Bayesian Analysis of Linear Combiners

Battista Biggio, Giorgio Fumera, and Fabio Roli

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

Abstract. A new theoretical framework for the analysis of linear combiners is presented in this paper. This framework extends the scope of previous analytical models, and provides some new theoretical results which improve the understanding of linear combiners operation. In particular, we show that the analytical model developed in seminal works by Tumer and Ghosh is included in this framework.

1 Introduction

One of the main open problems in the field of multiple classifier systems is the lack of a general theoretical framework which can give a unifying view of the large number of classifier combining rules and ensemble construction methods proposed so far in the literature [6]. With regard to combining rules, some theoretical results, with limited scope, are currently available for the majority voting and the linear combination of classifiers outputs. In particular, a theoretical framework for linear combiners, which are the focus of this paper, has been developed in seminal works by Tumer and Ghosh [8, 9], and was then exploited and extended in [1] and [2]. Theoretical analysis of linear combiners were also reported in [3–5]. The framework by Tumer and Ghosh was the first to provide useful insights into the behaviour of the linear combination by simple averaging, and some practical guidelines to the design of linearly combined classifier ensembles [8,9]. Fumera and Roli extended these results to the weighted average combining rule [2], and derived some guidelines for the choice between simple and weighted averaging. Although the theoretical predictions of the model by Tumer and Ghosh are derived under very strict and unrealistic assumptions, the authors noted that they were confirmed with good accuracy on many real data sets [2]. This raised an issue about the scope of the model by Tumer and Ghosh.

The work presented in this paper is a by-product of an attempt to provide an explanation to the above issue. We found that the theoretical analysis of the misclassification probability of individual and linearly combined classifiers given by Tumer and Ghosh can be developed under a new theoretical framework, which is presented in Sect. 2. The new theoretical framework has a broader scope than the one by Tumer and Ghosh, and includes it as a particular case, as explained in Sect. 3. We finally show in Sect. 4 that our framework provides some more insights into the operation of linear combiners, and also provides a partial answer to the open issue mentioned above about the prediction capability of Tumer and

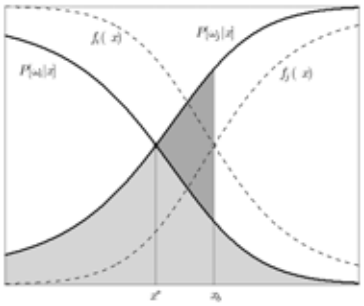

Fig. 1. True posteriors (solid lines) around the ideal boundary x_{opt} between ω_i and ω_j, and estimated posteriors (dashed lines) leading to the boundary x_b, and to an added error (dark gray area) over Bayes error (light gray area).

Ghosh model. We believe that the results presented in this paper can be a further step towards a more general framework for multiple classifier systems.

2 A Bayesian Framework for Generalization Error Analysis

Consider a given C-class classification problem, and a classifier which provides estimates $f_k(x), k = 1, \ldots, C$, of the a posteriori probabilities $\mathbb{P}[\omega_k|x]$, where x denotes a feature vector. The $f_k(x)$'s are considered random variables (their randomness depends for instance on the random choice of the training set). If Bayes decision rule is applied to the estimated posteriors, x is assigned to the class ω_i such that $i = \arg\max_k f_k(x)$, and non optimal decisions are taken if $\arg\max_k f_k(x) \neq \arg\max_k \mathbb{P}[\omega_k|x]$. This causes an additional misclassification probability (named *added error* in [8, 9]) over Bayes error. In the following we consider the case of a one-dimensional feature space: all the results can be extended to multi-dimensional feature spaces as described in detail in [7].

The framework by Tumer and Ghosh is based on analyzing the added error in a region of the feature space around an ideal boundary x_{opt} between two classes ω_i and ω_j, in the case in which the estimation errors lead to a boundary x_b between the *same* classes, which can be shifted from the ideal one. An example is given in Fig. 1. In this case, it is easy to see that the added error is given by

$$e_{\text{add}}(x_b) = \int_{x_{\text{opt}}}^{x_b} \left(\mathbb{P}[\omega_j|x] - \mathbb{P}[\omega_i|x]\right) \mathbb{P}[x] \mathrm{d}x. \tag{1}$$

Note that it depends on the posteriors of classes ω_i and ω_j only.

Our aim is instead to analyze the added error under more general conditions, namely in a region of the feature space around *any* given estimated boundary x_b between two classes ω_i and ω_j, without making any assumption on the true posteriors or on the presence of ideal boundaries in such region. To this aim, consider a given interval $[x_1, x_2]$ which contains an estimated boundary x_b. Assuming

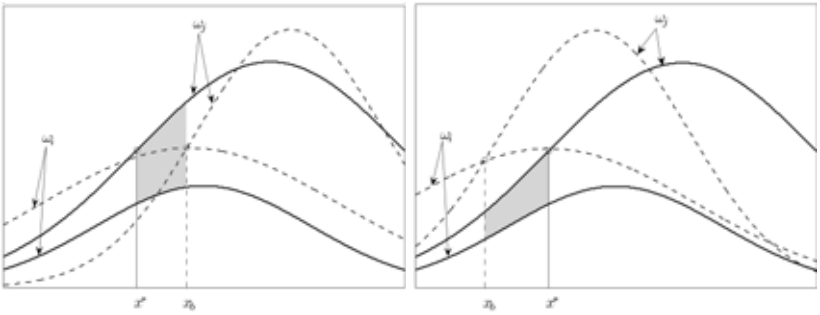

Fig. 2. Two possible realizations of the estimates of the posteriors of classes ω_i and ω_j (dashed lines), leading to an estimated class boundary x_b. The true posteriors are shown as solid lines. The difference $\Delta e_{\mathrm{add}}(x_{\mathrm{ref}}, x_b)$ (x_{ref} is the same in both plots) corresponds to the gray areas: it is positive in the left and negative in the right.

without loss of generality that $f_i(x_b) > f_j(x_b)$ for $x < x_b$, so that x is assigned to ω_i, if $x < x_b$, the added error in $[x_1, x_2]$ can be written as a function of x_b, as $e_{\mathrm{add}}(x_b) = \int_{x_1}^{x_b} (\mathbb{P}[\omega(x)|x] - \mathbb{P}[\omega_i|x])\mathbb{P}[x]\mathrm{d}x + \int_{x_b}^{x_2} (\mathbb{P}[\omega(x)|x] - \mathbb{P}[\omega_j|x])\mathbb{P}[x]\mathrm{d}x$, where $\omega(x) = \arg\max_{\omega_k} \mathbb{P}[\omega_k|x]$. Note that $e_{\mathrm{add}}(x_b)$ depends on the maximum of the posteriors in each x, which does not necessarily coincide with the posterior of ω_i or ω_j, contrary to the case considered by Tumer and Ghosh. To the purpose of our analysis, namely the comparison between the added error of individual classifiers and of their linear combination, it is convenient to remove the above dependence on $\mathbb{P}[\omega(x)|x]$. This can be achieved by considering any fixed reference point $x_{\mathrm{ref}} \in [x_1, x_2]$, and by rewriting $e_{\mathrm{add}}(x_b)$ as $e_{\mathrm{add}}(x_{\mathrm{ref}}) + [e_{\mathrm{add}}(x_b) - e_{\mathrm{add}}(x_{\mathrm{ref}})]$, where $e_{\mathrm{add}}(x_{\mathrm{ref}})$ is the added error that one would get if the estimated boundary x_b lay in x_{ref}. The term between square brackets is the difference between the added error when the estimated boundary lies in a point x_b, and $e_{\mathrm{add}}(x_{\mathrm{ref}})$, and will be denoted as $\Delta e_{\mathrm{add}}(x_{\mathrm{ref}}, x_b)$. It is easy to see that

$$\Delta e_{\mathrm{add}}(x_{\mathrm{ref}}, x_b) = \int_{x_{\mathrm{ref}}}^{x_b} (\mathbb{P}[\omega_j|x] - \mathbb{P}[\omega_i|x])\, \mathbb{P}[x]\mathrm{d}x \ . \tag{2}$$

An example is given in Fig. 2. Note now that $\Delta e_{\mathrm{add}}(x_{\mathrm{ref}}, x_b)$ depends on the posteriors of ω_i or ω_j only, contrary to both $e_{\mathrm{add}}(x_{\mathrm{ref}})$ and $e_{\mathrm{add}}(x_b)$. The main idea behind our framework is to express the added error of each individual classifier, as well as the one of the linear combiner, using the same reference point x_{ref}, as the sum of $e_{\mathrm{add}}(x_{\mathrm{ref}})$, which is a constant term *identical* for each classifier, and the term $\Delta e_{\mathrm{add}}(x_{\mathrm{ref}}, \cdot)$, which can be different for each classifier. This allows to evaluate the reduction of the added error which can be attained by the linear combination by comparing the latter term only.

We now formalize the main assumption on which our model is based. It is analogous to the main assumption of Tumer and Ghosh model reported in Sect. 3, although it was not explicitly phrased in this form in [8, 9].

Assumption 1. *Each realization of the random variables $f_k(x), k = 1, \ldots, C$, leads to an estimated boundary x_b between ω_i and ω_j in the considered interval $[x_1, x_2]$ of the feature space. No other estimated class boundary lies in $[x_1, x_2]$.*

As a consequence, x_b is a random variable whose distribution depends on the distribution of the $f_k(x)$'s.

2.1 Added Error of a Single Classifier

Following the approach in [8, 9], we start our analysis by writing the estimates $f_k(x)$ as $\mathbb{P}[\omega_k|x]+\varepsilon_k(x)$, where $\varepsilon_k(x)$ denotes the estimation error. An estimated boundary x_b between two classes ω_i and ω_j is characterized by $f_i(x_b) = f_j(x_b) > f_k(x_b), k \neq i, j$. We will denote with b the offset $x_b - x_{\text{ref}}$. As in [8,9], if b is small enough with respect to the changes in the posteriors and in $\mathbb{P}[x]$, a first order approximations of the posteriors and a zero order approximation of $\mathbb{P}[x]$ can be made around the reference point x_{ref}: $\mathbb{P}[\omega_k|x_{\text{ref}}+b] \simeq \mathbb{P}[\omega_k|x_{\text{ref}}]+b\mathbb{P}'[\omega_k|x_{\text{ref}}], k = i, j$, e $\mathbb{P}[x_{\text{ref}} + b] \simeq \mathbb{P}[x_{\text{ref}}]$. Substituting in Eq. 2 we obtain

$$\Delta e_{\text{add}}(x_{\text{ref}}, x_b) = \frac{\mathbb{P}[x_{\text{ref}}]t}{2}\left(\frac{2u}{t}b + b^2\right) \ , \tag{3}$$

where

$$u = \mathbb{P}[\omega_j|x_{\text{ref}}] - \mathbb{P}[\omega_i|x_{\text{ref}}], \ t = \mathbb{P}'[\omega_j|x_{\text{ref}}] - \mathbb{P}'[\omega_i|x_{\text{ref}}] \ . \tag{4}$$

The expected value of $\Delta e_{\text{add}}(x_{\text{ref}}, x_b)$ with respect to b is then

$$\Delta E_{\text{add}} = \mathbb{E}[\Delta e_{\text{add}}(x_{\text{ref}}, x_b)] = \frac{\mathbb{P}[x_{\text{ref}}]t}{2}\left[\frac{2u}{t}\beta_b + \beta_b^2 + \sigma_b^2\right] \ , \tag{5}$$

where β_b and σ_b^2 denote the expected value and the variance of b.

It is also possible to express b as a function of the estimation errors: this allows to rewrite Eq. 5 in a form which will be useful to compare the expected added error of an individual classifier with the one of linearly combined classifiers. From $f_i(x_b) = f_j(x_b)$, rewriting $f_k(x), k = i, j$, as $\mathbb{P}[\omega_k|x] + \varepsilon_k(x)$, and using the first order approximation of the posteriors, we obtain

$$b = \frac{\varepsilon_i(x_b) - \varepsilon_j(x_b)}{t} - \frac{u}{t} \ . \tag{6}$$

Assuming as in [8, 9] that the estimation errors on different classes $\varepsilon_i(x)$ and $\varepsilon_j(x)$ are uncorrelated, from Eq. 6 we obtain

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

where β_k and $\sigma_k^2, k = i, j$, denote the expected value (named *bias* in [8, 9]) and the variance of $\varepsilon_k(x_b)$. Substituting the above expression of β_b into Eq. 5 we obtain

$$\Delta E_{\text{add}} = \frac{\mathbb{P}[x_{\text{ref}}]t}{2}\left[-\frac{u^2}{t^2} + \frac{1}{t^2}(\beta_i - \beta_j)^2 + \frac{1}{t^2}(\sigma_i^2 + \sigma_j^2)\right] \ . \tag{8}$$

The expected added error in $[x_1, x_2]$ is then given by

$$E_{\text{add}} = e_{\text{add}}(x_{\text{ref}}) + \Delta E_{\text{add}} \ . \tag{9}$$

It is easy to see that the expected added error is the sum of three terms: a constant term $e_{\text{add}}(x_{\text{ref}}) - \frac{\mathbb{P}[x_{\text{ref}}]u^2}{2t}$, whose value depends only on the choice of the reference point x_{ref}; a term depending on the bias of estimation errors, and the other on their variance.

2.2 Added Error of Linearly Combined Classifiers

Consider now a linear combination of the posteriors estimates provided by an ensemble of N classifiers, $f_k^n(x), k = 1, \ldots, C; n = 1, \ldots, N$, using positive weights w_n which sum up to 1, as in [2]:

$$f_k^{\text{ave}}(x) = \sum_{n=1}^{N} w_n f_k^n(x) = \mathbb{P}[\omega_k|x] + \varepsilon_k^{\text{ave}}(x) = \mathbb{P}[\omega_k|x] + \sum_{n=1}^{N} w_n \varepsilon_k^n(x) \ . \tag{10}$$

To proceed with our analysis, we extend the assumption 1 to the estimates of each individual classifier, and of their linear combination. Now, as in Sect. 2.1, we rewrite the added error $e_{\text{add}}^{\text{ave}}(x_{b^{\text{ave}}})$ in $[x_1, x_2]$ as $e_{\text{add}}^{\text{ave}}(x_{\text{ref}}) + [e_{\text{add}}^{\text{ave}}(x_{b^{\text{ave}}}) - e_{\text{add}}^{\text{ave}}(x_{\text{ref}})]$, using the same reference point x_{ref} as in each individual classifier. With the same approximations, assumptions and steps as in Sect. 2.1, we obtain:

$$b^{\text{ave}} = \frac{\varepsilon_i^{\text{ave}}(x_{b^{\text{ave}}}) - \varepsilon_j^{\text{ave}}(x_{b^{\text{ave}}})}{t} - \frac{u}{t} \ , \tag{11}$$

while the expected value of $\Delta e_{\text{add}}^{\text{ave}}(x_{\text{ref}}, x_{b^{\text{ave}}})$ is

$$\Delta E_{\text{add}}^{\text{ave}} = \frac{\mathbb{P}[x_{\text{ref}}]t}{2} \left\{ -\frac{u^2}{t^2} + \frac{1}{t^2}(\beta_i^{\text{ave}} - \beta_j^{\text{ave}})^2 + \frac{1}{t^2}\left[(\sigma_i^{\text{ave}})^2 + (\sigma_j^{\text{ave}})^2\right] \right\} \ , \tag{12}$$

where

$$\beta_k^{\text{ave}} = \sum_{n=1}^{N} w_n \beta_k^n, \ (\sigma_k^{\text{ave}})^2 = \sum_{n=1}^{N} w_n^2 (\sigma_k^n)^2 + \sum_{n=1}^{N} w_n^2 \sum_{m \neq n} \rho_k^{mn} \sigma_k^m \sigma_k^n, k = i, j \ , \tag{13}$$

ρ_k^{mn} denotes the correlation coefficient between $\epsilon_k^m(x)$ and $\varepsilon_k^n(x)$, and σ_k^m is the standard deviation of $\varepsilon_k^m(x)$. Finally, the expected added error in $[x_1, x_2]$ is

$$E_{\text{add}}^{\text{ave}} = e_{\text{add}}(x_{\text{ref}}) + \Delta E_{\text{add}}^{\text{ave}}. \tag{14}$$

Eqs. 14,12 show that the expected added error of the linear combiner, as the one of individual classifiers (see Eqs. 9 and 8), is given by the *same* constant term $e_{\text{add}}(x_{\text{ref}}) - \frac{\mathbb{P}[x_{\text{ref}}]u^2}{2t}$, plus a bias term and a variance term. The error reduction attainable by the linear combination can thus be evaluated taking into account only the bias and variance terms.

In the next Section we will point out the different scopes of the two frameworks in their capability of modeling the added error, and show that Tumer and Ghosh framework is included in ours. In Sect. 4 we will then compare the predictions about the behaviour of linear combiners which can be obtained from the two frameworks.

3 Comparison with Tumer and Ghosh Framework

As explained in Sect. 2, Tumer and Ghosh framework differs from ours since it evaluates the added error in an interval of the feature space containing an *ideal* boundary x_{opt} between two classes ω_i and ω_j, which is characterized by $\mathbb{P}[\omega_i|x_{\text{opt}}] = \mathbb{P}[\omega_j|x_{\text{opt}}] > \mathbb{P}[\omega_k|x_{\text{opt}}], k \neq i, j$. The main assumption of Tumer and Ghosh framework can be phrased as follows.

Assumption 2. *Each realization of the random variables $f_k^n(x), k = 1, \ldots, C; n = 1, \ldots, N$, leads to an estimated boundary between ω_i e ω_j, in a given interval $[x_1, x_2]$ which contains an ideal boundary x_{opt} between the same classes, both for each individual classifier and for their linear combination. Furthermore, there are no other estimated or ideal class boundaries in the considered interval.*

This is more restrictive than assumption 1 of our framework, which does not require the presence of such an ideal boundary, and thus allows to model the added error only for a subset of cases which can be modeled by our framework. Under assumption 2, the added error in $[x_1, x_2]$ is given by Eq. 1. Denoting the offset $x_b - x_{\text{opt}}$ with b, making a first order approximation of the posteriors and a zero order approximation of $\mathbb{P}[x]$ around x_{opt}, and assuming that the estimation errors on different classes are uncorrelated as in Sect. 2, it turns out [8,9] that

$$b^n = \frac{\varepsilon_i(x_{b^n}) - \varepsilon_j(x_{b^n})}{t}, \; b^{\text{ave}} = \frac{\varepsilon_i(x_{b^{\text{ave}}}) - \varepsilon_j(x_{b^{\text{ave}}})}{t} \;\; , \tag{15}$$

while the expected added error in $[x_1, x_2]$ is

$$\begin{array}{c} E_{\text{add}}^n = \frac{\mathbb{P}[x_{\text{opt}}]t}{2} \left\{ \frac{1}{t^2}(\beta_i^n - \beta_j^n)^2 + \frac{1}{t^2}\left[(\sigma_i^n)^2 + (\sigma_j^n)^2\right] \right\}, \\ E_{\text{add}}^{\text{ave}} = \frac{\mathbb{P}[x_{\text{opt}}]t}{2} \left\{ \frac{1}{t^2}(\beta_i^{\text{ave}} - \beta_j^{\text{ave}})^2 + \frac{1}{t^2}\left[(\sigma_i^{\text{ave}})^2 + (\sigma_j^{\text{ave}})^2\right] \right\} \;\; , \end{array} \tag{16}$$

where t, $\varepsilon_k^{\text{ave}}(x)$, β_k and $\sigma_k^2, k = i, j$, are defined exactly as in Sect. 2. It is worth noting that the two expressions of the expected added error are the sum of a bias and a variance term formally identical to the ones derived from our model (see Eqs. 13 and 8 for an individual classifier, and 13, 12 for a linear combiner): the only difference is that our model leads to a further constant additive term due to the fact that in our model the reference point x_{ref} needs not to coincide with an ideal boundary x_{opt} (which could even not exist in $[x_1, x_2]$).

We now show that Tumer and Ghosh framework is included in ours, in the sense that, under the more restrictive assumption 2, they both lead to the same expression of the expected added error in the considered interval, provided that the reference point x_{ref} is chosen equal to the ideal boundary x_{opt} between ω_i

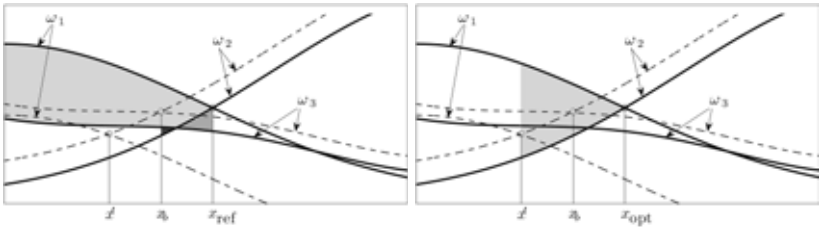

Fig. 3. True (solid lines) and estimated posteriors (dashed lines) for a three-class problem, leading to an ideal boundary x_{opt} between ω_1 and ω_2, and to an estimated boundary x_b between ω_3 and ω_2. The added error corresponds to the light gray plus the dark gray area in the left panel, while it would be erroneously evaluated by Tumer and Ghosh framework as the gray area in the right panel (see text for a complete explanation).

and ω_j. To this aim, it is sufficient to note that in this case the term $u = \mathbb{P}[\omega_j|x_{\text{opt}}] - \mathbb{P}[\omega_i|x_{\text{opt}}]$ is null, since the true posteriors are equal in the ideal boundary x_{opt}, and the term $e_{\text{add}}(x_{\text{opt}})$ is null as well, since, when the estimated boundary coincides with the ideal one ($x_b = x_{\text{opt}}$), the added error vanishes. It immediately follows that the expressions of b and of the expected added error given by the two frameworks are identical.

We finally point out that there are cases in which the added error can be correctly modeled by our framework only. This happens when assumption 1 holds while 2 does not, namely when there is no ideal boundary between ω_i and ω_j in the considered interval, or equivalently when the effect of the estimation errors on the posteriors is not a shift of an ideal class boundary. It is worth noting that these are cases of practical interest: as pointed out in [6], in complex pattern recognition problems it is likely that estimation errors cause different effects besides the shift of ideal boundaries. To clear up this point, we consider an example taken from [6] for a three-class problem, which is illustrated in Fig. 3, left. In the considered interval there is an ideal boundary x_{opt} between ω_1 and ω_2, while estimation errors lead to a boundary x_b between ω_3 and ω_2. Note that there is also a point x' such that $f_1(x') = f_2(x')$, which however is not an estimated boundary between ω_1 and ω_2. The true added error corresponds to the light gray area in Fig. 3, left. Tumer and Ghosh framework would erroneously model it with reference to the point x', as the grey area in Fig. 3, right. Consider instead $x_{\text{ref}} = x_{\text{opt}}$ as the reference point for our framework, for the sake of simplicity (any other point could be used as well). Our framework correctly models the added error as the sum of $e_{\text{add}}(x_{\text{ref}})$, namely the one corresponding to $x_b = x_{\text{ref}}$ (the sum of the light gray and of the intermediate gray areas in Fig. 3, left), and of $\Delta e_{\text{add}}(x_{\text{ref}}, x_b)$ (the dark gray area minus the intermediate gray area), which results in the sum of the light and dark gray areas in Fig. 3, left.

To sum up, our model has a broader scope than the one by Tumer and Ghosh, since it allows to model the added error under more general conditions. As a result, one may expect that our framework gives more accurate predictions on

the behaviour of linear combiners, which in turn could provide better guidelines for their design. This issue is discussed in the next Section.

4 Analysis of Simple and Weighted Averaging

In [8,9] the model by Tumer and Ghosh was exploited to evaluate the reduction of the added error attainable by the simple average combining rule (from now on, SA) with respect to individual classifiers. Some further results were pointed out in [2]. The main results were the following:

1. SA reduces the variance component of the expected added error by an amount which depends on the correlation between estimation errors ρ_k^{mn} of the different classifiers (see Eqs. 16 and 13); for negatively correlated errors, the variance component can be reduced up to zero.
2. SA guarantees at least a bias component not greater than the maximum one exhibited by the individual classifiers.

These results suggested in [8,9] that the design of individual classifiers should focus on obtaining low bias and correlation, while the variance can be reduced by averaging classifiers.

What does the model presented in Sect. 2 add to the above results? Note first that in our model the bias and variance components of the expected added error, in Eqs. 8 ($\frac{\mathbb{P}[x_{\text{ref}}]t}{2}[\frac{1}{t^2}(\beta_i-\beta_j)^2+\frac{1}{t^2}(\sigma_i^2+\sigma_j^2)]$) and 12 ($\frac{\mathbb{P}[x_{\text{ref}}]t}{2}\{\frac{1}{t^2}(\beta_i^{\text{ave}}-\beta_j^{\text{ave}})^2+\frac{1}{t^2}[(\sigma_i^{\text{ave}})^2+(\sigma_j^{\text{ave}})^2]\}$), can be either positive or negative depending on the sign of the term t given by Eq. 4, while t is always positive in Tumer and Ghosh model. If $t > 0$, the bias and variance components of the two models are identical, and thus the above results provided by Tumer and Ghosh model hold also for ours. Instead, if $t < 0$, the bias and variance components derived from our model are *negative*. This implies that the expected added error of the SA can even be higher than the one of each individual classifier, and anyway it can never be lower than that of the best individual classifier. Furthermore, the reduction in the expected added error increases for increasing correlation between the estimation errors. Therefore, in presence of different estimated boundaries characterized by both positive and negative values of the corresponding t, the net effect of SA will be determined by the counterbalance of the two behaviours above. In other words, the advantage of SA over individual classifiers could be lower than the one predicted by Tumer and Ghosh model.

A further exploitation of Tumer and Ghosh model was carried out in [2], to compare the behaviour of the SA with that of the weighted average (from now on, WA) combining rule. An analytical comparison was possible only under the simplest case of unbiased and uncorrelated errors. The main theoretical results derived in [2] can be summarized as follows:

1. SA is the optimal linear combining rule, only if the individual classifiers exhibit the same misclassification rate.
2. WA can always perform at worst as the best classifier of the ensemble.

3. The improvement in misclassification rate which can be attained by WA over SA depends on the *error range* of the ensemble, namely on the misclassification rates of the best and worst individual classifiers: the broader the error range, the higher the improvement; being equal the error range, the improvement strongly depends on the degree of performance *imbalance*, namely on the distribution of the misclassification rates of the other individual classifiers. In particular, for classifiers exhibiting a narrow error range (say, below 0.05), the advantage of WA over SA is quite small (say, below 0.01).

This suggested in [2] some new simple guidelines for the choice between SA and WA in real applications. Basically, it could be worth using WA only if the individual classifiers exhibit a broad error range (say, above 5%), unless the weights can be estimated with high reliability; otherwise the small ideal advantage can be canceled out by weight estimations from small and noisy data sets. Although the assumption of unbiased and uncorrelated errors, as well as the main assumption of Tumer and Ghosh model (being the effect of estimation errors the shift of ideal boundaries) are likely to be violated in practice, it turned out that the derived predictions about the behaviour of SA and WA, and thus the validity of the above guidelines, were confirmed by experimental results on real data sets reported in [2]. It was left as an open problem to understand why theoretical predictions derived under assumptions which were apparently very restrictive were confirmed on real data sets. A partial answer can be given thanks to the new model described in this paper. By carrying out the same analysis described in detail in [2] for the case of unbiased and uncorrelated errors, it turns out that our model gives the *same* predictions above about the behaviour of WA and SA, for the case in which $t > 0$. Moreover, for $t < 0$ *only* prediction 2 above changes: in this case WA performs at best (instead of at worst) as the best individual classifier. This is thus an indication that the predictions derived from Tumer and Ghosh model were confirmed on real data sets since they actually hold under more general conditions. We point out that experimental results analogous to [2] (not reported here due to space limits) were obtained on six more real data sets taken from the UCI repository, namely Optdigits, DNA, Ionosphere, Satellite, Satimage and Segmentation.

5 Conclusions

In this paper we presented a new theoretical framework for the analysis of the reduction in misclassification probability which can be attained by linearly combining an ensemble of classifiers which provide estimates of the a posteriori probabilities. Our framework has a broader scope than the one developed in works by Tumer and Ghosh, and includes it as a particular case. It allows to analyze the added error around any class boundary provided by the estimated posteriors, not only around boundaries which are shifted from ideal ones as in [8, 9]. This gives a more general understanding of the operation of linear combiners. In particular, this allowed us to point out some behaviours of linear combiners

(technically, the cases in which the term t is negative) which were not contemplated by Tumer and Ghosh model. Nevertheless, we found that many of the predictions of our model, in particular the ones from which practical guidelines for the design of linear combiners can be derived, are nearly identical to the predictions derived from the previous model: this gives a partial explanation to an open issue pointed out in [2], raised by the fact that theoretical prediction derived by Tumer and Ghosh model under strict and unrealistic assumptions turned out to be experimentally confirmed on real data sets.

To sum up, the main contribution of this paper is the development of a theoretical framework which allows a more general understanding of linear combiners. We are also investigating whether the ideas behind the theoretical frameworks considered in this work could suggest new theoretical models for other combining rules, which would be a useful step towards a more general framework for multiple classifier systems.

Acknowledgment The authors would like to thank Gavin Brown for his valuable comments and suggestions on this work.

References

1. Alexandre, L.A., Campilho, A.C. Kamel, M.: Combining Independent and Unbiased Classifiers Using Weighted Average. In: Proc. 15th Int'l Conf. on Pattern Recognition, Vol. 2. IEEE Press (2000) 495–498
2. Fumera, G., Roli, F.: A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. IEEE Trans. Pattern Analysis Machine Intelligence **27** (2005) 942–956
3. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Analysis Machine Intelligence **20** (1998) 226–239
4. Kittler, J., Alkoot, F. M.: Sum versus Vote Fusion in Multiple Classifier Systems. IEEE Trans. Pattern Analysis Machine Intelligence **25** (2003) 110–115
5. Kuncheva, L.I.: A theoretical study on six classifier fusion strategies. IEEE Trans. Pattern Analysis Machine Intelligence **24** 281–286
6. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Hoboken, N.J., Wiley (2004)
7. Tumer, K.: Linear and order statistics combiners for reliable pattern classification. PhD dissertation, The University of Texas, Austin (1996)
8. Tumer, K., Ghosh, J.: Analysis of Decision Boundaries in Linearly Combined Neural Classifiers. Pattern Recognition **29** (1996) 341–348
9. Tumer, K., Ghosh, J.: Linear and order statistics combiners for pattern classification. In: Sharkey, A.J.C. (ed.): Combining Artificial Neural Nets. Springer (1999) 127–155