A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems

Giorgio Fumera, *Member, IEEE,* Fabio Roli, *Member, IEEE*

Department of Electrical and Electronic Engineering, University of Cagliari

Piazza d'Armi, 09123 Cagliari, Italy

Giorgio Fumera: e-mail fumera@diee.unica.it, phone +39 070 657 5754, fax +39 070 675 5782

Fabio Roli (corresponding author): e-mail roli@diee.unica.it, phone +39 070 657 5779, fax +39 070 675 5782

Abstract

In this paper, a theoretical and experimental analysis of linear combiners for multiple classifier systems is presented. Although linear combiners are the most frequently used combining rules, many important issues related to their operation for pattern classification tasks lack a theoretical basis. After a critical review of the framework developed in works by Tumer and Ghosh, on which our analysis is based, we focus on the simplest and most widely used implementation of linear combiners, which consists in assigning a non-negative weight to each individual classifier. Moreover, we consider the ideal performance of this combining rule, i.e., that achievable when the optimal values of the weights are used. We do not consider the problem of weights estimation, which has been extensively addressed in the literature. Our theoretical analysis shows how the performance of linear combiners, in terms of misclassification probability, depends on the performance of individual classifiers, and on the correlation between their outputs. In particular, we evaluate the ideal performance improvement that can be achieved using the weighted average over the simple average combining rule, and investigate in what way it depends on the individual classifiers. Experimental results on real data sets show that the behaviour of linear combiners agrees with the predictions of our analytical model. Finally, we discuss the contribution to the state of the art and the practical relevance of our theoretical and experimental analysis of linear combiners for multiple classifier systems.

Index Terms

Multiple Classifier Systems, Linear Combiners, Classifier Fusion, Pattern Classification.

I. INTRODUCTION

It is now widely accepted that combining multiple classifiers can provide advantages over the traditional monolithic approach to pattern classifier design [27]. Many experimental works have shown the improvement in performance that can be achieved by multiple classifiers in several applications [18], [26], [34], [27], [21]. Of the various combining rules proposed in the literature, linear combiners are the most frequently used [22], [12], [2], [30], [16], [31], [32], [28], [20], [19]. Simple and weighted averaging of classifiers' outputs are used in popular ensemble learning algorithms such as Bagging [5], Random Subspace Method [15], AdaBoost [7], and represent the baseline, and first choice combiners in the majority of applications. In spite of their wide use and the success of linear combiners, many important issues related to their operation for pattern classification tasks lack a theoretical explanation. To date just partial results are available, and only for classification problems. In particular, only the fraction of the misclassification probability for a single class boundary [30], [31],

and the closed-form expression of the conditional misclassification probability for a given point of the feature space [20], [19], have been derived. Moreover, so far theoretical analyses of classification problems have focussed on the simple average (hereafter, SA), rather than on the performance of the weighted average combining rule (hereafter, WA). To the best of our knowledge, the only exception is [1], where the author extended the results of [31] to the WA, limitedly to the simplest case of unbiased, uncorrelated and identically distributed estimation errors. Other works on WA have addressed the problem of weights estimation, albeit in a regression setting [22], [2], except for [32]. Theoretically speaking, WA is always able to outperform SA. But this is not guaranteed in practice, where weights must be estimated from training data. In real applications, the theoretical superiority of WA can be rapidly negated by weight estimations from small and noisy data sets, to the extent that WA can actually perform worse than SA. In fact, the experimental results reported in the literature do not show any clear superiority over SA. This can be observed in particular for the simplest implementation of WA, that is, when a single non-negative weight is assigned to each classifier [30], [33]. This is the most widely used form of WA, suitable for applications where the weights are intended to represent probabilities (usually, the probability that the corresponding classifier gives the correct answer), or if the resulting linear combination is to be interpreted itself as a class posterior probability estimate. Nevertheless, in [31] it was argued that such an implementation of the WA is not sufficiently more flexible than the SA combining rule. More flexible implementations, at least in principle, have been proposed in [2], [32], where a different set of weights is used for each class, with no sign restrictions. Weights are usually computed by minimising an estimate of the misclassification probability [32], or of the MSE (when neural networks are used as individual classifiers) [22], [2], although it is known that the MSE is not a suitable performance measure for classification problems [3]. It is worth noting that, although using weights which are unrestricted in sign leads in principle to better results, they cannot be reliably estimated for regression problems, especially when the individual regressors are highly correlated. One regularisation method suggested for avoiding this problem is to constrain the weights to be non-negative [17], [11], [4], [13]. The lack of guidelines based on sound theoretical foundations, implies that the practical use of linear combiners for classifier fusion relies on empirically derived heuristic rules. For example, SA is commonly believed to work well for classifiers with similar error rates, while WA is claimed to outperform SA when classifiers exhibit different error rates. However, these rules of thumb are not completely supported by experimental results [24], [25]. Moreover,

the conditions under which WA is able to significantly outperform SA are not clear, neither is the role played by classifier correlations, although the importance of such correlations is quite obvious.

The main goal of this paper is to provide a theoretical analysis of linear combiners which improves the understanding of these classifier fusion rules, and provides some well grounded guidelines for their practical use. Preliminary work on this topic was presented by the authors in [24], [8], [9]. We exploited the analytical framework developed by Tumer and Ghosh [30], [31], for analysing the SA combining rule and order statistics combiners. This framework is critically reviewed in Section II. For readers which are not familiar with this framework, an excellent introduction can be found in [21]. In this paper, we focus on the simplest and most widely used implementation of WA, which consists in assigning a non-negative weight to each individual classifier. Moreover, we consider the ideal performance of this combining rule, i.e. that achievable when the optimal values of the weights are used. We do not consider the problem of weights estimation, which has been addressed in the works cited above. Our theoretical analysis, presented in Section III, shows how, in terms of misclassification probability, the performance of the SA and WA combining rules depends on the performance of individual classifiers, and on the correlation between their outputs. In particular, our analysis evaluates the ideal improvement in performance that can be achieved by WA over SA, and provides a better understanding of its dependency on the individual classifiers of the ensemble. In Section IV, we describe the experimental investigations conducted to assess the extent to which the behaviour of SA and WA, evaluated on real data sets, agrees with the predictions of the analytical model presented in Section III, which is based on assumptions that are likely to be violated in real applications. Finally, we discuss the contribution to the state of the art and the practical relevance of our theoretical and experimental analysis of linear combiners for multiple classifier systems (Sections V, VI).

II. A Critical Review of the Analytical Framework by Tumer and Ghosh

A. Basic Concepts and Main Results

According to the Bayesian decision theory, the minimum of the misclassification probability (the so called Bayes error) is obtained by assigning an input pattern, characterised by the feature vector $\mathbf{x}$, to the class ω_i exhibiting the maximum posterior probability: $i = \arg\max_k P(\omega_k|\mathbf{x})$. However, real classifiers can only provide estimates $f_k(\mathbf{x})$ of the posterior probabilities $P(\omega_k|\mathbf{x})$. Therefore, when Bayes rule is applied to these estimates, non-optimal

decisions are taken for patterns for which $\arg\max_k f_k(\mathbf{x}) \neq \arg\max_k P(\omega_k|\mathbf{x})$. This results in an additional misclassification probability over Bayes error. Tumer and Ghosh analysed the case in which the effect of estimation errors consists in a shift of the ideal class boundaries, and calculated the expected value of the additional misclassification probability, named "added error", around a single ideal class boundary. For readers which are not familiar with this framework, an excellent introduction can be found in [21].

The estimated posterior probability for the k-th class can be written as[1]

$$f_k(x) = P(\omega_k|x) + \epsilon_k(x) \ , \tag{1}$$

where $\epsilon_k(x)$ denotes the estimation error, which is regarded as a random variable with mean β_k (named "bias") and variance σ_k^2. In [30], [31] it was assumed that, for any given x, the estimation errors on *different* classes, i.e. $\epsilon_i(x)$ and $\epsilon_j(x)$, $i \neq j$, are uncorrelated.[2] The ideal boundary between any two classes ω_i and ω_j is the point x^* such that

$$P(\omega_i|x^*) = P(\omega_j|x^*) > \max_{k \neq i,j} P(\omega_k|x^*) \ . \tag{2}$$

As pointed out above, Tumer and Ghosh analysed the case in which the effect of estimation errors consists in a shift of the ideal class boundaries. This situation is illustrated in Fig. 1. The boundary obtained from the estimated posteriors, denoted with x_b, is characterised by $f_i(x_b) = f_j(x_b)$, and may differ from the ideal boundary by an amount $b = x_b - x^*$. Without loss of generality, in Fig. 1 it is assumed that $b > 0$. It is worth pointing out that, as discussed in [21] (section 9.2.4), in the one-dimensional case the estimation errors can cause two other effects, besides the shift of a class boundary: a boundary could be created where there is none, or an existing boundary could be not detected. These two situations are not unlikely to occur when more than two classes are highly overlapping [21]. However, following the analysis by Tumer and Ghosh, in this paper we shall consider only the case of a shitf of a class boundary, which is perhaps the more likely to occur, and is also easier to analyse.

Since we assumed $b > 0$, from Fig. 1 it is easy to see that the added error is caused by the patterns falling in the interval (x^*, x_b), which are assigned to class ω_i instead of ω_j. The

---

[1]In this paper, we will consider the case of a one-dimensional feature space, as in [30], [31]. This is not a limitation, since in [29] it was shown that the same results hold for the case of multi-dimensional feature spaces.

[2]Note that this assumption does not hold if the estimates of the posterior probabilities are sum-to-one constrained, i.e. $\sum_k f_k(x) = 1$. In actual fact, from (1), this would imply $\sum_k \epsilon_k(x) = 0$, and therefore the $\epsilon_k(x)$'s cannot be uncorrelated.
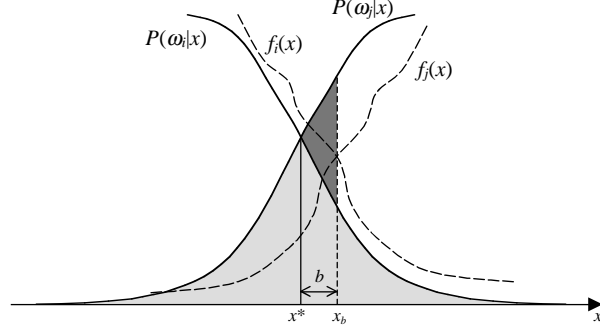
Fig. 1. "True" posterior probabilities around the boundary x^* between classes ω_i and ω_j (solid lines), and estimated posteriors, leading to the boundary x_b, (dashed lines). Lightly and darkly shaded areas represent contribution of this class boundary to Bayes error and to added error, respectively.

added error depends on the offset b, and is given by

$$\int_{x^*}^{x^*+b} [P(\omega_j|x) - P(\omega_i|x)]\, p(x)\mathrm{d}x \quad, \tag{3}$$

where $p(x)$ is the probability distribution of the feature x. To compute the above integral, assuming small values of the shift $b = x_b - x^*$, a first-order approximation is used in [30], [31] for the posteriors of classes ω_i and ω_j around the ideal boundary x^*:

$$P(\omega_k|x^* + b) = P(\omega_k|x^*) + bP'(\omega_k|x^*) \quad. \tag{4}$$

Furthermore, $p(x)$ is approximated by the constant value $p(x^*)$. Substituting (4) into (3), and $p(x^*)$ for $p(x)$, the added error (3) becomes $\frac{p(x^*)t}{2}b^2$, where $t = P'(\omega_j|x^*) - P'(\omega_i|x^*)$. The expected value of the added error with respect to b, denoted with E_{add}, is then given by

$$E_{\mathrm{add}} = \frac{p(x^*)t}{2}(\beta_b^2 + \sigma_b^2) \quad, \tag{5}$$

where β_b and σ_b^2 denote respectively the bias and the variance of b. The value of b, and thus of E_{add}, can be expressed as a function of the estimation errors. From (1) and (4), the equation $f_i(x_b) = f_j(x_b)$ can be rewritten as $P(\omega_i|x^*) + bP'(\omega_i|x^*) + \epsilon_i(x_b) = P(\omega_j|x^*) + bP'(\omega_j|x^*) + \epsilon_j(x_b)$. Since $P(\omega_i|x^*) = P(\omega_j|x^*)$, from the above equation one obtains

$$b = \frac{\epsilon_i(x_b) - \epsilon_j(x_b)}{t} \quad. \tag{6}$$

As the estimation errors on different classes are uncorrelated [30], [31], it follows that the bias and variance of b in (5) are given by

$$\beta_b = \frac{\beta_i - \beta_j}{t}, \ \sigma_b^2 = \frac{\sigma_i^2 + \sigma_j^2}{t^2} \quad. \tag{7}$$

Consider now a linear combination, by simple averaging, of the estimated posterior probabilities provided by an ensemble of N classifiers. In the following, we shall denote quantities related to the m-th individual classifier and to SA, with the superscripts m and "sa" respectively. From (1), the averaged estimates for the k-th class can be written as

$$f_k^{\text{sa}}(x) = \frac{1}{N} \sum_{m=1}^{N} f_k^m(x) = P(\omega_k|x) + \epsilon_k^{\text{sa}}(x) \ , \tag{8}$$

where

$$\epsilon_k^{\text{sa}}(x) = \frac{1}{N} \sum_{m=1}^{N} \epsilon_k^m(x) \ . \tag{9}$$

With reference to the same class boundary of Fig. 1, the class boundary $x_{b^{\text{sa}}}$ obtained from the posteriors estimated by simple averaging of classifier outputs (8) has an offset b^{sa} from x^*, and is characterised by $f_i^{\text{sa}}(x^* + b^{\text{sa}}) = f_j^{\text{sa}}(x^* + b^{\text{sa}})$. Following the same steps as above, one obtains

$$b^{\text{sa}} = \frac{\epsilon_i^{\text{sa}}(x_{b^{\text{sa}}}) - \epsilon_j^{\text{sa}}(x_{b^{\text{sa}}})}{t} \ , \tag{10}$$

$$E_{\text{add}}^{\text{sa}} = \frac{p(x^*)t}{2} \left(\beta_{b^{\text{sa}}}^2 + \sigma_{b^{\text{sa}}}^2 \right) \ . \tag{11}$$

Also the expected added error of the average fusion rule (11) can be expressed as a function of the estimation errors of individual classifiers. First, from (10) and (9), the bias of b^{sa} is

$$\beta_{b^{\text{sa}}} = \frac{\beta_i^{\text{sa}} - \beta_j^{\text{sa}}}{t} = \frac{1}{N} \sum_{m=1}^{N} \frac{\beta_i^m - \beta_j^m}{t} = \frac{1}{N} \sum_{m=1}^{N} \beta_{b^m} \ , \tag{12}$$

where β_{b^m} is given in (7). Then, $\sigma_{b^{\text{sa}}}^2$ can be computed by first noting that (9) implies

$$(\sigma_k^{\text{sa}})^2 = \frac{1}{N^2} \sum_{m=1}^{N} (\sigma_k^m)^2 + \frac{1}{N^2} \sum_{m=1}^{N} \sum_{n \neq m} \rho_k^{mn} \sigma_k^m \sigma_k^n \ , \tag{13}$$

where ρ_k^{mn} denotes the correlation coefficient between $\epsilon_k^m(x)$ and $\epsilon_k^n(x)$, and σ_k^m is the standard deviation of $\epsilon_k^m(x)$. The assumption that the estimation errors on different classes are uncorrelated was extended in [30], [31] to errors of different classifiers, i.e. $\epsilon_i^m(x)$ and $\epsilon_j^n(x), i \neq j$, are assumed to be uncorrelated. Accordingly, from (10) it follows that $\sigma_{b^{\text{sa}}}^2 = \frac{1}{t^2} \left[(\sigma_i^{\text{sa}})^2 + (\sigma_j^{\text{sa}})^2 \right]$. Substituting (13) in the above expression, one obtains

$$\sigma_{b^{\text{sa}}}^2 = \frac{1}{N^2} \sum_{m=1}^{N} \sigma_{b^m}^2 + \frac{1}{t^2} \frac{1}{N^2} \sum_{m=1}^{N} \sum_{n \neq m} \left(\rho_i^{mn} \sigma_i^m \sigma_i^n + \rho_j^{mn} \sigma_j^m \sigma_j^n \right) \ , \tag{14}$$

where $\sigma_{b^m}^2$ is given in (7).

The above results show that the expected added error on a given class boundary is given by the sum of two components: one depending on the bias of estimation errors, the other on their

variance. In particular, for linearly combined classifiers, the expected added error also depends on classifiers' pair-wise correlations. In [30], [31], these results were exploited to quantify the reduction of the expected added error that can be achieved by simple averaging (11) with respect to the individual classifiers (5). The results obtained can be summarised as follows. First, note that the reduction of the bias and variance components can be evaluated separately. From (12) it follows that, though the bias component is not necessarily reduced with respect to each individual classifier, it is at least guaranteed to be no greater than the maximum bias exhibited by the individual classifiers. The variance component (14) was analysed only for the case of estimation errors with identical variances, i.e. $(\sigma_k^m)^2 = \sigma^2 \; \forall m, k$. In this case, (7) becomes $\sigma_b^2 = \frac{1}{t^2} 2\sigma^2$, while (13) can be rewritten as

$$(\sigma_k^{\mathrm{sa}})^2 = \frac{1 + (N-1)\delta_k}{N} \sigma^2 \; , \tag{15}$$

where $\delta_k = \frac{1}{N(N-1)} \sum_{m=1}^{N} \sum_{n \neq m} \rho_k^{mn}$. It follows that (14) can be rewritten as

$$\sigma_{b^{\mathrm{sa}}}^2 = \frac{1 + (N-1)\delta_{ij}}{N} \frac{1}{t^2} 2\sigma^2 = \frac{1 + (N-1)\delta_{ij}}{N} \sigma_b^2 \; , \tag{16}$$

where $\delta_{ij} = \frac{\delta_i + \delta_j}{2}$. It is easy to see that $\frac{1+(N-1)\delta_{ij}}{N} \leq 1$. This means that, if the estimation errors have identical variances, then the averaging rule reduces the variance component of the added error of each individual classifier, by a factor depending on the correlation between the estimation errors: the lower the correlation, the greater the reduction. In particular, if the estimation errors are uncorrelated (i.e. $\rho_k^{mn} = 0 \; k = i, j, \; \forall m, n$), then $\delta_{ij} = 0$, and therefore the reduction factor is exactly N, the number of combined classifiers. It should be noted that, in practice, it is very difficult to obtain a large number of uncorrelated classifiers. From a practical standpoint, these results suggest using the following approach to the problem of bias-variance trade-off [10]: the design of individual classifiers should focus on obtaining low bias and correlation, rather than low variance, since the variance can be reduced by averaging classifiers.

B. Some Generalizations of Tumer and Ghosh Results

Let us now show how the results reported in [30], [31], and summarised above, can be extended to the more general case of estimation errors with negative correlation and non-identical variances. First, if the estimation errors are negatively correlated, besides having identical variances, then the variance component (16) of the added error can be reduced by a factor greater than N. Theoretically, it can be reduced up to zero for *any finite* N: this

occurs when $\delta_i = \delta_j = -\frac{1}{N-1}$ (note that $\delta_k \geq -\frac{1}{N-1} \ \forall k$, since the variance in (15) must be non-negative). For instance, it is easy to see that this happens when the correlations are all equal to $-\frac{1}{N-1}$. Moreover, it is easy to see that a low correlation is advantageous also when the estimation errors have non-identical variances: in fact, from (14) it is evident that the lower the correlations, the lower is the variance component of the expected added error. This result confirms that the design of individual classifiers can focus on obtaining low biases and low pair-wise correlations, even if this design produces classifiers with high and different variances, that is, classifiers which perform differently on unseen test data.

Finally, it can be shown that, according to the above framework, the expected added error of the averaging rule, on a given class boundary, is never greater than the maximum expected added error of individual classifiers, $\max_m E_{\mathrm{add}}^m$. Indeed, from (10) and (9), it is easy to see that $b^{\mathrm{sa}} = \frac{1}{N} \sum_{m=1}^N b^m$. This implies that $\beta_{b^{\mathrm{sa}}}^2 \leq \max_m \beta_{b^m}^2$, and $\sigma_{b^{\mathrm{sa}}}^2 \leq \max_m \sigma_{b^m}^2$. It follows from (5) that $E_{\mathrm{add}}^{\mathrm{sa}} \leq \max_m E_{\mathrm{add}}^m$, where the equality holds only in the limit case in which all the correlations are equal to one, and all the biases β_{b^m} are identical. The practical relevance of this result becomes apparent when one considers that one of the main reasons for combining classifiers is to avoid the "worst case" of the traditional "evaluation and selection" approach to classifier design: selecting the "apparent" best classifier from a given ensemble, on the basis of validation data, involves the risk of obtaining the worst classifier on unseen test data [6], [3]. Accordingly, a desirable property for a combining rule is that it guarantees a better test set performance than the worst classifier of the ensemble. The above result theoretically supports the fact that the averaging rule exhibits this property.

III. THEORETICAL ANALYSIS AND COMPARISON OF SIMPLE AND WEIGHTED AVERAGING

In this section, we present a theoretical analysis of the WA rule. First we derive the expected added error of the WA rule. We then quantify the reduction that can be obtained over individual classifiers, and compare it with the reduction obtained using the SA rule.

A. Expected Added Error of the Weighted Average

As explained in Section I, we consider the simplest form of WA, which consists in assigning a non-negative weight to each individual classifier w_m, with at least one weight greater than zero. This results in the following approximation of the posterior probabilities for the k-th

class (hereafter, the superscript "wa" will denote quantities related to the WA):

$$f_k^{\mathrm{wa}}(x) = \sum_{m=1}^{N} w_m f_k^m(x) = P(\omega_k|x) + \epsilon_k^{\mathrm{wa}}(x) \ , \tag{17}$$

where

$$\epsilon_k^{\mathrm{wa}}(x) = \sum_{m=1}^{N} w_m \epsilon_k^m(x) \ . \tag{18}$$

The constraints on the weights can be written as

$$w_m \geq 0 \ m = 1, \ldots, N, \quad \sum_{m=1}^{N} w_m > 0 \ . \tag{19}$$

Following the same procedure described in Section II, the expected added error of the WA rule, on a given boundary between classes ω_i and ω_j (see Fig. 1), can be written

$$E_{\mathrm{add}}^{\mathrm{wa}} = \frac{p(x^*)t}{2} \left(\beta_{b^{\mathrm{wa}}}^2 + \sigma_{b^{\mathrm{wa}}}^2 \right) \ , \tag{20}$$

where the offset b^{wa} is given by

$$b^{\mathrm{wa}} = \frac{\epsilon_i^{\mathrm{wa}}(x_{b^{\mathrm{wa}}}) - \epsilon_j^{\mathrm{wa}}(x_{b^{\mathrm{wa}}})}{t} \ . \tag{21}$$

Assuming as in [30], [31] that errors on different classes are uncorrelated, from (21) and (18) we obtain:

$$\beta_{b^{\mathrm{wa}}}^2 = \frac{1}{t^2} \sum_{m=1}^{N} w_m^2 \left(\beta_i^m - \beta_j^m \right)^2 + \frac{1}{t^2} \sum_{m=1}^{N} \sum_{n \neq m} w_m w_n \left(\beta_i^m - \beta_j^m \right) \left(\beta_i^n - \beta_j^n \right), \tag{22}$$

$$\sigma_{b^{\mathrm{wa}}}^2 = \frac{1}{t^2} \sum_{m=1}^{N} w_m^2 \left[(\sigma_i^m)^2 + (\sigma_j^m)^2 \right] + \frac{1}{t^2} \sum_{m=1}^{N} \sum_{n \neq m} w_m w_n \left(\rho_i^{mn} \sigma_i^m \sigma_i^n + \rho_j^{mn} \sigma_j^m \sigma_j^n \right). \tag{23}$$

As pointed out in Section I, we are interested in the "optimal" weights, i.e. those that minimise $E_{\mathrm{add}}^{\mathrm{wa}}$ under the constraints (25), and therefore maximise the reduction of the expected added error with respect to the individual classifiers. Obviously, when the optimal weights are used, then $E_{\mathrm{add}}^{\mathrm{wa}} \leq E_{\mathrm{add}}^{\mathrm{sa}}$. In actual fact, the SA rule is a particular case of WA, when $w_m = \frac{1}{N} \ \forall m$. It is also easy to see that the optimal weights lead to an overall misclassification probability no higher than that of the best classifier of the ensemble: indeed, the same performance of the best classifier is achieved when the corresponding weight equals 1 and all the others 0. Instead, we have shown in Section II that SA only guarantees an expected added error no greater than that of the worst classifier, limited to a single class boundary.

Substituting (22) and (23) into (20), $E_{\text{add}}^{\text{wa}}$ can be written in a compact form as $w^{\text{T}}\mathbf{M}w$, where w is the vector $(w_1, \ldots, w_N)^{\text{T}}$, and $\mathbf{M}$ is a symmetric N by N matrix, whose elements are (from (22) and (23)):

$$
\begin{aligned}
M_{mm} &= \frac{1}{t^2} \left[\left(\beta_i^m - \beta_j^m\right)^2 + (\sigma_i^m)^2 + (\sigma_j^m)^2 \right], \ m = 1, \ldots, N \ , \\
M_{mn} &= \frac{1}{t^2} \left[\left(\beta_i^m - \beta_j^m\right)\left(\beta_i^n - \beta_j^n\right) + \left(\rho_i^{mn}\sigma_i^m\sigma_i^n + \rho_j^{mn}\sigma_j^m\sigma_j^n\right) \right], \ m \neq n \ .
\end{aligned}
\tag{24}
$$

Note that, for classification problems, it is possible to replace (19) with the following equivalent constraints, which are computationally more convenient:[3]

$$
w_m \geq 0 \ m = 1, \ldots, N, \quad \sum_{m=1}^{N} w_m = 1 \ .
\tag{25}
$$

The optimal weights are thus the solution of the following optimisation problem:

$$
\text{minimise} \quad E_{\text{add}}^{\text{wa}} = w^{\text{T}}\mathbf{M}w, \quad \text{subject to} \quad w_m \geq 0 \ i = 1, \ldots, N, \quad \sum_{m=1}^{N} w_m = 1 \ .
\tag{26}
$$

This is a quadratic programming problem, that can be solved using standard optimisation techniques. However, the analytical solution can be obtained only when the matrix $\mathbf{M}$ is diagonal. A particular case of theoretical interest in which $\mathbf{M}$ is diagonal, is when the estimation errors are unbiased and uncorrelated, i.e. $\beta_k^m = 0$, $\rho_k^{mn} = 0$, $\forall m, n, k$ (see (24)). For this reason, in the following we will focus our analysis on the case of unbiased and uncorrelated errors. We will then show how some results can be extended to the case of unbiased and *correlated* errors.

B. Unbiased and Uncorrelated Estimation Errors

If the estimation errors are unbiased and uncorrelated, the expected added error of WA (from (22) and (23)) is

$$
E_{\text{add}}^{\text{wa}} = \frac{p(x^*)}{2t} \sum_{m=1}^{N} w_m^2 \left[(\sigma_i^m)^2 + (\sigma_j^m)^2 \right] \ .
\tag{27}
$$

It follows that the matrix $\mathbf{M}$ is diagonal, and the problem (26) can be analytically solved using the Lagrange multiplier technique. Moreover, from (5) and (12) it follows that

$$
E_{\text{add}}^m = \frac{p(x^*)}{2t} \left[(\sigma_i^m)^2 + (\sigma_j^m)^2 \right] \ ,
\tag{28}
$$

---

[3]We recall that a pattern x is assigned to the class corresponding to $\arg\max_k f_k^{\text{wa}}(x)$. Therefore, for any set of non-negative weights with positive sum (19), a set of non-negative weights that sum to one (25) can be found, such that the class assigned to each pattern does not change. The latter weights can be obtained simply by rescaling the former ones by a positive constant.

and therefore $E_{\text{add}}^{\text{wa}}$ can be simply written as the linear combination of the added errors of individual classifiers:

$$E_{\text{add}}^{\text{wa}} = \sum_{m=1}^{N} w_m^2 E_{\text{add}}^m \quad . \tag{29}$$

Hence we have $\mathbf{M} = \text{diag}(E_{\text{add}}^1, \ldots, E_{\text{add}}^N)$. In this way we can analyse the performance of the SA and WA rules as a function of the expected added errors of individual classifiers, instead of the variances of their estimation errors. The optimal weights are the following:

$$w_m = \left(\sum_{n=1}^{N} \frac{1}{E_{\text{add}}^n} \right)^{-1} \frac{1}{E_{\text{add}}^m} \quad . \tag{30}$$

This shows that, for unbiased and uncorrelated errors, the optimal weights are inversely proportional to the expected added error of the corresponding classifiers. This implies that SA, i.e. $w_m = \frac{1}{N}$, is the optimal linear combining rule, only if the individual classifiers exhibit identical expected added errors. This result provides a theoretical support to the common claim that the SA rule is appropriate for combining classifiers with similar performances (see for instance [30]), while different weights should be used for classifiers of different strength.

Substituting (30) into (29), we obtain the value of $E_{\text{add}}^{\text{wa}}$ corresponding to the optimal weights:

$$E_{\text{add}}^{\text{wa}} = \frac{1}{\frac{1}{E_{\text{add}}^1} + \cdots + \frac{1}{E_{\text{add}}^N}} \quad . \tag{31}$$

On the other hand, if the SA rule is used, setting $w_m = \frac{1}{N}$ in (29) we obtain:

$$E_{\text{add}}^{\text{sa}} = \frac{1}{N^2} \sum_{m=1}^{N} E_{\text{add}}^m \quad . \tag{32}$$

This shows that $E_{\text{add}}^{\text{wa}}$ and $E_{\text{add}}^{\text{sa}}$ are equal to $\frac{1}{N}$ times the harmonic mean and the arithmetic mean, respectively, of the E_{add}^m's.

In the following section we evaluate the improvement of the SA and WA over individual classifiers. Note that our analysis of SA extends the analysis presented in [30], [31], which was limited to the case of estimation errors with identical variances, and thus, in the case of unbiased and uncorrelated errors, to the case of classifiers with identical expected added errors (see (5) and (12)), that is, identical performances. Without losing generality, we will assume in the following that the classifier errors are arranged as follows:

$$E_{\text{add}}^1 \leq E_{\text{add}}^2 \leq \cdots \leq E_{\text{add}}^N \quad . \tag{33}$$

Accordingly, we will call classifiers 1 and N the "best" and "worst" classifier respectively, relatively to the class boundary considered, and the interval $\left[E_{\mathrm{add}}^1, E_{\mathrm{add}}^N\right]$ as the "error range" of the classifier ensemble. We will also assume $E_{\mathrm{add}}^1 > 0$, that is, none of the individual classifiers achieves the Bayes error. Finally, we point out that (33) implies that the condition that the individual classifiers exhibit identical expected added errors can be simply written as $E_{\mathrm{add}}^1 = E_{\mathrm{add}}^N$.

1) Comparison with Individual Classifiers: When the individual classifiers exhibit identical expected added errors (i.e., $E_{\mathrm{add}}^1 = E_{\mathrm{add}}^N$), we showed that the optimal weights are $w_m = \frac{1}{N}$. It follows that $E_{\mathrm{add}}^{\mathrm{wa}} = E_{\mathrm{add}}^{\mathrm{sa}}$, and therefore (29) implies that a reduction of the expected added error by a factor N is achieved over each individual classifier. On the other hand, if the individual classifiers exhibit non-identical added errors (i.e., if $E_{\mathrm{add}}^1 < E_{\mathrm{add}}^N$), then $E_{\mathrm{add}}^{\mathrm{wa}} < E_{\mathrm{add}}^{\mathrm{sa}}$. In this case, from (32), the error reduction achieved by SA over the generic m-th classifier is

$$\frac{E_{\mathrm{add}}^{\mathrm{sa}}}{E_{\mathrm{add}}^m} = \frac{1}{N^2}\left(1 + \sum_{n \neq m} \frac{E_{\mathrm{add}}^n}{E_{\mathrm{add}}^m}\right) \quad . \tag{34}$$

In particular, taking into account that $0 < E_{\mathrm{add}}^m < 1 \ \forall m$, it follows that the reduction factors over the best and worst classifiers take values in the following ranges:

$$\frac{E_{\mathrm{add}}^{\mathrm{sa}}}{E_{\mathrm{add}}^1} \in \left(\frac{1}{N}, +\infty\right), \quad \frac{E_{\mathrm{add}}^{\mathrm{sa}}}{E_{\mathrm{add}}^N} \in \left(\frac{1}{N^2}, \frac{1}{N}\right) \tag{35}$$

(The lower bound for the reduction factor over E_{add}^1 is obtained when all the N classifiers exhibit the same added error, while the upper bound is obtained for $E_{\mathrm{add}}^1 \to 0$, for any fixed value of the added error of the other classifiers. Similar considerations lead to the lower and upper bounds for $\frac{E_{\mathrm{add}}^{\mathrm{sa}}}{E_{\mathrm{add}}^N}$.) From (31), the reduction factor of WA is

$$\frac{E_{\mathrm{add}}^{\mathrm{wa}}}{E_{\mathrm{add}}^m} = \left(1 + \sum_{n \neq m} \frac{E_{\mathrm{add}}^m}{E_{\mathrm{add}}^n}\right)^{-1} \quad , \tag{36}$$

and therefore:

$$\frac{E_{\mathrm{add}}^{\mathrm{wa}}}{E_{\mathrm{add}}^1} \in \left(\frac{1}{N}, 1\right), \quad \frac{E_{\mathrm{add}}^{\mathrm{wa}}}{E_{\mathrm{add}}^N} \in \left(0, \frac{1}{N}\right) \quad . \tag{37}$$

From (34)–(37) it follows that, if $E_{\mathrm{add}}^1 < E_{\mathrm{add}}^N$, then both SA and WA achieve a reduction *lower* than N over the best classifier, and *greater* than N over the worst one. However, (35) and (37) show that $E_{\mathrm{add}}^{\mathrm{wa}}$ is always smaller than E_{add}^1, i.e., WA always performs better than the best individual classifier. Instead, $E_{\mathrm{add}}^{\mathrm{sa}}$ can become arbitrarily larger than E_{add}^1, depending on the particular values of $E_{\mathrm{add}}^1, \ldots, E_{\mathrm{add}}^N$. Moreover, the reduction achieved by WA over the

worst classifier E_{add}^N can be arbitrarily large, while the maximum reduction achievable by SA is $\frac{1}{N^2}$.

We will now analyse further the dependence of the performance of the two combining rules on the expected added errors of individual classifiers. Consider ensembles of fixed size N and fixed error range $\left[E_{\text{add}}^1, E_{\text{add}}^N\right]$. Which values of $E_{\text{add}}^2, \ldots, E_{\text{add}}^{N-1}$ yield the minimum and maximum values of $E_{\text{add}}^{\text{sa}}$ and $E_{\text{add}}^{\text{wa}}$ (equivalently, the maximum and minimum reduction factors over E_{add}^1 and E_{add}^N)? The answer can be easily obtained from (35) and (37): the minimum of $E_{\text{add}}^{\text{sa}}$ and $E_{\text{add}}^{\text{wa}}$ is achieved when $E_{\text{add}}^m = E_{\text{add}}^1$, $m = 2, \ldots, N-1$; the maximum is achieved when $E_{\text{add}}^m = E_{\text{add}}^N$, $m = 2, \ldots, N-1$. The above analytical results basically confirm the following intuitive result for linear combiners: of all the ensembles of N classifiers exhibiting a given error range, SA and WA are equally effective as the performance of $N-2$ classifiers $2, \ldots, N-1$ approaches that of the best classifier.

Finally, let us consider what happens when a new classifier is added to a given ensemble. From (37) it follows that this *always* leads to an improvement in the performance of the WA rule, whatever the expected added error of the new classifier. On the other hand, this does not always hold for the SA rule: (32) implies that $E_{\text{add}}^{\text{sa}}$ improves only if the expected added error of the new classifier is smaller than $\frac{2N+1}{N^2} \sum_{m=1}^N E_{\text{add}}^m$.

Summing up, the analytical results obtained for the case of unbiased and uncorrelated errors provide a theoretical support to the common claim that the WA rule can counterbalance the effects of uneven performances of individual classifiers, and can always outperform the best classifier, even when classifiers with arbitrarily high error probability are added to a given ensemble.

2) Comparison between Simple and Weighted Averaging: Let us now consider the following question: given an ensemble of classifiers, how much theoretical reduction of error probability can be obtained using the WA instead of the SA rule, and how does this reduction depend on the individual classifiers? The practical relevance of this problem becomes apparent when one considers that small reductions in error can be negated by weight estimations from small and noisy data sets (Section I), resulting in poorer WA performances compared to SA. The improvement ΔE achievable for a given class boundary by WA is given by the difference $E_{\text{add}}^{\text{sa}} - E_{\text{add}}^{\text{wa}}$, since the Bayes error is obviously the same for both rules. If the estimation errors are unbiased and uncorrelated, and the optimal weights are used, from (31) and (32) we obtain $\Delta E = \frac{1}{N^2} \sum_{m=1}^N E_{\text{add}}^m - \frac{1}{\sum_{m=1}^N \frac{1}{E_{\text{add}}^m}}$. We already know that, when the added error of individual classifiers is different, then WA outperforms SA, i.e. $\Delta E > 0$,

otherwise $\Delta E = 0$. To analyse the behaviour of ΔE as a function of the E_{add}^m's, we first consider classifier ensembles with fixed size N, and fixed error range $\left[E_{\text{add}}^1, E_{\text{add}}^N\right]$. In this case, it is interesting to analyse the conditions under which ΔE is minimum and maximum, with respect to the values of $E_{\text{add}}^2, \ldots, E_{\text{add}}^{N-1}$. We found (for brevity the proof is reported in [9]) that the minimum of ΔE is obtained when the expected added errors of classifiers $2, \ldots, N-1$ are all equal to the harmonic mean of E_{add}^1 and E_{add}^N:

$$E_{\text{add}}^m = \frac{2E_{\text{add}}^1 E_{\text{add}}^N}{E_{\text{add}}^1 + E_{\text{add}}^N} \quad m = 2, \ldots, N-1 \ . \tag{38}$$

The corresponding value of ΔE is

$$\min_{E_{\text{add}}^2, \ldots E_{\text{add}}^{N-1}} \Delta E = \frac{1}{N^2} \frac{\left(E_{\text{add}}^1 - E_{\text{add}}^N\right)^2}{E_{\text{add}}^1 + E_{\text{add}}^N} \ . \tag{39}$$

On the other hand, ΔE is maximum when the expected added error of k classifiers is equal to E_{add}^N, while that of the other $N-k$ classifiers is equal to E_{add}^1:

$$E_{\text{add}}^m = E_{\text{add}}^1, \ m = 2, \ldots, N-k; \quad E_{\text{add}}^m = E_{\text{add}}^N, \ m = N-k+1, \ldots, N-1 \ . \tag{40}$$

We found that the value of k is given either by $\lfloor k^* \rfloor$ or by $\lceil k^* \rceil$, where

$$k^* = N \frac{\sqrt{E_{\text{add}}^1 E_{\text{add}}^N} - E_{\text{add}}^1}{E_{\text{add}}^N - E_{\text{add}}^1} \ . \tag{41}$$

The maximum of ΔE w.r.t. $E_{\text{add}}^2, \ldots E_{\text{add}}^{N-1}$ is then given by

$$\frac{1}{N^2} \left[(N-k)E_{\text{add}}^1 + kE_{\text{add}}^N\right] - \frac{1}{(N-k)\frac{1}{E_{\text{add}}^1} + k\frac{1}{E_{\text{add}}^N}} \ , \tag{42}$$

which can be approximated using (41) as

$$\max_{E_{\text{add}}^2, \ldots E_{\text{add}}^{N-1}} \Delta E \approx \frac{1}{N} \frac{\sqrt{E_{\text{add}}^1 E_{\text{add}}^N}(\sqrt{E_{\text{add}}^1} - \sqrt{E_{\text{add}}^N})^2}{E_{\text{add}}^1 + E_{\text{add}}^N - \sqrt{E_{\text{add}}^1 E_{\text{add}}^N}} \ . \tag{43}$$

Note that for $N = 3$, we found that $k = 2$, whatever the values of E_{add}^1 and E_{add}^N.

The two conditions (38) and (40) can be described by introducing the concept of performance "imbalance", related to the improvement that can be achieved by WA with respect to SA. If the expected added errors (and thus the overall error probabilities) are identical, then the performances of the individual classifiers are said to be balanced, otherwise they are said to be imbalanced. Accordingly, the WA rule outperforms the SA rule for classifier ensembles with imbalanced performances. Now, let us consider two ensembles A and B with the same size and error range, but having different values of $E_{\text{add}}^2, \ldots, E_{\text{add}}^{N-1}$. If $\Delta E_{\text{A}} > \Delta E_{\text{B}}$, then we can say that the performances of A are more imbalanced than those of B, in the sense that the improvement achieved by WA for ensemble A with respect to SA is greater than that
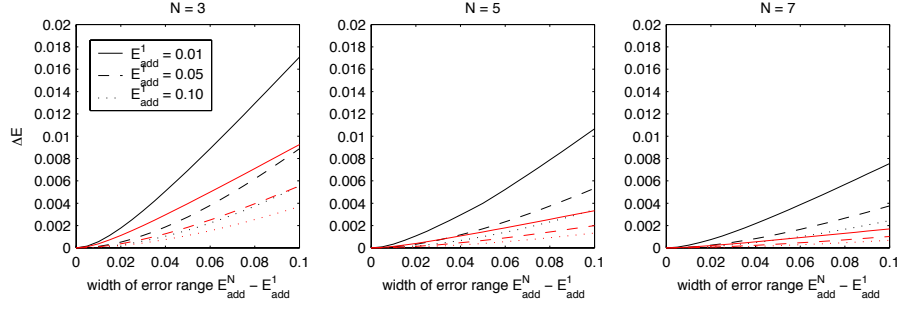
Fig. 2. Maximum and minimum values of $\Delta E = E_{\mathrm{add}}^{\mathrm{sa}} - E_{\mathrm{add}}^{\mathrm{wa}}$ versus error range width $E_{\mathrm{add}}^N - E_{\mathrm{add}}^1$, for $N = 3, 5, 7$, $E_{\mathrm{add}}^1 = 0.01, 0.05, 0.10$, and E_{add}^N ranging from E_{add}^1 to $E_{\mathrm{add}}^1 + 0.10$.

achieved for B. Accordingly, when the condition (38) is verified, we can say that a classifier ensemble exhibits the smallest performance imbalance, with respect to all the other ensembles with identical size and error range, since the corresponding ΔE is minimum, among all such ensembles. For this reason, we can denote (38) as the condition of *minimum performance imbalance*. Similarly, we can say that the largest performance imbalance is achieved under condition (40), which can be denoted as the condition of *maximum performance imbalance*. It is not surprising that the condition of minimum performance imbalance corresponds to identical values of E_{add}^m, $m = 2, \ldots, N-1$, in light of the fact that $\Delta E = 0$ when all the N classifiers exhibit balanced performances, i.e. identical expected added errors. On the other hand, the condition of maximum performance imbalance (40), which is the less favourable to the SA rule, is not so obvious, and can be described as follows: the expected added errors of individual classifiers are divided into two compact "clusters" of values, characterised by maximum intra-class distance (equal to the width of the error range).

Let us now complete the analysis of the behaviour of ΔE, varying the values of N, E_{add}^1 and E_{add}^N. For simplicity, we will conduct an experimental analysis of the behaviour of ΔE for some values of N and different error ranges. Fig. 2 shows the maximum and minimum values of ΔE for different error ranges $\left[E_{\mathrm{add}}^1, E_{\mathrm{add}}^N\right]$, and for $N = 3, 5, 7$. Each curve corresponds to a fixed value of E_{add}^1, and represents the value of ΔE under the condition of minimum or maximum performance imbalance (i.e., the minimum or maximum ΔE with respect to $E_{\mathrm{add}}^2, \ldots, E_{\mathrm{add}}^{N-1}$), for values of E_{add}^N ranging from E_{add}^1 to $E_{\mathrm{add}}^1 + 0.10$. Three different values of E_{add}^1 have been considered: 0.01, 0.05 and 0.10. From (39) and (43), and referring to Fig. 2, three characteristics of the behaviour of ΔE emerge. First and most notable is the fact that the maximum (and also the minimum) improvement achievable using WA in the place of

SA decreases as the size N of the classifier ensemble increases, over the same error range. Second, for the same N and E_{add}^{1}, the minimum and maximum values of ΔE increase for increasing values of E_{add}^{N}. This is reasonable, since the WA rule is affected to a lesser extent than the SA rule by the performance of the worst classifier, as shown in Section III-B.1. Third, for the same N and $E_{\mathrm{add}}^{N} - E_{\mathrm{add}}^{1}$, the minimum and maximum values of ΔE decrease as E_{add}^{1} and E_{add}^{N} increase. This indicates that, as the performances of the best and worst individual classifiers worsen by an identical amount, so the advantage of using the WA rule diminishes.

Summing up, the results obtained in this section allow to describe the relative behaviour of the two combining rules as a function of the size N of the classifier ensemble, of its error range $[E_{\mathrm{add}}^{1}, E_{\mathrm{add}}^{N}]$, and of the degree of performance imbalance. As two of the above three parameters are equal, the improvement of WA over SA increases as the size of the ensemble decreases, the expected added error of the worst classifier (and thus the error range width) increases, or the performance of individual classifiers approaches the condition of maximum performance imbalance. The conditions of maximum and minimum performance imbalance, and the behaviour of ΔE, are summarised respectively in Fig. 1 and Table 1 of the addendum to this paper.

Let us now consider the above results from a quantitative viewpoint, again referring to Fig. 2. As can be observed, the improvement of WA over SA is always fairly small. Note that the values of the error range considered in Fig. 2 are representative of most cases of practical interest. We considered expected added error of the best classifier of between 0.01 and 0.10, and a difference between the error probability of the worst and best classifiers (i.e. error range width) of between 0 and 0.10. For these values, Fig. 2 shows that ΔE never attains 0.02. Also, when five or more classifiers are combined, ΔE never exceeds 0.01. Moreover, even when $N < 5$, it can be seen that ΔE is greater than 0.01, only under the following conditions: E_{add}^{1} does not exceed 0.05, the error range width is greater than 0.06, and the condition of minimum performance imbalance is not verified.

To gain a better understanding of the behaviour of the WA rule, and to evaluate the improvement that can be achieved with respect to SA, under, as far as is possible, more realistic assumptions than those considered in this section, in the next section we extend our analysis by relaxing the assumption of uncorrelated estimation errors.

C. Unbiased and Correlated Estimation Errors

In the general case of biased and correlated estimation errors, we have seen that the optimal weights, and thus the corresponding $E_{\text{add}}^{\text{wa}}$, can only be computed by numerical analysis. In other words, no analytical investigation is possible. Nevertheless, in light of the analytical results obtained in Section III-B, concerning the behaviour of SA and WA versus the expected added errors of individual classifiers, it is interesting to verify, by means of numerical analysis, whether the same behaviour also holds when the errors are biased or correlated. To this end, we can rewrite the general expression of $E_{\text{add}}^{\text{wa}}$, given by (20), (22) and (23), by expliciting the E_{add}^m's, given by (5) and (7):

$$E_{\text{add}}^{\text{wa}} = \sum_{m=1}^{N} w_m^2 E_{\text{add}}^m + \sum_{m=1}^{N} \sum_{n \neq m} w_m w_n \left[\left(\beta_i^m - \beta_j^m \right) \left(\beta_i^n - \beta_j^n \right) + \rho_i^{mn} \sigma_i^m \sigma_i^n + \rho_j^{mn} \sigma_j^m \sigma_j^n \right] \ . \tag{44}$$

However, numerical analysis of the above expression is still impractical, as it involves too many parameters, besides the E_{add}^m's (that is, the biases, variances and correlations of the estimation errors). To simplify this analysis, we consider simply the case of unbiased errors, with class independent variances and correlations:

$$\beta_k^m = 0 \ \forall k, m, \quad (\sigma_k^m)^2 = (\sigma^m)^2 \ \forall k, m, \quad \rho_k^{mn} = \rho^{mn} \ \forall k, m, n \ . \tag{45}$$

In this case, it is easy to see that (44) can be rewritten as:

$$E_{\text{add}}^{\text{wa}} = \sum_{m=1}^{N} w_m^2 E_{\text{add}}^m + \sum_{m=1}^{N} \sum_{n \neq m} w_m w_n \rho^{mn} \sqrt{E_{\text{add}}^m E_{\text{add}}^n} \ . \tag{46}$$

In this way it is possible to verify, at least for small values of N, whether the conditions of maximum and minimum performance imbalance, found in Section III-B, also hold in the case of correlated errors. Let us consider classifier ensembles having the same size N, error range, and values of the correlation coefficients ρ^{mn}. Numerical analysis of (46) for $N = 3$ and $N = 5$, reveals that the maximum of ΔE is obtained when the expected added error of k ($1 \leq k < N$) individual classifiers is equal to E_{add}^N, that of the other $N - k$ classifiers being equal to E_{add}^1 (in this case, the value of k cannot be analytically determined). This means that the condition of maximum performance imbalance also holds in the case of correlated estimation errors, whatever the values of the correlation coefficients. On the contrary, we found that the condition of minimum performance imbalance does not hold for correlated classifiers. Moreover, the minimum of ΔE, for correlated errors, does not seem to exhibit a clear pattern of values of the E_{add}^m's.

Let us now consider the behaviour of SA and WA with respect to the correlation between estimation errors, for fixed values of all the E_{add}^m's. We already know that, in the most general case of biased and correlated errors, the smaller the correlation coefficients ρ^{mn}'s are, the smaller the $E_{\mathrm{add}}^{\mathrm{sa}}$ and $E_{\mathrm{add}}^{\mathrm{wa}}$. Obviously, this also holds true in the particular case represented by (46). On the other hand, a very interesting result emerges from the analysis of the behaviour of ΔE. First, if all the correlation coefficients are identical, and likewise the expected added errors of individual classifiers, then the optimal weights are $w_m = \frac{1}{N} \ \forall m$, i.e. $\Delta E = 0$ (note that this is the only case where the optimal weights related to (46) can be obtained analytically, using the Lagrange multiplier technique). Consider now fixed values of the E_{add}^m's (even non-identical), of N, and of the correlation range $[\rho^{\min}, \rho^{\max}]$ (i.e., $\rho^{\min} \leq \rho^{mn} \leq \rho^{\max} \ \forall m, n$). Numerical analysis for $N = 3$ and $N = 5$ shows that ΔE is maximum when p of the $\frac{N(N-1)}{2}$ correlation coefficients ρ^{mn} are equal to $\rho^{\max}$ (with $1 \leq p < \frac{N(N-1)}{2}$), while the other $\frac{N(N-1)}{2} - p$ are equal to $\rho^{\min}$ (the value of p cannot be analytically determined). Surprisingly, this condition on the correlation coefficients values is analogous to the condition of maximum performance imbalance. By analogy, we can define it as the condition of maximum correlation imbalance.

One notable consequence of the above results is that ΔE is greater than zero even when the individual classifiers exhibit identical expected added errors, but different correlation coefficients between the estimation errors. In this way we can extend the results obtained for unbiased and uncorrelated estimation errors as follows: SA is the optimal linear combining rule, only if the individual classifiers exhibit both identical expected added errors and identical correlations between the estimation errors.

The behaviour of ΔE with respect to the E_{add}^m's and to the ρ^{mn}'s can be better understood from Fig. 3, which refers to $N = 3$ (Fig. 3(a)), and $N = 5$ (Fig. 3(b)). Analogously to Fig. 2, the minimum (only for $N = 3$) and maximum of ΔE are shown, with respect to $E_{\mathrm{add}}^2, \ldots, E_{\mathrm{add}}^{N-1}$, and to the ρ^{mn}'s, for different error and correlation ranges. Fig. 3 shows that with respect to the E_{add}^m's, ΔE exhibits a similar behaviour to the case of uncorrelated estimation errors: ΔE increases for increasing error range width, and decreases for increasing N. Note that for correlated errors a similar trend has been observed only for $N = 3, 5$. However, the fact that it has been analytically proven to hold for each N for uncorrelated errors (see Section III-B.2) provides reasonable evidence that it holds for each N also for correlated errors.

With respect to the correlations, it can be seen that, for the same E_{add}^m's and correlation
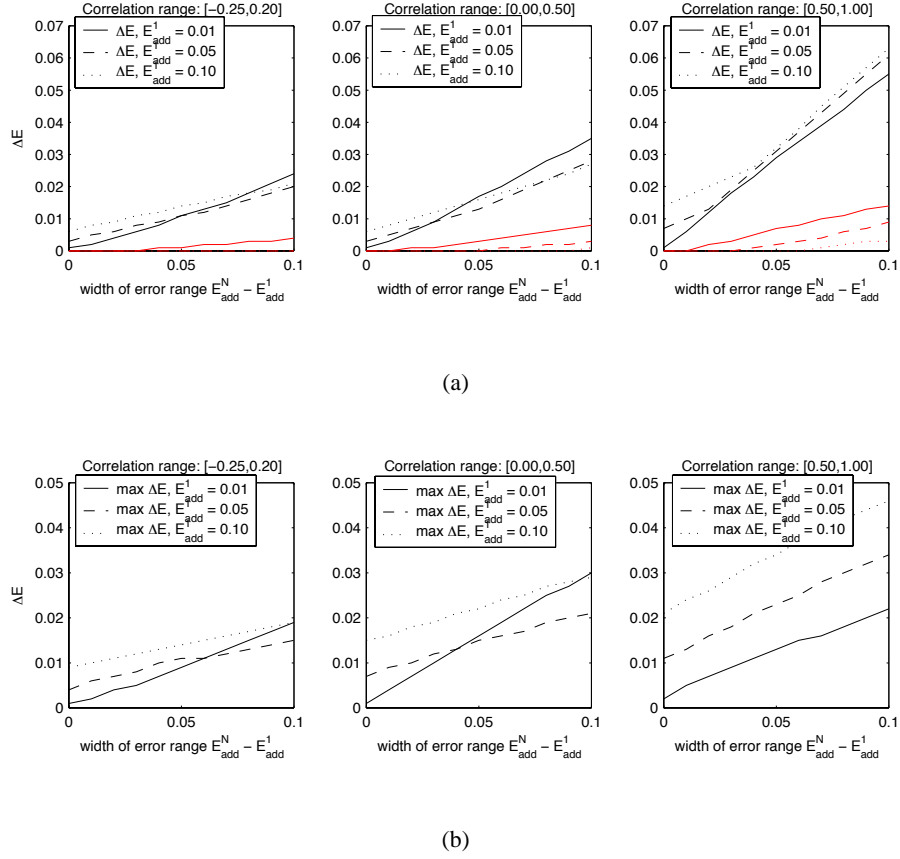
(a)

(b)

Fig. 3. Maximum values of $\Delta E = E_{\text{add}}^{\text{sa}} - E_{\text{add}}^{\text{wa}}$, versus error range width $E_{\text{add}}^N - E_{\text{add}}^1$, for $N = 3$ (3(a)) and $N = 5$ (3(b)), for $E_{\text{add}}^1 = 0.01, 0.05, 0.10$, and E_{add}^N ranging from E_{add}^1 to $E_{\text{add}}^1 + 0.10$. The maximum is computed with respect to $E_{\text{add}}^2, \ldots, E_{\text{add}}^{N-1}$, and to correlations ρ_k^{mn} for three different ranges. The minimum of ΔE is also shown, only for $N = 3$ (3(a)).

range width $\rho^{\max} - \rho^{\min}$, ΔE is greater for larger correlation values. Comparison of Fig. 3 and Fig. 2, also shows that, for non-identical correlations between the estimation errors of individual classifiers, WA outperforms SA to a greater extent than for uncorrelated errors. This indicates that, though a higher correlation diminishes the improvement achievable by both rules over individual classifiers, WA is affected to a lesser extent than SA. Note in particular that, as pointed out above, ΔE can be greater than zero even when $E_{\text{add}}^1 = E_{\text{add}}^N$, i.e., when all the individual classifiers exhibit identical expected added errors. For the correlation ranges considered in Fig. 3, the maximum improvement achievable by WA combining three classifiers with identical performances but different correlations, is about 0.015 (see Fig. 3(a), correlation range $[0.50, 1.00]$), which is very close to the maximum improvement achievable for different performances, for uncorrelated estimation errors (see Fig. 2).

Finally, Fig. 3(a) clearly shows that, for the same error correlation ranges, ΔE can take very different values, depending on the particular values of $E_{\mathrm{add}}^2, \ldots, E_{\mathrm{add}}^{N-1}$, and on the correlation coefficients. In other words, the value of ΔE depends strongly on the conditions of performance and correlation imbalance.

The conditions of maximum performance and correlation imbalance, and the behaviour of ΔE, are summarised respectively in Fig. 2 and Table 2 of the addendum to this paper.

In conclusion, we believe that the results of this section contribute to a better understanding of linear combiners for multiple classifier systems. Simple averaging is commonly believed to work well for classifiers with similar error rates, while weighted averaging is claimed to outperform simple averaging when classifiers exhibit different error rates. However, no previous work has analysed in detail the key role played by error correlations. The results of this section show that SA can be considered the optimal combiner only if the individual classifiers exhibit both identical error rates and identical correlations between the estimation errors. In particular, we have shown that WA is required to be used when pair-wise correlation coefficients are different, that is, the use of weights is necessary to counterbalance both the differences in classifiers' error rates and the correlation differences among classifier pairs.

D. Discussion

This section provides a critical discussion of the limitations of the assumptions under which some of the above theoretical results have been obtained. First, our analytical framework assumes that the estimation errors on the posterior probabilities of different classes are uncorrelated. As pointed out in Section II-A, this assumption is violated if the estimated posteriors are sum-to-one constrained. This happens, for instance, when the posterior probabilities are estimated using parametric methods, or the k-nearest neighbours classifier, but not when neural network classifiers are used. Second, the analytical framework does not allow to evaluate the overall added error of a classifier, but only that obtained on a single class boundary, which could result in underestimating the overall expected added error of the WA rule. The third limitation derives from the fact that the analytical results of Section III-B have been obtained under the assumption of unbiased and uncorrelated estimation errors, which is clearly unrealistic. However, it should be noted that these results are valid under slightly more general conditions. Indeed, as explained in Section III-A, these results are valid when the matrix $\mathbf{M}$ is diagonal, that is, from (22) and (23), when

$$\left(\beta_i^m - \beta_j^m\right)\left(\beta_i^n - \beta_j^n\right) + \left(\rho_i^{mn}\sigma_i^m\sigma_i^n + \rho_j^{mn}\sigma_j^m\sigma_j^n\right) = 0 \ \forall m, n \ . \tag{47}$$

The case of unbiased and uncorrelated errors is only a particular, though the most meaningful, case in which (47) holds, In addition, it should also be noted that some of these results were found to hold also for correlated estimation errors, though only for the case when variance and correlation are class independent (see Section III-C).

Considering the above limitations, it was interesting to evaluate the behaviour of the SA and WA rules experimentally, using real data, where such assumptions are likely to be violated, so as to investigate the extent to which their behaviour agrees with the predictions of our theoretical model, and to experimentally compare SA and WA.

IV. Experimental Results

The experiments presented in this section have been conducted on three real data sets: Feltwell, Letter, and Pendigits. The Feltwell data set consists of a set of multisensor remote-sensing images for an agricultural area near the village of Feltwell (UK) [23]. From a section $(250 \times 350$ pixels) of a scene acquired by an optical sensor and a radar sensor, 10,944 pixels belonging to five agricultural classes (i.e., sugar beet, stubble, bare soil, potatoes, and carrots) were selected, and divided into a training set and a test set of 5,124 and 5,820 pixels respectively. Each pixel is characterised by a fifteen-element feature vector containing the grey-level brightness values in six optical bands and over nine radar channels. The Letter and Pendigits data sets were taken from the UCI machine learning repository.[4] Letter consists of 20,000 images of the 26 printed capital letters of the English alphabet, each characterised by sixteen numeric features. We used the first 15,000 characters as training set, and the remaining 5,000 as the test set. Pendigits consists of 10,992 images of the ten handwritten digits, characterised by sixteen numeric features, and divided into a training set of 7,494 images, and a test set of 3,498.

The aim of the experiments was to investigate to what extent the behaviour of the SA and WA rules on real data agreed with the predictions of the theoretical model. In the experiments, we focused on the behaviour of the two combining rules as a function of the performances of individual classifiers, and of the correlations between their outputs,[5] the quantities considered in our theoretical analysis (Section III). Since, in an experimental setting, the performance

---

[4] http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] Note that the correlation coefficient $\rho_k^{mn}(x)$ between the estimation errors $\epsilon_k^m(x)$ and $\epsilon_k^n(x)$ of classifiers m and n on the posterior probability of the k-th class $P(\omega_k|x)$, for a pattern x, is equal to the correlation between the outputs $f_k^m(x)$ and $f_k^n(x)$ of the classifiers, since both outputs approximate the same quantity $P(\omega_k|x)$.

of individual classifiers can be controlled much more easily than correlations, we chose to construct several classifier ensembles having different error ranges, and different conditions of performance imbalance for the same error range. For this purpose, we trained several MLPs with a varying number of hidden neurons and of training epochs, and random values of the initial weights, using the standard backpropagation algorithm, until a range of test set error rates of width 0.15 was obtained for each data set. Next we constructed sixteen ensembles of three MLPs each, by selecting the MLPs with error rates nearest to predefined values. Denoting the error rates of the three classifiers of each ensemble with E_1, E_2 and E_3 (with $E_1 < E_2 < E_3$), the predefined values were chosen as follows. First, we constructed four ensembles of classifiers with identical error rates $E_1 = E_2 = E_3$, that we shall denote as "balanced ensembles"; the predefined error rates are shown in rows 1–4 of Table I. Next we constructed three pairs of ensembles having three different error ranges of the same width 0.05 (Table I, rows 5–10). The two ensembles of each pair have the same error range, but two different conditions of performance imbalance: one ensemble is characterised by the maximum performance imbalance,[6] i.e. $E_2 = E_3$ (rows $5, 7$ and 9), the other by $E_2 = E_1$ (rows $6, 8$ and 10). Finally, we constructed two groups of ensembles having two different error ranges, of the same width 0.10 (rows 11–16). Each group consists of three ensembles with the same error range, and three different conditions of performance imbalance: one ensemble is characterised by the maximum performance imbalance, i.e. $E_2 = E_3$ (rows 11 and 14), another by $E_2 = E_1$ (rows 13 and 16), and the third by a value of E_2 equidistant to E_1 and E_3 (rows 12 and 15), which should be close to the condition of minimum performance imbalance (38).[7]

We should point out that in these experiments we were interested in assessing the *ideal* performance of WA, i.e. the performance obtained using the optimal weights for the test set, and comparing it with that predicted by the theoretical model described in Section III. As it was not our intention to explore the effects of weight estimation, we did not estimate the weights from a validation set, using one of the methods proposed in the literature. Instead, for each classifier ensemble, we computed the optimal weights by performing an exhaustive search on the test set. A discretisation step of 0.01 was used to reduce computational

---

[6]We recall that, for ensembles of three classifiers, the theoretical condition of maximum performance imbalance found in Section III-B.2 was $E_{\mathrm{add}}^2 = E_{\mathrm{add}}^3$.

[7]It is not possible to find the exact value of the error rate E_2 corresponding to (38), since the Bayes error of the real data sets considered in the experiments is unknown.

TABLE I

PREDEFINED VALUES OF ERROR RATES FOR THE THREE CLASSIFIERS (E_1,E_2,E_3) OF EACH OF THE SIXTEEN

ENSEMBLES CONSTRUCTED FOR THE EXPERIMENTS.

	Feltwell, Pendigits			Letter		
	E_1	E_2	E_3	E_1	E_2	E_3
1	0.10	0.10	0.10	0.15	0.15	0.15
2	0.15	0.15	0.15	0.20	0.20	0.20
3	0.20	0.20	0.20	0.25	0.25	0.25
4	0.25	0.25	0.25	0.30	0.30	0.30
5	0.10	0.15	0.15	0.15	0.20	0.20
6	0.10	0.10	0.15	0.15	0.15	0.20
7	0.15	0.20	0.20	0.20	0.25	0.25
8	0.15	0.15	0.20	0.20	0.20	0.25
9	0.20	0.25	0.25	0.25	0.30	0.30
10	0.20	0.20	0.25	0.25	0.25	0.30
11	0.10	0.20	0.20	0.15	0.25	0.25
12	0.10	0.15	0.20	0.15	0.20	0.25
13	0.10	0.10	0.20	0.15	0.15	0.25
14	0.15	0.25	0.25	0.20	0.30	0.30
15	0.15	0.20	0.25	0.20	0.25	0.30
16	0.15	0.15	0.25	0.20	0.20	0.30

complexity. While this method was suitable for the purposes of our experiments, in real applications an exhaustive search (on a validation set) is clearly unfeasible.

The experiments were repeated ten times, by training the same three MLPs of each ensemble on ten different training sets, composed by randomly extracting 60% of the patterns of the original training sets. Tables II–IV show the test set error rates of the SA and WA rules, averaged over the ten runs of the experiments. The average correlation ρ^{mn} between the outputs of each pair of MLPs ($m, n = 1, 2, 3, \ m \neq n$) is also given. More precisely, we computed the correlation coefficient $\rho_k^{mn}(x)$ between the outputs $f_k^m(x)$ and $f_k^n(x)$ over the ten test runs, for each test pattern x, and for each class k. The value of ρ^{mn} was obtained as the average of the $\rho_k^{mn}(x)$ over all classes and test patterns. In the following, we will denote the error rates of SA and WA respectively with E_{sa} and E_{wa}, and their difference with ΔE.

Let us first compare the qualitative behaviour of the two combining rules for the three data sets, with that predicted by the theoretical model of Section III. As expected, since the optimal weights were used, WA always outperformed SA, i.e. $\Delta E > 0$ for all the ensembles considered. Nevertheless, it is worth noting that the SA rule always gave a lower misclassification probability than the worst classifier of the ensemble. Looking at Tables II–IV, we can see that the experimental results agree with the theoretical predictions on four main points. The first one concerns the behaviour of E_{sa} and E_{wa}, while the other

TABLE II

RESULTS FOR THE FELTWELL DATA SET. TEST SET ERROR RATES OF THE THREE INDIVIDUAL CLASSIFIERS (E_1, E_2 AND E_3), AND OF SA AND WA (E_{sa} AND E_{wa}), FOR EACH OF THE SIXTEEN ENSEMBLES CONSIDERED, LISTED IN THE SAME ORDER AS IN TABLE I. COLUMN ΔE SHOWS THE VALUES OF $E_{\text{sa}} - E_{\text{wa}}$. ALL REPORTED VALUES ARE AVERAGED OVER TEN REPETITIONS. STANDARD DEVIATION IS SHOWN IN BRACKETS. AVERAGE CORRELATION BETWEEN OUTPUTS OF EACH PAIR OF CLASSIFIERS IS ALSO SHOWN (ρ^{mn}, $m, n = 1, 2, 3$, $m \neq n$).

	E_1	E_2	E_3	ρ^{12}	ρ^{13}	ρ^{23}	E_{sa}	E_{wa}	ΔE
1	0.099 (0.007)	0.109 (0.009)	0.102 (0.006)	0.36	0.43	0.24	0.098 (0.005)	0.095 (0.005)	0.004 (0.002)
2	0.148 (0.009)	0.144 (0.007)	0.130 (0.012)	0.31	0.31	0.34	0.118 (0.007)	0.111 (0.007)	0.007 (0.005)
3	0.190 (0.006)	0.194 (0.008)	0.193 (0.010)	0.93	0.88	0.82	0.191 (0.008)	0.185 (0.008)	0.005 (0.005)
4	0.251 (0.017)	0.242 (0.010)	0.242 (0.010)	0.38	0.43	0.77	0.237 (0.010)	0.231 (0.010)	0.006 (0.004)
5	0.099 (0.007)	0.144 (0.007)	0.130 (0.012)	0.38	0.45	0.34	0.112 (0.006)	0.096 (0.006)	0.016 (0.005)
6	0.099 (0.007)	0.109 (0.009)	0.130 (0.012)	0.36	0.45	0.23	0.105 (0.007)	0.097 (0.005)	0.009 (0.004)
7	0.148 (0.009)	0.194 (0.008)	0.193 (0.010)	0.28	0.31	0.82	0.158 (0.006)	0.130 (0.011)	0.029 (0.012)
8	0.148 (0.009)	0.144 (0.007)	0.193 (0.010)	0.31	0.31	0.35	0.142 (0.006)	0.127 (0.011)	0.015 (0.008)
9	0.190 (0.006)	0.242 (0.010)	0.242 (0.010)	0.08	0.14	0.77	0.211 (0.012)	0.176 (0.007)	0.035 (0.008)
10	0.190 (0.006)	0.194 (0.008)	0.242 (0.010)	0.93	0.14	0.16	0.179 (0.010)	0.173 (0.008)	0.007 (0.003)
11	0.099 (0.007)	0.194 (0.008)	0.193 (0.010)	0.23	0.31	0.82	0.147 (0.007)	0.096 (0.005)	0.052 (0.007)
12	0.099 (0.007)	0.144 (0.007)	0.193 (0.010)	0.38	0.31	0.35	0.123 (0.007)	0.096 (0.005)	0.027 (0.007)
13	0.099 (0.007)	0.109 (0.009)	0.193 (0.010)	0.36	0.31	0.11	0.108 (0.007)	0.095 (0.004)	0.013 (0.006)
14	0.148 (0.009)	0.242 (0.010)	0.242 (0.010)	0.10	0.13	0.77	0.215 (0.008)	0.146 (0.009)	0.069 (0.010)
15	0.148 (0.009)	0.194 (0.008)	0.242 (0.010)	0.28	0.13	0.16	0.164 (0.010)	0.133 (0.011)	0.031 (0.013)
16	0.148 (0.009)	0.144 (0.007)	0.242 (0.010)	0.31	0.13	0.14	0.161 (0.011)	0.127 (0.010)	0.034 (0.011)

three points concern the improvement achievable by WA with respect to SA, ΔE. First, let us examine the five groups of ensembles having the same error range $[E_1, E_3]$ (i.e., ensembles (5,6), (7,8), (9,10), (11,12,13) and (14,15,16)). As predicted in Section III-B.1, E_{sa} and E_{wa} increase for increasing E_2, with the only exceptions of E_{wa} for ensembles (5,6) of Table II, and both E_{sa} and E_{wa} for ensembles (14,15) of Table IV. Second, for the balanced ensembles 1–4, the values of ΔE are smaller than those obtained for the imbalanced ensembles 5–16, with the only exceptions of ensemble 10 of Table II, and ensembles (9,10,16) of Table IV. In other words, the improvement obtained with WA was almost always lower for ensembles of classifiers with similar performances. Third, looking again at the five groups of ensembles with the same error range $[E_1, E_3]$, it is easy to see that, with the exception of ensembles (14,15,16) for Pendigits, the maximum of ΔE is obtained when $E_2 = E_3$, which corresponds to the condition of maximum performance imbalance described in Sections III-B.2 for ensembles of three classifiers. Finally, consider the ensembles having the same value of E_1 and E_2, and increasing values of E_3 (for instance, ensembles (2,8,16) and (3,10)). Except for ensembles (8,16) and (7,15) for Pendigits, we can see that ΔE increases for increasing E_3 (that is, increasing width of the error range), in agreement with the theoretical

TABLE III

RESULTS FOR LETTER DATA SET. SEE CAPTION TABLE 2 FOR DETAILS.

	E_1	E_2	E_3	ρ^{12}	ρ^{13}	ρ^{23}	E_{sa}	E_{wa}	ΔE
1	0.149 (0.007)	0.149 (0.004)	0.152 (0.006)	0.15	0.15	0.15	0.116 (0.004)	0.115 (0.004)	0.001 (0.001)
2	0.197 (0.004)	0.200 (0.007)	0.202 (0.008)	0.12	0.11	0.11	0.157 (0.005)	0.156 (0.005)	0.002 (0.001)
3	0.250 (0.011)	0.251 (0.004)	0.248 (0.009)	0.04	0.06	0.05	0.202 (0.006)	0.200 (0.006)	0.002 (0.002)
4	0.295 (0.014)	0.301 (0.010)	0.296 (0.015)	0.04	0.01	0.03	0.244 (0.004)	0.241 (0.004)	0.002 (0.001)
5	0.149 (0.007)	0.200 (0.007)	0.202 (0.008)	0.13	0.13	0.11	0.142 (0.006)	0.136 (0.006)	0.006 (0.002)
6	0.149 (0.007)	0.149 (0.004)	0.202 (0.008)	0.15	0.13	0.12	0.126 (0.005)	0.122 (0.005)	0.004 (0.001)
7	0.197 (0.004)	0.251 (0.004)	0.248 (0.009)	0.10	0.09	0.05	0.180 (0.003)	0.175 (0.003)	0.005 (0.002)
8	0.197 (0.004)	0.200 (0.007)	0.248 (0.009)	0.12	0.09	0.08	0.166 (0.005)	0.162 (0.005)	0.004 (0.001)
9	0.298 (0.152)	0.301 (0.010)	0.296 (0.015)	0.04	0.04	0.03	0.231 (0.008)	0.226 (0.009)	0.005 (0.003)
10	0.298 (0.152)	0.274 (0.072)	0.296 (0.015)	0.04	0.04	0.05	0.216 (0.006)	0.213 (0.006)	0.004 (0.002)
11	0.149 (0.007)	0.251 (0.004)	0.248 (0.009)	0.10	0.11	0.05	0.158 (0.006)	0.144 (0.007)	0.014 (0.004)
12	0.149 (0.007)	0.200 (0.007)	0.248 (0.009)	0.13	0.11	0.08	0.149 (0.007)	0.139 (0.006)	0.009 (0.002)
13	0.149 (0.007)	0.149 (0.004)	0.248 (0.009)	0.15	0.11	0.12	0.130 (0.005)	0.124 (0.004)	0.006 (0.001)
14	0.197 (0.004)	0.301 (0.010)	0.296 (0.015)	0.08	0.07	0.03	0.202 (0.005)	0.187 (0.004)	0.015 (0.002)
15	0.197 (0.004)	0.274 (0.072)	0.296 (0.015)	0.10	0.07	0.05	0.191 (0.004)	0.181 (0.004)	0.010 (0.002)
16	0.197 (0.004)	0.200 (0.007)	0.296 (0.015)	0.12	0.07	0.07	0.174 (0.006)	0.165 (0.005)	0.009 (0.002)

results of Sections III-B.2 and III-C (see Figs. 2 and 3).

Let us now consider the behaviour of the two combining rules from a quantitative viewpoint. As pointed out above, the lower values of ΔE were almost always obtained for the ensembles 1–4, containing classifiers exhibiting similar performances. Tables II–IV show that, for these ensembles, the values of ΔE are fairly small, ranging from 0.004 to 0.007 for Feltwell, from 0.001 to 0.002 for Letter, and from 0.002 to 0.006 for Pendigits. Higher values were obtained for the imbalanced ensembles 5–16. In particular, the maximum values of ΔE were obtained for ensembles with the greatest error range width (0.10): ΔE attained 0.069 for Feltwell, 0.015 for Letter, and 0.031 for Pendigits. However, for all imbalanced ensembles with identical error range $[E_1, E_3]$, the value of ΔE depends strongly on the value of E_2, i.e. on what we called the condition of "performance imbalance" in Section III-B.2. For instance, consider the three imbalanced ensembles 11–13 with the same error range $[0.10, 0.20]$, for Feltwell (Table II) and Pendigits (Table IV): as the value of E_2 decreases from 0.20 to 0.10, so too does the value of ΔE from 0.031 to 0.011 (Feltwell) and from 0.052 to 0.013 (Pendigits). This means that the improvement achievable using WA may be minor even for ensembles with a large error range width.

Lastly we will consider the correlation between classifier outputs. In Section II we showed that, according to the theoretical model, the lower the correlation is, the better the performance of the SA rule. The results of Tables II–IV agree with this prediction. Table III, shows that very low correlation values were found for the Letter data set, ranging from 0.01 to 0.15.

TABLE IV

RESULTS FOR PENDIGITS DATA SET. SEE CAPTION TABLE 2 FOR DETAILS.

	E_1	E_2	E_3	ρ^{12}	ρ^{13}	ρ^{23}	E_{sa}	E_{wa}	ΔE
1	0.099 (0.003)	0.102 (0.003)	0.103 (0.002)	0.27	0.31	0.31	0.098 (0.002)	0.095 (0.002)	0.002 (0.001)
2	0.149 (0.002)	0.150 (0.002)	0.151 (0.008)	0.29	0.17	0.18	0.147 (0.003)	0.143 (0.003)	0.004 (0.002)
3	0.199 (0.005)	0.202 (0.005)	0.198 (0.004)	0.06	0.21	0.16	0.180 (0.004)	0.176 (0.003)	0.004 (0.002)
4	0.248 (0.028)	0.252 (0.021)	0.252 (0.052)	0.05	0.19	-0.03	0.184 (0.016)	0.179 (0.016)	0.006 (0.005)
5	0.099 (0.003)	0.150 (0.002)	0.151 (0.008)	0.27	0.19	0.18	0.123 (0.004)	0.099 (0.003)	0.024 (0.005)
6	0.099 (0.003)	0.102 (0.003)	0.151 (0.008)	0.27	0.19	0.17	0.106 (0.002)	0.096 (0.002)	0.009 (0.003)
7	0.149 (0.002)	0.202 (0.005)	0.198 (0.004)	0.17	0.15	0.16	0.167 (0.002)	0.149 (0.002)	0.019 (0.002)
8	0.149 (0.002)	0.150 (0.002)	0.198 (0.004)	0.29	0.15	0.17	0.155 (0.001)	0.145 (0.002)	0.010 (0.002)
9	0.199 (0.005)	0.252 (0.021)	0.252 (0.052)	0.09	0.07	-0.03	0.179 (0.009)	0.173 (0.010)	0.006 (0.003)
10	0.199 (0.005)	0.202 (0.005)	0.252 (0.052)	0.06	0.07	0.07	0.175 (0.004)	0.170 (0.004)	0.005 (0.001)
11	0.099 (0.003)	0.202 (0.005)	0.198 (0.004)	0.14	0.13	0.16	0.131 (0.003)	0.099 (0.003)	0.031 (0.004)
12	0.099 (0.003)	0.150 (0.002)	0.198 (0.004)	0.27	0.13	0.17	0.127 (0.002)	0.099 (0.003)	0.027 (0.004)
13	0.099 (0.003)	0.102 (0.003)	0.198 (0.004)	0.27	0.13	0.16	0.108 (0.002)	0.096 (0.002)	0.011 (0.003)
14	0.149 (0.002)	0.252 (0.021)	0.252 (0.052)	0.14	0.08	-0.03	0.159 (0.008)	0.147 (0.003)	0.012 (0.006)
15	0.149 (0.002)	0.202 (0.005)	0.252 (0.052)	0.17	0.08	0.07	0.162 (0.005)	0.148 (0.003)	0.014 (0.003)
16	0.149 (0.002)	0.150 (0.002)	0.252 (0.052)	0.29	0.08	0.07	0.151 (0.004)	0.145 (0.003)	0.006 (0.003)

For this data set, the SA rule outperforms the best individual classifier (i.e., $E_{\text{sa}} < E_1$) on fourteen out of sixteen ensembles, even for ensembles with a large error range width. Slightly higher correlation values (up to 0.31) were observed for Pendigits. In this case, $E_{\text{sa}} < E_1$ on six out of sixteen ensembles. The largest correlation values were observed for the Feltwell data set (up to 0.93), where it can be seen that $E_{\text{sa}} < E_1$ only on five ensembles. Moreover, in Section III-C we showed that, theoretically, the improvement achievable by WA over SA increases for increasing values of the correlation between estimation errors. This behaviour also emerges from the above results: the values of ΔE, across the sixteen classifier ensembles, do not exceed 0.015 for Letter, but attain 0.031 for Pendigits, and 0.069 for Feltwell.

Summing up, the qualitative behaviour of the two combining rules, versus the performance of individual classifiers, and the correlation between their outputs, was found to agree with the predictions of the theoretical model, for the real data sets tested, in spite of the fact that the model is based on strict assumptions. Moreover, the ideal improvement achievable by WA over SA was often found to be quite small.

V. CONTRIBUTION TO THE STATE OF THE ART

In this section, we discuss the contribution to the state of the art of our theoretical and experimental analysis of linear combiners. We review the contribution provided by our analysis of SA and WA (sections V-A and V-B), and by the analytical and numerical comparison of these two combiners (section V-C). In particular, we discuss how the results

presented in the previous sections improve our understanding of the operation of linear combiners, and their practical relevance in the design of linearly combined multiple classifiers.

A. Simple Averaging

As pointed out in Section II, the seminal work by Tumer and Ghosh analysed the expected added error of simple averaging assuming estimation errors with identical variances [30], [31]. This implies that their results do not provide information on the expected added error of SA for classifiers with different variances, that is, classifiers which perform differently on unseen test data. In Section II-B, we showed that the results obtained by Tumer and Ghosh also hold for the more general case of estimation errors with non-identical variances. The contribution to the state of the art of this result emerges quite clearly. As pointed out in Section II-A, SA allows to handle the bias-variance trade-off by designing classifiers with low bias and low correlations, as the variance can be reduced by averaging classifiers' outputs. However, this kind of design often produces classifiers with different variances. The result of Section II-B guarantees that the variance can still be reduced by averaging classifiers' outputs. In Section II-B, we also showed that the expected added error of the SA is never larger than the maximum expected added error of individual classifiers. The contribution to the state of the art and the practical relevance of this result becomes apparent when one considers that one of the main reasons for combining classifiers is to avoid the "worst" case of the traditional evaluation and selection approach to classifier design. In fact, selecting the apparent best classifier from a given ensemble, on the basis of validation data, involves the risk of obtaining the worst classifier on unseen test data [6], [3]. Our result provides a theoretical support to the experimental evidence that simple averaging can guard the designer against this worst case, as the test set error of SA is theoretically guaranteed to be smaller than that of the worst classifier in the ensemble. For the case of unbiased and uncorrelated estimation errors, Tumer and Ghosh quantified the error reduction achievable by SA under the assumption of classifiers with identical expected added errors [30], [31]. As this assumption is likely to be violated in real cases, in section III-B.1, we provided formulas for quantifying the error reduction in the general case of classifiers with different errors. In particular, we found the conditions on the error rates of the individual classifiers which determine the maximum and the minimum error reduction. Finally, so far SA has been believed to work well for classifiers with similar error rates, in spite of correlations among the classifiers. No previous work has analysed the effect of different classifiers' correlations

on the performance of SA, though, in real cases, classifiers often exhibit different pair-wise correlations. As an example, in multi-sensor applications, classifiers using data from different sensors are usually uncorrelated to a far greater degree than classifiers which use data from the same sensor. Moreover, it should be noted that experimental results reported in previous works have demonstrated that SA may perform differently for ensembles containing classifiers with similar error rates but different pair-wise correlations [24], [25]. In section III-C, we analysed the effect of classifiers' correlations on the performance of SA. In particular, we showed that SA is the optimal combiner only if the individual classifiers exhibit both identical error rates and identical correlations between the estimation errors. Therefore, the design of classifiers to be combined by SA should also take into account the differences among the pair-wise correlations exhibited by classifiers, and not simply, as believed until now, the differences in error rates. It is easy to see the relevance that this result could have for popular methods, such as Bagging, which use the SA combiner [5]. For example, the result of section III-C suggests that classifiers generated by Bagging should exhibit similar error rates and similar correlations between the estimation errors, in order to guarantee good performance of this ensemble learning method. Moreover, this result could also be exploited to stop the generation of new classifiers by Bagging, when classifiers with different error rates and correlations are being produced, or, after their creation, to select the classifiers with the most similar error rates and pair-wise correlations.

B. Weighted Averaging

As pointed out in Section I, previous works have only addressed the SA rule, and to date no theoretical analysis of the WA rule has been performed, the sole exception being [1], where the author extended some results of [31] to the WA rule, limitedly to the simplest case of unbiased, uncorrelated and identically distributed estimation errors. In sections III-A and III-B, we derived the general expression of the expected added error of WA, and then, for the case of unbiased and uncorrelated errors, we analytically determined the optimal weights, which were found to be inversely proportional to the added errors of the individual classifiers. It should be noted that this result suggests a simple method, albeit obtained under very strict assumptions, for computing the optimal weights. In section III-B.1, we found the conditions on the error rates of the individual classifiers which determine the maximum and minimum error reduction achievable by WA. Moreover, we showed that, when a new classifier is added to a given ensemble, the theoretical performance of the WA rule is always enhanced, whatever

the expected added error of the new classifier. The latter result has clear relevance for the design of multiple classifier systems.

C. *Simple versus Weighted Averaging*

As discussed in Section I, because of the lack of guidelines with clear theoretical foundations, the choice between the use of SA and WA is currently based on experimentally derived heuristic rules. Simple averaging is commonly believed to work well for classifiers with similar error rates, while weighted averaging is claimed to outperform simple averaging when classifiers exhibit different error rates. However, experimental results do not completely support these rules of thumb [24], [25], and the conditions under which WA can significantly outperform SA are not clear. In sections III-B.2 and III-C, our analytical and numerical comparison of the SA and WA has improved the understanding of these two combiners and provided new guidelines for the practical choice between SA and WA. First, we showed that SA is the optimal linear combining rule only if the individual classifiers exhibit identical error rates and identical correlations between estimation errors. Second, we showed that, when the optimal weights are used, the improvement achievable by the WA over the SA increases, all other factors being equal, as any of the following conditions hold: the size of the classifier ensemble decreases, the width of the error range increases, the correlation between estimation errors increase, and the performance of individual classifiers, or the correlations between the estimation errors, approach the condition of maximum "imbalance" defined in Sections III-B.2 and III-C. All the above guidelines are supported by the experimental evidence reported in Section IV for real data sets. Before discussing the practical relevance of our guidelines to the design of multiple classifier systems, we recall that previous works on linear combiners for classifier fusion, and for regression, have focused on weight estimation methods. Instead, the theoretical analysis presented in this paper addresses a different issue, namely, what *ideal* improvement (i.e., when the optimal weights are used) can be achieved when WA is used in the place of SA in classification problems, and how such improvement depends on the performance of individual classifiers and on the correlation between their outputs. Obviously, in the ideal case, the performance of WA can never be worse than SA, but, in real cases, it can, on account of the fact that the weights are estimated from validation data. It is therefore interesting to evaluate what improvement can ideally be achieved using WA, as SA is to be preferred when only a minor improvement can be achieved. In fact, small improvements can be negated by the weight estimation issue, in which case WA may perform worse than SA.

Besides providing a better understanding of these two combining rules, the theoretical and experimental results described in Sections III and IV also suggest some practical guidelines, which will assist the designer in choosing between the two. In the following, we sum up the results pertinent to this specific aim. First, we found that, for classifiers exhibiting similar performances, the ideal improvement in misclassification probability achievable using WA over SA is fairly small. In our experiments, it was always below 0.01. The improvement increases as the difference between the performance of individual classifiers (i.e., the width of the error range) increases. However, for identical width of error range, the experiments showed that the improvement depends strongly on how the error rates of individual classifiers are distributed within the error range (i.e., on what we called the condition of performance imbalance): we found that the ideal improvement could be small even when the width of the error range is as large as 0.10. Moreover, it should be noted that the condition of maximum performance imbalance is known only for ensembles of three classifiers, as explained in Section III-B.2. This means that, even if the error rates of individual classifiers can be reliably estimated, it is not possible to know whether they are close to the condition corresponding to the maximum ideal improvement achievable by the WA, for the given error range, unless only three classifiers are combined. The theoretical results also showed that, all other factors being equal, the ideal improvement decreases as the size of the classifier ensemble increases (although no tests have been conducted with more than three classifiers, because of the computational complexity involved in the exhaustive search for the optimal weights). Finally, we found that, while the performance of linear combiners improves as the correlation between classifier outputs decreases, at the same time the improvement achievable by WA decreases.

Bearing in mind that weight estimation can worsen the performance of WA with respect to the ideal case, our results suggest that when WA is implemented by assigning one positive weight to each classifier, its use is recommendable only in particular cases. In particular, WA can be expected to perform significantly better than SA (provided that suitable validation data are available for weight estimation) only for small classifier ensembles, if the individual classifiers exhibit a range of error rates with non-negligible width (say, at least 0.05), and if the outputs of the individual classifiers are highly correlated. Otherwise, the SA combining rule seems to be a valid alternative, from the viewpoint of both computational complexity, since no training data are required, and of the achievable performance. Concerning this point, it is worth recalling that, as pointed out in Section IV, in our experiments SA always outperformed the worst classifier of the ensemble.

The above results appear to confirm that the implementation of WA considered in this work (one positive weight for each classifier) is not sufficiently more flexible than SA, as argued in [31]. In principle, performance can be enhanced using weights which are unrestricted in sign (for instance, using a neural network as a trained combiner), or using more flexible implementations like those proposed in [2], [32], where different weights are assigned to each classifier and to each class. However, it should be taken into account that a greater quantity of validation data is required for a reliable estimate of a larger number of weights.

D. Directions for future work

As pointed out in Section IV, reported results agree with the main qualitative predictions of the theoretical model, even if some of our assumptions, such as unbiasedness of estimation errors, are likely to be violated in real cases. We believe that the observed agreement between experimental results and the theoretical predictions is due to the fact that the extent to which our assumptions are violated in the considered real data sets does not affect the predictive capability of the model up to invalidate it. In addition, one should note that our predictions are only qualitative, which make them more robust to violations of model assumptions. A possible way to verify whether this explanation is correct would be to extend the analysis of our model by relaxing some of the assumptions (for instance, considering biased estimation errors, or the different sources of added error described in [21]). This could allow to see how the theoretical predictions change, and investigate some of the conditions which invalidate our predictions. However, as explained in Section III-C, this step is quite difficult, since it requires to resort to a complex numerical analysis. Another way to assess further the predictive capability of our theoretical model is obviously to extend the experiments to a larger number of data sets, and then to analyse carefully the characteristics of the data sets for which the qualitative predictions of our model strongly fail. Both these tasks should be considered as a natural and necessary follow up of this work, which obviously is not intended to have the last word on the analysis of linear combiners.

VI. Conclusions

In this paper, we have presented a theoretical and experimental analysis of linear combiners for classifier fusion. To this end, we extended the scope of the analytical framework developed by Tumer and Ghosh [30], [31], in order to treat WA and draw comparisons with SA. In our analysis, we considered the simplest and most widely used implementation of WA, where

one non-negative weight is assigned to each individual classifier. Analytical results have been obtained for unbiased and uncorrelated, but not identically distributed, estimation errors. Some results have been extended to the case of correlated errors, through a numerical analysis. These results show how the expected added error, over Bayes error, of the SA combining rule, and that which can be achieved adopting the WA combining rule when optimal weights are used, depends on the expected added errors of individual classifiers, and on the correlation between their estimation errors. In particular, we showed that SA is the optimal linearly combining rule, only if the individual classifiers exhibit both identical performances, and identical correlations between estimation errors. Otherwise, WA can provide better results. However, our theoretical and experimental analysis indicated that the improvement which can be achieved by WA over SA is smaller than one would expect. The improvement was found to increase with increasing width of the error range exhibited by the classifier ensemble, and also to depend on the manner in which their misclassification probabilities are distributed within the error range, but in our experiments, the improvement was always lower than 0.01. Moreover, we showed that, while a low correlation between estimation errors is beneficial for a linear combiner, the improvement achievable by WA over SA, for individual classifiers with identical performance, actually diminishes. Therefore, our theoretical and experimental results suggest that the practical use of WA (implemented with one positive weight for each classifier) is to be recommended only in the special cases discussed in Section V-C. Finally, it should be noted that, although our theoretical results are based on strict assumptions, they are confirmed by the experiments carried out on three real data sets where such assumptions are likely to be violated. These results suggest some interesting directions for future work on this subject, as explained in Sections V-A and V-D. As discussed in Section V, we believe that this theoretical and experimental analysis of linear combiners provides an important contribution to the state of the art of multiple classifier systems, as it provides a deeper insight into these classifier fusion rules and some well grounded guidelines for their practical use.

REFERENCES

[1] L.A. Alexandre, A.C. Campilho, and M. Kamel, "Combining Independent and Unbiased Classifiers Using Weighted Average," *Int'l Conf. Pattern Recognition*, pp. 495-498, 2000.

[2] J.A. Benediktsson, J.R. Sveinsson, O.K. Ersoy, and P.H. Swain, "Parallel Consensual Neural Networks," *IEEE Trans. Neural Networks*, vol. 8, pp. 54-64, 1997.

[3] C.M. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[4] L. Breiman, "Stacked Regressions," *Machine Learning*, vol. 24, pp. 49-64, 1996.

[5] L.Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.

[6] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed., Wiley & Sons, 2000.

[7] Y. Freund, and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Computer and System Sciences*, vol. 55, pp. 119-139, 1999.

[8] G. Fumera, and F. Roli, "Performance analysis and comparison of linear combiners for classifier fusion," *Proc. Int'l Workshop Statistical Pattern Recognition*, pp. 424-432, LNCS vol. 2396, Springer, 2002.

[9] G. Fumera, and F. Roli, "Linear Combiners for Classifier Fusion: Some Theoretical and Experimental Results," *Proc. Int'l Workshop Multiple Classifier Systems*, pp. 74-83, LNCS vol. 2709, Springer, 2003.

[10] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, pp. 1-58, 1992.

[11] S. Hashem, "Optimal Linear Combination of Neural Networks," PhD dissertation, Purdue University, 1993.

[12] S. Hashem, and B. Schmeiser, "Improving model accuracy using optimal linear combinations of trained neural networks," *IEEE Trans. Neural Networks*, vol. 6, pp. 792-794, 1995.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[14] T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66-75, 1994.

[15] T.K. Ho, "The random subspace method for constructing decision forests", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832-844, 1998.

[16] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226-239, 1998.

[17] M. Le Blanc, and R. Tibshirani, "Combining Estimates in Regression and Classification," Technical Report 9318, Dept. of Statistics, University of Toronto, 1993.

[18] *Multiple Classifier Systems*, J. Kittler, and F. Roli, eds., LNCS vol. 1857 (2000) and 2096 (2001), Springer.

[19] J. Kittler, and F.M. Alkoot, "Sum versus Vote Fusion in Multiple Classifier Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, pp. 110-115, 2003.

[20] L.I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 281-286, 2002.

[21] L.I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Hoboken, N.J.: Wiley, 2004.

[22] M.P. Perrone, and L.N. Cooper, "When Networks Disagree: Ensemble Methods for Hybrid Neural Networks," *Neural Networks for Speech and Vision*, R.J. Mammone, ed., pp. 126-142. New York: Chapman-Hall, 1993.

[23] F. Roli, "Multisensor image recognition by neural networks with understandable behaviour," *Int'l J. Pattern Recognition Artificial Intelligence*, vol. 10, pp. 887-917, 1996.

[24] F. Roli, and G. Fumera, "Analysis of Linear and Order Statistics Combiners for Fusion of Imbalanced Classifiers," *Proc. Int'l Workshop Multiple Classifier Systems*, pp. 252-261, LNCS vol. 2364, Springer, 2002.

[25] F. Roli, G. Fumera, and J. Kittler, "Fixed and Trained Combiners for Fusion of Unbalanced Pattern Classifiers", *Proc. Int'l Conf. Information Fusion*, pp. 278-284, 2002.

[26] *Multiple Classifier Systems*, F. Roli, and J. Kittler, eds., LNCS vol. 2364, Springer, 2002.

[27] *Multiple Classifier Systems*, F.Roli, J. Kittler, and T. Windeatt, eds., LNCS vol. 3077, Springer, 2004.

[28] D.M.J. Tax, M. van Breukelen, R.P.W. Duin, and J. Kittler, "Combining Multiple Classifiers by Averaging or by Multiplying," *Pattern Recognition*, vol. 33, pp. 1475-1485, 2000.

[29] K. Tumer, "Linear and order statistics combiners for reliable pattern classification," PhD dissertation, The University of Texas, Austin, 1996.

[30] K. Tumer, and J. Ghosh, "Analysis of Decision Boundaries in Linearly Combined Neural Classifiers," *Pattern Recognition*, vol. 29, pp. 341-348, 1996.

[31] K. Tumer, and J. Ghosh, "Linear and order statistics combiners for pattern classification," *Combining Artificial Neural Nets*, A.J.C. Sharkey, ed., pp. 127-155. London: Springer, 1999.

[32] N. Ueda, "Optimal linear combination of neural networks for improving classification performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 207-215, 2000.

[33] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis, "Soft combination of neural classifiers: a comparative study," *Pattern Recognition Letters*, vol. 20, pp. 429-444, 1999.

[34] *Multiple Classifier Systems*, T. Windeatt, and F. Roli, eds., LNCS vol. 2709, Springer, 2003.

Giorgio Fumera obtained his M.S. degree, with honours, and Ph.D. degrees in Electronic Engineering from the University of Cagliari in 1997 and 2002. He is now assistant professor of computer engineering at the Department of Electrical and Electronic Engineering of the University of Cagliari, and a member of the research group on pattern recognition and applications. His current research interests are in the field of statistical pattern recognition, and include multiple classifier systems, classification with the reject option, support vector machines, and document categorization. Giorgio Fumera is a member of IAPR and IEEE.

Fabio Roli received his M.S. degree, with honours, and Ph.D. degree in Electronic Engineering from the University of Genoa, Italy. He was a member of the research group on Image Processing and Understanding of the Dept. of Biophysical and Electronic Engineering of the University of Genoa, Italy, from 1988 to 1994. He was adjunct professor at the University of Trento, Italy, in 1993 and 1994. In 1995, he joined the Dept. of Electrical and Electronic Engineering of the University of Cagliari, Italy, where he is now professor of computer engineering and head of the research group on pattern recognition and applications. Prof. Rolis current research activity is focused on multiple classifier systems and their applications to biometric personal identification, text categorization, and intrusion detection in computer networks. On such topics, he has published more than one hundred papers at conferences and on journals and has given lectures and tutorials. Prof. Roli has organised and co-chaired the five editions of the International Workshop on Multiple Classifier Systems. He is member of the IEEE and Fellow of the International Association for Pattern Recognition, editor of the Journal of Advances in Information Fusion, associate editor of the Electronic Letters on Computer Vision and Image Analysis and the Information Fusion journal, and member of the editorial Board of the International Journal of Computational Intelligence. Prof. Roli is the chairman of the IAPR Technical Committee on Statistical Techniques in Pattern Recognition and member of the Executive Board of the International Computational Intelligence Society.