Performance Analysis and Comparison of Linear Combiners for Classifier Fusion

Giorgio Fumera and Fabio Roli

Dept. of Electrical and Electronic Eng., University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{fumera,roli}@diee.unica.it

Abstract. In this paper, we report a theoretical and experimental comparison between two widely used combination rules for classifier fusion: simple average and weighted average of classifiers outputs. We analyse the conditions which affect the difference between the performance of simple and weighted averaging and discuss the relation between these conditions and the concept of classifiers' "imbalance". Experiments aimed at assessing some of the theoretical results for cases where the theoretical assumptions could not be hold are reported.

1 Introduction

In the past decade, several rules for fusion of classifiers outputs have been proposed [10]. Some theoretical works also investigated the conditions which affect the performance of specific combining rules [1,2,3]. For the purposes of our discussion, the combining rules proposed in the literature can be classified on the basis of their "complexity". Simple rules are based on fixed combining methods, like the majority voting [1] and the simple averaging [2,3]. Complex rules use adaptive or trainable techniques, like the weighted voting [4] and the Behaviour Knowledge Space rule [5]. Researchers agree that simple combining rules work well for ensembles of classifiers exhibiting similar performance ("balanced" classifiers). On the other hand, experimental results showed that complex combining rules can outperform simple ones for ensembles of classifiers exhibiting different performance ("imbalanced" classifiers), supposed that a large and independent validation set is available for training such rules [10]. From the application viewpoint, it would be very useful to evaluate the maximum performance improvement achievable by trained rules over fixed ones for a classifier ensemble exhibiting a certain degree of imbalance. If such improvement is not significant for the application at hand, the use of a trained rule could be not worth, since the quality and the size of the training set can strongly reduce the theoretical improvement. However, no theoretical framework has been developed so far, which allows a clear quantitative comparison between different combining rules.

In this paper, we focus on two widely used combining rules, namely, simple and weighted averaging of classifiers outputs. Weighted averaging is often claimed to

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 424-432, 2002.
© Springer-Verlag Berlin Heidelberg 2002

perform better than simple averaging for unbalanced classifier ensembles. However, to the best of our knowledge, no work clearly analysed the conditions which affect the difference between the performance of simple and weighted averaging. The performance improvement achievable by weighted averaging was not clearly quantified so far [2,3,11]. Moreover, experimental results, for instance, the ones reported in [6], showed a small improvement.

In the following, we report a theoretical and experimental comparison between weighted averaging and simple averaging. For our theoretical comparison, we used an analytical framework developed by Tumer and Ghosh [2,3] for the simple averaging rule, and extended it to the weighted averaging rule (Section 2). In Section 3, we quantify the theoretical performance improvement achievable by weighted averaging over simple averaging. We also discuss the conditions under which such improvement can be achieved, and the connection with the concept of classifier "imbalance". In Section 4, experiments aimed at assessing some of the theoretical results for cases where the theoretical assumptions could not be hold are reported.

2 An Analytical Framework for Linear Combiners

Following the work of Tumer and Ghosh [2,3], the outputs of an individual classifier approximating the a posteriori probabilities can be denoted as:

$$\hat{p}_i(x) = p_i(x) + \varepsilon_i(x) \ , \tag{1}$$

where $p_i(x)$ is the "true" posterior probability of the i-th class, and $\varepsilon_i(x)$ is the estimation error. We consider here a one-dimensional feature vector x. The multi-dimensional case is discussed in [7]. The main hypothesis made in [2,3] is that the decision boundaries obtained from the approximated a posteriori probabilities are close to the Bayesian decision boundaries. This allows focusing the analysis of classifier performance around the decision boundaries. Tumer and Ghosh showed that the expected value of the added error (i.e., the error added to the Bayes one due to estimation erros), denoted as E_{add}, can be expressed as:

$$E_{add} = \frac{1}{2s} E\left\{\left(\varepsilon_i(x_b) - \varepsilon_j(x_b)\right)^2\right\}, \tag{2}$$

where $E\{\}$ denotes the "expected" value, and s is a constant term depending on the values of the probability density functions in the optimal decision boundary. Let us assume that the estimation errors $\varepsilon_i(x)$ on different classes are i.i.d. variables [2,3], with zero mean (note that we are not assuming that the estimated a posteriori probabilities sum up to 1). Denoting their variance with σ_ε^2, we obtain from Eq. 2:

$$E_{add} = \frac{\sigma_\varepsilon^2}{s} \ . \tag{3}$$

Let us now evaluate the expected value E_{add}^{ave} of the added error for the weighted averaging of the outputs of an ensemble of N classifiers. We consider the case of normalised weights w_k:

$$\sum_{k=1}^{N} w_k = 1, \quad w_k \geq 0 \quad k = 1,\ldots,N \ . \tag{4}$$

The outputs of the combiner can be expressed as:

$$\hat{p}_i^{ave}(x) = \sum_{k=1}^{N} w_k \hat{p}_i^k(x) = \sum_{k=1}^{N} w_k \left(p_i(x) + \varepsilon_i^k(x)\right) = p_i(x) + \bar{\varepsilon}_i(x) \ , \tag{5}$$

where

$$\bar{\varepsilon}_i(x) = \sum_{k=1}^{N} w_k \varepsilon_i^k(x) \tag{6}$$

is the estimation error of the combiner. By proceeding as shown above for an individual classifier, one obtains the following expression for E_{add}^{ave}:

$$E_{add}^{ave} = \frac{1}{2s} E\left\{\left(\bar{\varepsilon}_i\left(x_b^{ave}\right) - \bar{\varepsilon}_j\left(x_b^{ave}\right)\right)^2\right\}, \tag{7}$$

where x_b^{ave} denotes the decision boundary estimated by the combiner. We assume again that, for any individual classifier, the estimation errors $\varepsilon_i^k(x)$ on different classes are i.i.d. variables with zero mean, and denote their variances with $\sigma_{\varepsilon^k}^2$. We also assume that the errors $\varepsilon_i^m(x)$ and $\varepsilon_i^n(x)$ of different classifiers on the same class are correlated [2,3], with correlation coefficient ρ_i^{mn}, while they are uncorrelated on different classes. Under these assumptions, we obtain from Eq. 7:

$$E_{add}^{ave} = \frac{1}{s}\sum_{k=1}^{N} \sigma_{\varepsilon^k}^2 w_k^2 + \frac{1}{s}\sum_{m=1}^{N}\sum_{n\neq m} \left(\rho_i^{mn} + \rho_j^{mn}\right) \sigma_{\varepsilon^m} \sigma_{\varepsilon^n} w_m w_n \ . \tag{8}$$

This expression generalises the result obtained in [2,3] for simple averaging to the case of weighted averaging. For the purposes of our discussion, let us assume that the correlation coefficients of the different classes are equal: $\rho_i^{mn} = \rho_j^{mn} = \rho^{mn}$. From Eq. 3 it follows that E_{add}^{ave} can be rewritten as follows:

$$E_{add}^{ave} = \sum_{k=1}^{N} E_{add}^k w_k^2 + \sum_{m=1}^{N}\sum_{n\neq m} 2\rho^{mn} \sqrt{E_{add}^m E_{add}^n}\, w_m w_n \ . \tag{9}$$

Let us now analyse Eq. 9. We first consider the case of uncorrelated estimation errors (i.e., ρ^{mn}=0 for any $m\neq n$). In this case Eq. 9 reduces to:

$$E_{add}^{ave} = \sum_{k=1}^{N} E_{add}^k w_k^2 \ . \tag{10}$$

Taking into account the constraints of Eq. 4, it is easy to see that the optimal weights of the linear combination, that is, the ones which minimise the above E_{add}^{ave}, are:

$$w_k = \left(\sum_{m=1}^{N} \frac{1}{E_{add}^m}\right)^{-1} \frac{1}{E_{add}^k} \ . \tag{11}$$

Eq. 11 shows that the optimal weights are inversely proportional to the expected added error of the corresponding classifiers. Accordingly, for equal values of the

expected added error, the optimal weights are w_k=1/N. This means that simple averaging is the optimal combining rule in the case of classifiers with equal performance ("balanced" classifiers).

Consider now the case of correlated estimation errors (Eq. 9). In this case it is not easy to derive an analytical expression for the optimal weights. However, from Eq. 9 it turns out that the optimal weights are w_k=1/N if all classifiers exhibit both equal average performance and equal correlation coefficients. Otherwise, different weights are needed to minimise the expected added error E_{add}^{ave} of the combiner. This means that even for equal average accuracies, simple averaging is not the optimal rule, if the estimation errors of different classifiers exhibit different correlations.

3 Performance Analysis and Comparison

In this section, we quantitatively evaluate the theoretical improvement achievable by weighted averaging over simple averaging. To this end, we use the theoretical model described in Sect. 2. In the following we denote with ΔE_{add}^{ave} the difference between the expected added error achieved by simple averaging and the one achievable by weighted averaging using the optimal weights given in Eq. 11. Without loss of generality, we consider the N classifiers ordered for decreasing values of their expected added error E_{add}^{k}, so that $E_{add}^{1} \geq E_{add}^{2} \geq \ldots \geq E_{add}^{N}$.

3.1 Combining Uncorrelated Classifiers

Let us first consider the case of uncorrelated estimation errors (i.e., ρ^{mn}=0 for any $m \neq n$). According to Eq. 10, ΔE_{add}^{ave} can be written as:

$$\Delta E_{add}^{ave} = \frac{1}{N^2} \sum_{k=1}^{N} E_{add}^{k} - \left(\sum_{k=1}^{N} \frac{1}{E_{add}^{k}} \right)^{-1} . \tag{12}$$

By a mathematical analysis of Eq. 12 we proved that, for any given value of the difference $E_{add}^{1} - E_{add}^{N}$, the maximum of ΔE_{add}^{ave} is achieved when the N-2 classifiers 2,…,N-1 exhibit the same performance of the worst individual classifier, that is, $E_{add}^{1} = E_{add}^{2} = \ldots = E_{add}^{N-1} > E_{add}^{N}$. For the sake of brevity, we omit this proof. According to our model, this is therefore the condition under which, for a given value of the difference $E_{add}^{1} - E_{add}^{N}$, the advantage of weighted averaging over simple averaging is maximum. Hereafter we will denote this condition as performance "imbalance".

Under the above condition, in Fig. 1 we reported the values of ΔE_{add}^{ave} for values of $E_{add}^{1} - E_{add}^{N}$ ranging from 0 to 25%. Three different values of E_{add}^{N} for the best classifier were considered (1%, 5%, and 10%), and two values of the ensemble size, N=3,5. From Fig. 1, two conclusions can be drawn. First, weighted averaging significantly outperforms simple averaging (say, more than 1%) only if the performance of the individual classifiers are highly imbalanced (that is, for high values of $E_{add}^{1} - E_{add}^{N}$), and if the performance of the best individual classifier is very

high (that is, for low values of E^N_{add}). Moreover, the advantage of weighted averaging decreases for increasing values of N (note that in practice it is unlikely to have a high number of uncorrelated classifiers [8]).

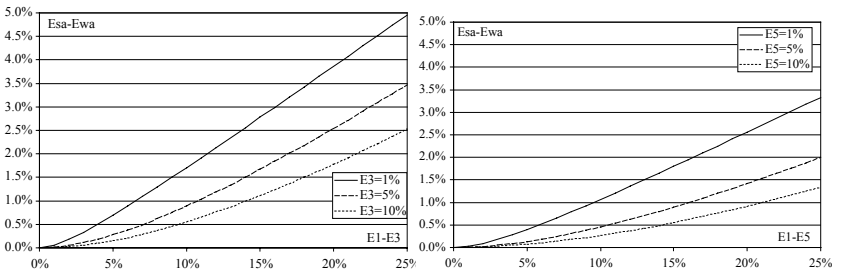

Fig. 1. Values of ΔE^{ave}_{add} (denoted as E^{sa}-E^{wa}) for uncorrelated classifiers, for N=3 (left) and N=5 (right). The values of E^i_{add} are denoted as E_i

Consider now the optimal weights given in Eq. 11. It is easy to see that the highest weight is assigned to the best individual classifier. Moreover, the weights of classifiers 1,…,N-1 are equal, as these classifiers have equal values of the expected added error. Their weight is reported in Fig. 2, plotted against $E^1_{add} - E^N_{add}$, for the same values of E^N_{add} and N as in Fig. 1.

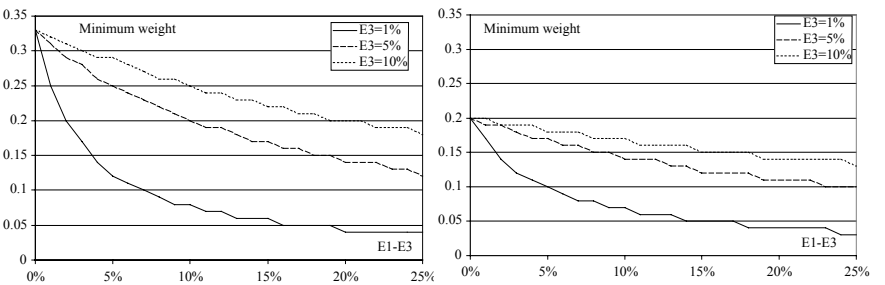

Fig. 2. Values of the minimum of the optimal weights, for N=3 (left) and N=5 (right)

The comparison of Figs. 1 and 2 shows that higher values of ΔE^{ave}_{add} correspond to lower weights for classifiers 1,…,N—1. In particular, if the best individual classifier performs very well (i.e., E^N_{add} is close to 0), a value of ΔE^{ave}_{add} greater than 1% can be achieved only by assigning to the other classifiers a weight lower than 0.1. This means that the performance of weighted averaging gets close to the one of the best individual classifier, as the other classifier are discarded.

To sum up, the theoretical model predicts that weighted averaging can significantly outperform simple averaging only if a classifier with very high performance is combined with few other classifiers exhibiting much worse performance. However, in this case, using only the best classifier could be a better choice than combining.

3.2 Combining Correlated Classifiers

Let us now consider the case of correlated estimation errors (Eq. 9). We evaluated ΔE_{add}^{ave} by first computing numerically the optimal weights from Eq. 9. As in the case of uncorrelated errors, it turned out that, for any given value of $E_{add}^{1} - E_{add}^{N}$, the maximum of ΔE_{add}^{ave} is achieved for $E_{add}^{1} = E_{add}^{2} = \ldots = E_{add}^{N-1}$. Under this condition, in Fig. 3 we report the values of ΔE_{add}^{ave} for N=3, and for values of ρ^{mn} in the range [—0.4, 0.8]. Fig. 3 shows that the advantage of weighted averaging over simple averaging is greater than in the case of uncorrelated errors. However, note that achieving a significant advantage still requires that the performance of the individual classifiers are highly imbalanced. Moreover, it turns out that the weight of one of the worst classifiers is always zero.

Let us now consider the values of the correlation coefficients. For given values of E_{add}^{3} and $E_{add}^{1} - E_{add}^{3}$, it turned out that the maximum of ΔE_{add}^{ave} is achieved when the best individual classifier is little correlated with one of the others (in our case, $\rho^{13} = —0.4$), while the other correlation coefficients must be as high as possible ($\rho^{12} = \rho^{23} = 0.8$). It seems therefore that the correlations must be imbalanced in an analogous way as the performance.

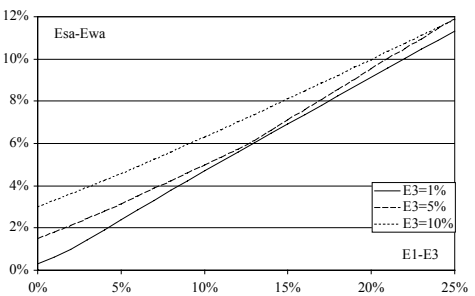

Fig. 3. Values of ΔE_{add}^{ave} for correlated classifiers, for N=3

To better analyse the effects of correlation, we evaluated ΔE_{add}^{ave} for varying values of the correlations coefficients. We considered imbalanced values in the sense defined above, that is, $\rho^{13}<\rho^{12}=\rho^{23}$. In Fig. 4 the values of ΔE_{add}^{ave} are plotted against the value of ρ^{12}—ρ^{13}, for three different values of ρ^{13}. Two different cases are considered for the expected added errors: $E_{add}^{1} = E_{add}^{2} = E_{add}^{3} = 5\%$ and $E_{add}^{1} = E_{add}^{2} = 10\%, \quad E_{add}^{3} = 5\%$.

Fig. 4 shows that imbalanced correlations significantly affect the performance of simple averaging only when the individual classifiers have imbalanced performance. Moreover, it turned out that the weight assigned to one of the worst individual classifiers drops to zero as the imbalancing in performance or in correlation increases. This means that discarding such classifier would not affect the performance of weighted averaging, while simple averaging would perform significantly better. We found that the classifier whose weight is minimum is the one highly correlated with the best individual classifier. In Fig. 5 the value of the lowest of the optimal weights is reported, with reference to the cases shown in Fig. 4.

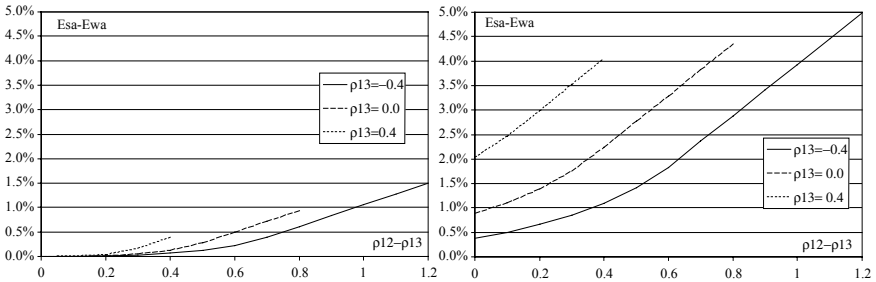

Fig. 4. Values of ΔE_{add}^{ave} for $E_{add}^1 = E_{add}^2 = E_{add}^3 = 5\%$ (balanced performance, left), and $E_{add}^1 = E_{add}^2 = 10\%, E_{add}^3 = 5\%$ (imbalanced performance, right)

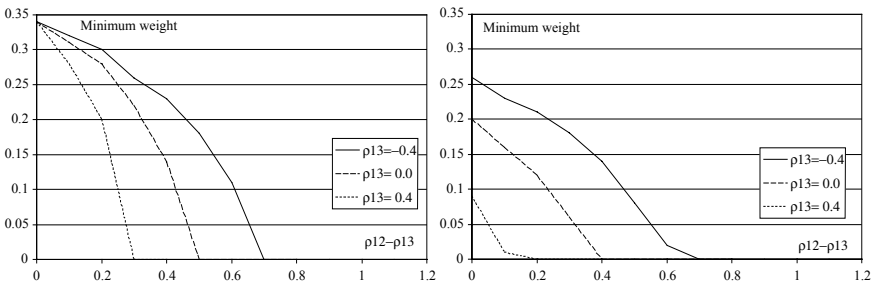

Fig. 5. Values of the minimum of the optimal weights for balanced (left), and imbalanced (right) performance, as in Fig. 5

To sum up, according to our model, weighted averaging significantly improves the performance of simple averaging only for ensembles of classifiers with highly imbalanced performance and correlations. However, such improvement is often achieved by discarding one of the worst classifiers, that is, assigning to it a weight close to zero. When the optimal weights are significantly greater than zero, the advantage of weighted averaging over simple averaging is quite small. It is worth noting that the advantage of weighted averaging over simple averaging is smaller than one can think of. This conclusion is in agreement with experimental results recently reported [6]. Moreover, it should be noted that, in practical applications, it can be very difficult to obtaining good estimates of the optimal weights.

4 Experimental Results

In this section, we report experiments aimed at comparing the performance of simple averaging and weighted averaging for ensembles with different degrees of imbalance.

We used a data set of remote-sensing images related to an agricultural area near the village of Feltwell (UK) [9]. 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.

We considered ensembles made up of a *k*-nearest neighbours classifier (*k*-NN), with a value of *k* equal to 15, and two multi-layer perceptron (MLP1 and MLP2) neural networks. Two ensembles were selected so that the performance of individual classifiers were imbalanced as defined in Sect. 3.1. We used MLPs with 15 hidden units for ensemble 1, and two hidden units for ensemble 2. The test set error percentages of the individual classifiers are shown in Table 1. The difference between the error percentages of the worst and the best classifier is shown in the last column as E_1—E_3. These values are averaged over ten runs corresponding to ten training set / validation set pairs, obtained by sampling without replacement from the original training set. The validation set contained the 20% of patterns of the original training set, and was used for the stopping criterion of the training phase of the MLPs.

Table 1. Error percentages of the individual classifiers on the test set. E_1-E_3 indicates the difference between the error percentages of the worst and the best classifier

	k-NN	MLP1	MLP2	E_1-E_3
ensemble 1	10.01	18.20	18.00	7.99
Ensemble 2	10.01	25.97	26.23	16.22

Table 2. Error percentages of weighted averaging (E^{wa}) and simple averaging (E^{sa}) on the test set

	ensemble performance			optimal weights		
	E^{sa}	E^{wa}	E^{sa}-E^{wa}	*k*-NN	MLP1	MLP2
Ensemble 1	12.09	9.69	2.40	0.689	0.080	0.231
Ensemble 2	16.81	9.79	7.02	0.838	0.006	0.156

In both ensembles, the *k*-NN was the best classifier. The two MLPs exhibited a similar error probability, which was higher than the one of the *k*-NN of about 8% (ensemble 1) and 16% (ensemble 2). As in these experiments we were not interested in the problem of weight estimation, the optimal weights of the linear combination were computed on the test set by "exhaustive" search. The average performance of simple and weighted averaging are reported in Table 2, respectively as E^{sa} and E^{wa}, together with the values of the optimal weights.

For ensemble 1, where the difference E_1—E_3 between the performance of the best and the worst classifier is about 8%, weighted averaging outperforms simple averaging of 2.4%. This value increases to about 7% for ensemble 2, where E_1-E_3 is about 16%. For both cases the weight of MLP1 is close to 0, and the performance of weighted averaging is very close to the one of the best individual classifier. These preliminary results are in agreement with the theoretical predictions. They show that weighted averaging significantly outperforms simple averaging only for highly imbalanced classifiers, and only by discarding one of the worst classifiers.

References

1. Lam, L., Suen, C. Y.: Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. IEEE Trans. on Systems, Man and Cybernetics - Part A 27 (1997) 553-568

2. Tumer, K., Ghosh, J.: Analysis of Decision Boundaries in Linearly Combined Neural Classifiers. Pattern Recognition 29 (1996) 341-348
3. Tumer, K., Ghosh, J.: Linear and Order Statistics Combiners for Pattern Classification. In: Sharkey, A. J. C. (ed.): Combining Artificial Neural Nets. Springer (1999) 127-161
4. Xu, L., Krzyzak, A., Suen, C. Y.: Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Trans. on Systems, Man, and Cybernetics 22 (1992) 418-435
5. Huang, Y. S., Suen, C. Y.: A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals. IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1995) 90-94
6. Ueda, N.: Optimal Linear Combination of Neural Networks for Improving Classification Performance. IEEE Trans. on Pattern Analisys and Machine Int. 22 (2000) 207-215
7. Tumer, K.: Linear and Order Statistics Combiners for Reliable Pattern Classification. PhD thesis. The University of Texas, Austin, TX (1996)
8. Perrone, M., Cooper, L. N.: When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. In: Mammone, R.J. (ed.): Neural Networks for Speech and Image Processing. Chapman-Hall, New York (1993)
9. Roli, F.: Multisensor Image Recognition by Neural Networks with Understandable Behaviour. Int. J. of Pattern Recognition and Artificial Intelligence 10 (1996) 887-917
10. Kittler, J., Roli, F. (eds.): Proc. of the 1st and 2nd Int. Workshop on Multiple Classifier Systems. Springer-Verlag, LNCS, Vol. 1857 (2000), and Vol. 2096 (2001)
11. Tumer, K., Ghosh, J.: Robust Combining of Disparate Classifiers through Order Statistics. To appear in: Pattern Analysis and Applications, special issue on "Fusion of Multiple Classifiers"