as a look-up table. For a given entry $(\delta_1(\mathbf{x}), \ldots, \delta_N(\mathbf{x}))$, the BKS rule assigns the input pattern $\mathbf{x}$ to the class exhibiting the highest number of patterns for that entry. In other words, for each entry (i.e., for each pattern of classifiers' decisions), the class with the greatest number of votes is chosen by the BKS rule. The values of the look-up table entries are computed by a validation set.

3 Experimental results

In this section, we report experiments aimed at comparing the fusion rules described in Sect. 2. We focus on the behaviour of these rules for ensembles of classifiers exhibiting different accuracy. In particular, for linear combiners (simple and weighted average), our experiments are guided by the theoretical results presented in [7,8], and summarised in Sect. 2.1.

In Sect. 3.1, we report experiments on a data set of remote-sensing images. We focused on the comparison between simple and weighted average rules. In Sect. 3.2, we present a comparison among all the fusion rules described in Sect. 2. To this end, the XM2VTS data base, containing multi-modal biometrics data, was used.

3.1 Results with the Feltwell data set

The experiments described in this section were carried out on a data set of remote-sensing images related to an agricultural area near the village of Feltwell (U.K.) [11]. This data set consists of 10,944 pixels belonging to five agricultural classes. It was randomly subdivided into a training set of 5,820 pixels, and a test set of 5,124 pixels. Each pixel is characterised by fifteen features, corresponding to the brightness values in the six optical bands, and over the nine radar channels considered.

For our experiments, we used ensembles made up of a k-nearest neighbours classifier (k-NN), with a k value of 15, and two multi-layer perceptron (MLP) neural networks with one hidden layer. In order to obtain imbalanced classifiers, we trained several MLPs with two different architectures, characterised by five and two hidden units. The number of input and output units was equal to the number of features and data classes, respectively. We then selected five ensembles characterized by various degrees of classifiers' imbalance. The percentage error rates of the individual classifiers on the test set are shown in Table 1.

The range of the error rates is shown in the last column, denoted as Δ. All the values were averaged over ten runs corresponding to ten training set / validation set pairs, obtained by a bootstrap procedure from the original training set. The validation set contained the 20% of patterns of the original training set. The training of MLPs was stopped as it reached the minimum error probability on the validation set.

Table 1. Average percentage errors on the test set of the individual classifiers forming the five ensembles. For each ensemble, the parameter Δ indicates the range of the error rate, that is, the difference between the best and the worst classifier.

	k-NN	MLP1	MLP2	Δ
Ensemble 1	10.01	11.68	12.05	2.04
Ensemble 2	10.01	18.20	18.00	8.19
Ensemble 3	10.01	13.27	17.78	7.77
Ensemble 4	10.01	25.97	26.23	16.22
Ensemble 5	10.01	17.78	26.23	16.22

According to the value of the parameter Δ, the above ensembles exhibit different degrees of classifiers' imbalance. Ensemble 1 is the most balanced set, since it exhibits the smallest value of the parameter Δ. All the other ensembles are more imbalanced. In particular, ensembles 4 and 5 exhibit the highest degree of imbalance (Δ = 16.22). However, it should be noted that ensemble 4 and 5 differ for the error rate exhibited by MLP1. The same observation holds for ensembles 2 and 3. For linear combiners, we proved that such difference strongly influences the degree of classifiers imbalance, and, consequently, the performance difference between simple and weighted averaging [7,8].

Since in these experiments we were interested in the ideal performance of weighted averaging, the optimal weights of the linear combination were computed on the test set by "exhaustive" search. The average performance of simple and weighted averaging on the test set are reported in Table 2, together with the values of the optimal weights.

Table 2. Average percentage error rates on the test set of simple average (E^{sa}) and weighted average (E^{wa}). For the weighted average rule, the optimal values of the weights assigned to the individual classifiers are shown.

	combiner error rates			optimal weights		
	E^{sa}	E^{wa}	E^{sa}-E^{wa}	*k*-NN	MLP1	MLP2
Ens. 1	10.00	9.37	0.63	0.576	0.200	0.224
Ens. 2	12.09	9.69	2.40	0.689	0.080	0.231
Ens. 3	10.69	9.63	1.06	0.681	0.231	0.088
Ens. 4	16.81	9.79	7.02	0.838	0.006	0.156
Ens. 5	12.44	9.73	2.71	0.752	0.103	0.143

Table 2 shows that weighted average always outperforms the best individual classifier. Simple average achieves this result only for the balanced Ensemble 1, but with a negligible performance improvement. Moreover, the performance of simple and weighted average is very similar for the balanced Ensemble 1, while weighted average outperforms simple average for the imbalanced ensembles. In particular, the higher is the value of the range Δ, the higher is the performance improvement achieved by weighted averaging. However, it is worth noting that the difference between the performance of simple and weighted averaging, E^{sa} - E^{wa}, does not depend on the range Δ of classifiers error rates only, but also on the error rate exhbited by MLP1. Ensembles 2 and 5 exhibit very different values of Δ , respectively 8.19% and 16.22%, but similar values of E^{sa} - E^{wa}. On the other hand, the performance difference E^{sa} - E^{wa} can widely vary even for equal values of Δ, as happens for Ensembles 4 and 5. In [7,8], we provided a theoretical explanation of these results. It is also worth noting that, for imbalanced ensembles, the minimum of the optimal weights is always very low. This seems to mean that weighted averaging can significantly outperform simple averaging only by discarding one of the worst classifiers.

3.2 Results with the XM2VTS data base

The experiments on the fusion of different biometrics modalities for personal identity verification were carried out using the XM2VTS data base. XM2VTS is a multimodal data base consisting of face images, video sequences, and speech recordings of 295 subjects, taken in four sessions. During each session two head rotation and speaking shots were taken. The subjects were randomly subdivided into 200 clients, 25 evaluation impostors and 70 test impostors [12]. In particular, the first shot of the first three sessions were taken as training data, and the second shots for evaluation. Therefore the evaluation set contains 600 client shots and 40,000 impostor cases, while the test set contains 400 client shots and 112,000 impostor cases.

Our experiments were focused on two biometrics modalities: speaker voice and frontal face image. Among the different classifiers designed for each modality, we considered two speech classifiers (denoted in the following as Classifiers 3 and 4) and six face classifiers (denoted as Classifiers 1, 2, 5, 6, 7 and 8). Further details about these classifiers can be found in [13]. Their decision thresholds were selected on the evaluation set using the Receiver Operating Curve (ROC), so that the false acceptance and false rejection error rates are equal. The client and impostor error rates on the test set, achieved using these thresholds, are reported in Table 3, together with the average error rates.

Table 3 shows that the performance of the individual classifiers is very different. This happens even for classifiers based on the same sensing modality. For example, the average error rates of the two speech Classifiers 3 and 4 differ by a factor 6. Similarly, a

difference by a factor 4 can be observed between the best and worst face classifiers (Classifier 2 and 8).

Table 3. Test set error rates for the eight individual classifiers.

Error rate	Classif. 1	Classif. 2	Classif. 3	Classif. 4
Average	7.185	3.105	4.205	0.740
Client	6.750	2.750	7.000	0.000
Impostor	7.620	3.460	1.410	1.480
Error rate	Classif. 5	Classif. 6	Classif. 7	Classif. 8
Average	7.055	7.510	7.310	12.940
Client	6.000	7.250	6.500	12.250
Impostor	8.110	7.770	8.120	13.630

Using the above eight classifiers, we designed four ensembles with different degrees of classifiers' imbalance. The classifiers of each ensemble and the corresponding test set average error rates are reported in Table 4. According to the values of the parameter Δ, Ensemble 1 is balanced, while the other ensembles are more imbalanced. In particular, Ensembles 2 and 3 are characterized by the same value of Δ, but different error rates of the "mid-range" classifier. Ensemble 4 exhibits the highest value of the parameter Δ.

Table 4. Average error rates of the four ensembles on the test set. For each ensemble, the parameter Δ indicates the range of the error rate, that is, the difference between the best and the worst classifier. Such parameter is used for characterizing the degree of classifiers' imbalance.

	Classifiers	Average Error Rates			Δ
Ens. 1	5,1,7	7.055	7.185	7.310	0.255
Ens. 2	2,7,6	3.105	7.310	7.510	4.405
Ens. 3	2,3,6	3.105	4.205	7.510	4.405
Ens. 4	2,6,8	3.105	7.510	12.940	9.835

The average test set error rates obtained using the five fixed rules are reported in Table 5, while Table 6 shows the results of the two trained rules. We remark that the optimal weights of the weighted averaging rule were computed by exhaustive search on the test set. Analogously, the BKS rule was trained on the test set. Therefore, the related results represent the ideal performance of these rules.

Table 5 shows that simple average outperforms the best individual classifier only for the balanced Ensemble 1. The performances of the majority voting rule are significantly good. In particular, majority voting exhibited an excellent performance for Ensemble 3. The performance of the three combining rules based on order statistics operators is often significantly worse than simple average and majority vote. The only exception is the *med* combiner for Ensemble 3. However, the *med* combiner exhibited an error rate very similar to that of the worst classifier for Ensemble 4. The *max* rule performed worse than all individual classifiers for the balanced Ensemble 1.

Table 5. Average test set error rates for the five fixed rules described in Sect. 2. S.A. denotes the simple average rule.

	S.A.	Majority	OS *min*	OS *med*	OS *max*
Ens. 1	6.014	5.696	6.465	6.725	9.049
Ens. 2	5.739	5.836	7.475	6.465	5.325
Ens. 3	4.420	1.035	7.470	1.587	5.325
Ens. 4	5.509	3.895	7.474	11.980	5.325

Table 6. Average test set error rates for the two trained rules. W.A. denotes the weighted average rule. BKS indicates the Behavior Knowledge Space method.

	W.A.	BKS
Ensemble 1	3.205	4.909
Ensemble 2	2.184	4.246
Ensemble 3	0.351	0.474
Ensemble 4	2.150	3.487

Let us now consider the performance of the trained rules (Table 6). Note that the weighted average rule always outperformed both the best individual classifier and all the other combining rules. In particular, the error rate improvements over the simple average rule are reported in Table 7. We point out that the weighted average rule significantly outperforms simple average even for the balanced Ensemble 1. However, the corresponding improvement is comparable to that achieved on the imbalanced ensembles, in particular for Ensemble 4. Concerning the optimal weights, which are reported in Table 8, it is possible to note that a very low weight is assigned to one of the worst classifiers. This seems to confirm that weighted average can significantly outperform simple average only by discarding one or more of the worst classifiers. Consider now the BKS rule. It outperforms the best individual classifiers only for Ensembles 1 and 3. Moreover, it is worth noting that its performance is quite similar to those of the majority voting rule. The difference between the error rates is greater than 1% only for Ensemble 2. The BKS rule achieves very good performance for Ensemble 3, like weighted average, majority vote and *med*.

To analyze further the above performance, we computed the correlation coefficients among the outputs of the classifiers of each ensemble. They are reported in Table 9. For Ensemble 3, the outputs of each pair of classifiers exhibit very low correlations, except for the outputs of classifiers 2 and 6 corresponding to clients. The correlation of the other ensembles are instead significantly greater, at least for some pairs of classifiers. This can explain the good performance achieved by the most of the rules for Ensemble 3. However, it is interesting to note that the simple average rule failed to

take advantage of low correlated classifiers, as can be seen from Table 5.

Table 7. Difference between the test set error rates achieved by the simple and weighted average rules for the four ensembles.

Ensembles	1	2	3	4
E^{sa}-E^{wa}	2.809	3.555	4.069	3.359

Table 8. Optimal weights of the weighted average rule for the four ensembles.

Ensemble 1	5	1	7
Optimal weights	0.010	0.780	0.210
Ensemble 2	2	7	6
Optimal weights	0.850	0.000	0.150
Ensemble 3	2	3	6
Optimal weights	0.140	0.830	0.030
Ensemble 4	2	6	8
Optimal weights	0.870	0.080	0.050

Table 9. Correlation coefficients among the outputs of the classifiers of the four ensembles on the test set. The correlations have been computed separately for the two classes of clients and impostors. For each ensemble, the correlation coefficients were computed for all the possible couples of classifiers. (*x*,*y*) indicates the pair of classifiers considered.

Ensemble 1	(1,5)	(1,7)	(5,7)
Clients	0.440	0.424	0.975
Impostors	0.091	0.026	0.960
Ensemble 2	(2,6)	(2,7)	(6,7)
Clients	0.421	0.452	0.977
Impostors	-0.050	0.041	0.964
Ensemble 3	(2,3)	(2,6)	(3,6)
Clients	0.036	0.421	0.067
Impostors	0.013	-0.050	0.057
Ensemble 4	(2,6)	(2,8)	(6,8)
Clients	0.421	0.363	0.548
Impostors	-0.050	0.068	0.505

4 Conclusions

In this paper, we reported an experimental comparison between fixed and trained combining rules for fusion of classifiers. The experiments were carried out using a data set of remote-sensing images, and a personal identity verification problem based on a multimodal biometrics data base. In particular, we focused on the behaviour of the considered combining rules for ensembles of imbalanced classifiers.

The results showed that the performance improvements of the two trained rules (weighted average and BKS) over fixed rules are lower than one can think of. In particular, the weighted average rule significantly outperformed simple average only for ensembles of classifiers exhibiting very different error rates. Moreover, such improvement was achieved only by discarding at least one classifier. This result is in agreement with the theoretical results presented in [7,8]. Concerning the BKS rule, its performance was not significantly better than the ones of majority voting, and it was always slightly worse than the one of weighted average.

These results point out that, even for ensembles of classifiers exhibiting very different performances, trained rules could not provide significant advantages over fixed rules. It should be also noted that we considered the ideal performance of trained rules. In practical applications, their performance can be affected by small or low representative training sets, reducing further the achievable improvement over fixed rules. Further work is therefore required to clearly identify the conditions under which trained rules can significantly outperform fixed rules.

References

[1] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, *Fusion of face and speech data for personal identity verification*, IEEE Trans. on Neural Networks, Vol. 10, No. 5, pp. 1065-1074, 1999.

[2] S. Ben-Yacoub, J. Luettin, K. Jonsson, J. Matas, and J. Kittler, *Audio-visual person verification*, in *Computer Vision and Pattern Recognition*, IEEE Computer Society, Los Alamitos, California, 1999, pp. 580-585.

[3] J. Kittler, M. Hatef, R. Duin, and J. Matas, *On combining classifiers*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, pp 226-239, 1998.

[4] J. Kittler, and F. Roli (eds.), Multiple Classifier Systems, Springer-Verlag, LNCS, Vol. 1857 (2000), and Vol. 2096 (2001).

[5] L. Xu, A. Krzyzak, and C.Y. Suen, *Methods of combining multiple classifiers and their applications to handwriting recognition*, IEEE Trans. on Systems, Man, and Cybernetics, Vol. 22, pp. 418-435, 1992.

[6] Y.S. Huang, and C.Y. Suen, *A method of combining multiple experts for the recognition of unconstrained handwritten numerals*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 17, pp. 90-94, 1995.

[7] G. Fumera, and F. Roli, *Performance analysis and comparison of linear combiners for classifier fusion*, IAPR Int. Workshop on Statistical Pattern Recognition

(SPR 2002), Windsor, Canada, August 2002, Springer-Verlag, LNCS, in press.

[8] F. Roli, and G. Fumera, *Analysis of linear and order statistics combiners for fusion of imbalanced classifiers*, 3rd Int. Workshop on Multiple Classifier Systems (MCS 2002), Cagliari, Italy, June 2002, Springer-Verlag, LNCS, in press.

[9] K. Tumer, and J. Ghosh, *Analysis of decision boundaries in linearly combined neural classifiers*, Pattern Recognition, Vol. 29, No. 2, pp. 341-348, February 1996.

[10] K. Tumer, and J. Ghosh, *Linear and order statistics combiners for pattern classification* in *Combining Artificial Neural Nets*, A.J.C. Sharkey (ed.), Springer, 1999, pp. 127-161.

[11] F. Roli, *Multisensor image recognition by neural networks with understandable behaviour*, Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 10, pp. 887-917, 1996.

[12] J. Luettin, and G. Matre, *Evaluation protocol for the extended m2vts database (xm2vtsdb)*, Technical Report IDIAP-COM 98-05, Dalle Mole Institute for Perceptual Artificial Intelligence, http://www.idiap.ch, July 1998.

[13] J. Kittler, M. Ballette, and F. Roli, *Decision level fusion in multimodal personal identity verification systems*, submitted to the Information Fusion journal, special issue on Fusion of Multiple Classifiers.