International Summer School on Neural Nets Eduardo R. Caianiello"
ENSEMBLE METHODS FOR LEARNING MACHINES
Vietri sul Mare(Salerno) – ITALY 22-28 September 2002

Linear Combiners for Fusion of Pattern Classifiers

Lecturer

Prof. Fabio ROLI

University of Cagliari

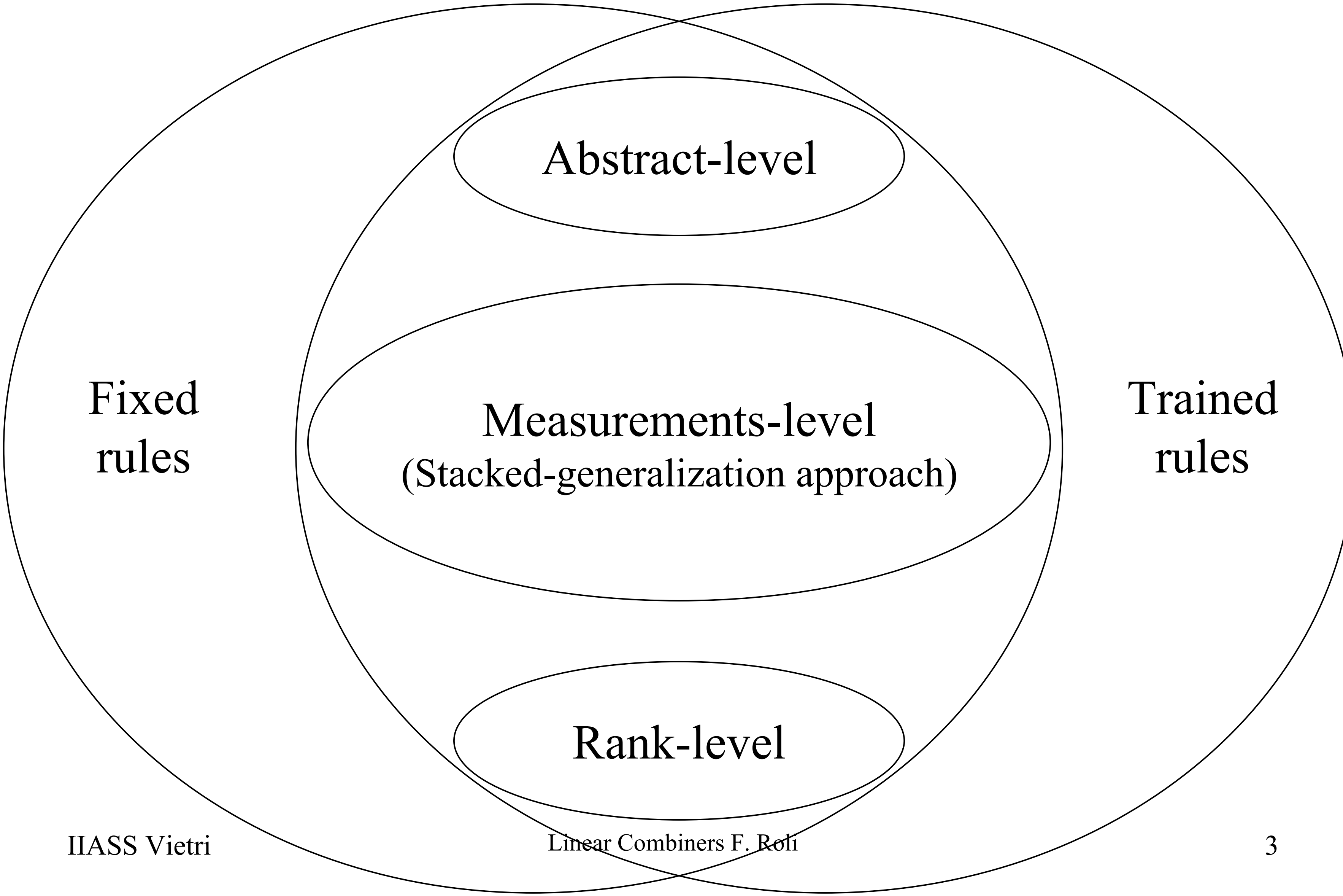
Dept. of Electrical and Electronics Eng., Italy

email roli@diee.unica.it

Methods for generating classifier ensembles

- The effectiveness of ensemble methods relies on combining *diverse*/*complementary* classifiers
- Several approaches have been proposed to construct ensembles made up of complementary classifiers. Among the others:
  - Injecting randomness
  - Varying the classifier architecture, parameters, or type
  - Manipulating the training data / input features / output features
- Examples:
  - Data Splitting
  - Bootstrap
  - Random Subspace Method

Methods for fusing multiple classifiers



State of the Art

- Despite observed successes in may experiments and real applications, there is no guarantee that a given method will work well for the task at hand
- For each method, we have evidences that it does not always work
- For a given task, the choice of the most appropriate combination method lies on the usual paradigm of model evaluation and selection
- Many key concepts (e.g., diversity) need to be formally defined
- Few theoretical explanations of observed successes and failures

State of the Art

- In particular, we have few theoretical studies that compared different combination rules (e.g., Kittler et al., PAMI 1998; L.I.Kuncheva, PAMI 2002)
- Surely, a general and unifying framework is very far to appear.
- However, "….we have to start somewhere…" (L.I.Kuncheva, PAMI 2002)
- Theoretical works aimed to compare a limited set of rules, even if under strict assumptions, are mandatory steps towards a general framework
- In addition, practical applications demands for some quantitative guidelines, under realistic assumptions *(there is nothing more useful than a good theory)*

A class of Fusers: Linear Combiners

- Linear combiners: Simple and Weighted averaging of classifiers' outputs
- Many observed successes of these simple combiners (Bagging, Random Subspace Method, Mixtures, etc.)
- However, many important aspects had for a long time (and still have) just qualitative explanations:
  - Effects of classifiers correlations on linear combiners performances
  - Effects of errors and correlations imbalance
  - Quantitative comparison between simple and weighed average
- So far, it is not completely clear when, and how much, simple averaging can perform well, and when weighted average can significantly outperform it

An example of unclear results
(Roli and Fumera, MCS 2002)

- Test set error rates (averaged over ten runs)

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

	combiner error rates			optimal weights		
	E^{sa}	E^{wa}	$\boldsymbol{E^{sa}-E^{wa}}$	*k*-NN	MLP1	MLP2
Ensemble1	10.00	9.37	**0.63**	0.576	0.200	0.224
Ensemble2	12.09	9.69	**2.40**	0.689	**0.080**	0.231
Ensemble3	10.69	9.63	**1.06**	0.681	0.231	**0.088**
Ensemble4	16.81	9.79	**7.02**	0.838	**0.006**	0.156
Ensemble5	12.44	9.73	**2.71**	0.752	**0.103**	0.143

Outline of the Lecture

1. An analytical framework for simple averaging of classifiers' outputs
2. Extension of the framework to weighted average
3. Analytical and numerical comparison between simple and weighted average
4. Conclusions

An analytical framework for simple averaging
(Tumer and Ghosh, 1996, 1999)

- An analytical framework to *quantify* the improvements in classification accuracy due to simple averaging of classifiers' outputs has been developed by Tumer and Ghosh
- This framework applies to classifiers which provide approximations of the posterior probabilities
- The framework shows that simple averaging can reduce the error "added" to the Bayes one
- In particular, Tumer and Ghosh analysis points out and quantifies the effect of output correlations on simple averaging accuracy

An analytical framework for simple averaging

- Consider the output of an individual classifier for class *i*, given an input pattern *x*:

$$f_i(x) = p_i(x) + \varepsilon_i(x)$$

  – $p_i(\mathbf{x})$: a posteriori probability of class *i*
  – $\varepsilon_i(\mathbf{x})$: estimation error

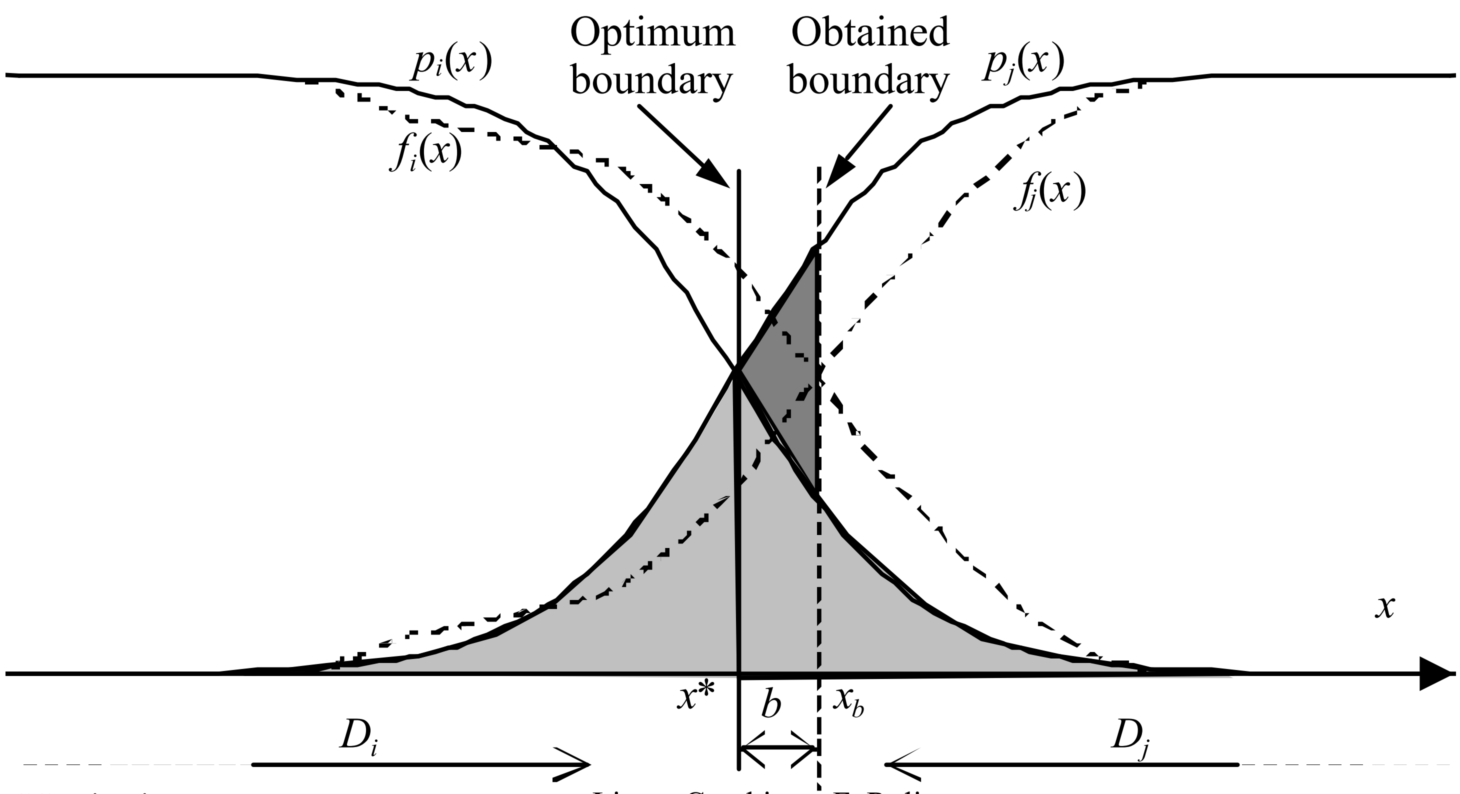
➢ Assume estimation errors are small and concentrated around the boundaries, so that the obtained decision boundaries are close to the optimal Bayesian boundaries

- This allows to focus the analysis around the decision boundaries

➢ The following analysis is made for a one-dimensional case, but it can be extended to the multi-dimensional case (Tumer, 1996)

Analysis around the decision boundaries

- Assume that the estimation errors cause a "small" shift of the optimal decision boundary by an amount *b*: $f_i(x^{*}+b) = f_j(x^{*}+b)$
- This shift produces the *added error* region shown in figure (darkly shaded area) over Bayes error (lightly shaded area)



Added error probability

- Tumer and Ghosh showed that the added error probability can be expressed as a function of the distribution of the estimation errors $\varepsilon_i(x)$ and $\varepsilon_j(x)$

- To compute the expected value of added error probability:
  - A first order approximation is used for $p_k(x)$ around x^*:

$$p_k(x^*+b) \cong p_k(x^*) + bp'_k(x^*)$$

  - The error $\varepsilon_k(x)$ is broken into a bias β_k and a noise term η_k with variance σ^2_k:

$$\varepsilon_k(x) = \beta_k + \eta_k(x)$$

  - ➢ Important hypothesis: the $\eta_k(x)$ are i.i.d. variables with variance σ^2
  - ➢ The most likely values of the shift "b" are small

Added error probability

The expected added error can be written as:

$$E_{add} = A(b) = \int_{-\infty}^{+\infty} A(b) f_b(b) db$$

Using the first order approximation $p_k(x^*+b) \cong p_k(x^*) + bp'_k(x^*)$:

$$A(b) = \frac{1}{2}(x^* - x_b)(p_j(x^* + x_b) - p_i(x^* + x_b)) = \frac{1}{2} b^2 s$$

$$s = p'_j(x^*) - p'_i(x^*)$$

Accordingly: $$E_{add} = \int_{-\infty}^{+\infty} \frac{1}{2} b^2 s f_b(b) db = \frac{s}{2} \sigma_b^2$$

For unbiased classifiers it is easy to show that: $\sigma_b^2 = \dfrac{2\sigma^2}{s^2}$

Therefore $\mathbf{E_{add}} = \dfrac{\mathbf{1}}{\mathbf{s}} \sigma^{\mathbf{2}}$

Added error for individual classifiers

- For the case of biased classifiers, Tumer and Ghosh showed that the expected value E_{add} is:

$$E_{add} = \frac{1}{s}\sigma^2 + \frac{1}{2s}\left(\beta_i - \beta_j\right)^2$$

where s is a constant term

- E_{add} is the sum of two terms:
  – the first term is proportional to the variance of the estimation errors
  – the second term is proportional to the squared difference of the biases of classes i and j

➢ Remind that the total error is the sum of the added error and the Bayes error: $E_{tot} = E_{add} + E_{bayes}$

Simple averaging of classifiers' outputs

- The approximation of $p_i(x)$ provided by averaging the outputs of N classifiers is:

$$f_i^{ave}(x) = \frac{1}{N}\sum_{m=1}^{N} f_i^m(x)$$
$$= p_i(x) + \overline{\beta}_i + \overline{\eta}_i(x)$$

where

$$\overline{\eta}_i(x) = \frac{1}{N}\sum_{m=1}^{N} \eta_i^m(x)$$

$$\overline{\beta}_i = \frac{1}{N}\sum_{m=1}^{N} \beta_i^m$$

- Uncorrelated classifiers:
  - the $\eta_i^m(x)$, $m = 1,\ldots,N$, are i.i.d. variables

Simple averaging of unbiased and uncorrelated classifiers

- Unbiased estimation errors: $\beta_i^m = 0$, m = 1,…,N
- Again, important hypothesis:
  - the $\eta_i^m(x)$, $m = 1,\ldots,N$, are i.i.d. variables

- Tumer and Ghosh showed that the variance of the estimation error is reduced by a factor N by averaging:

$$\sigma_{\bar{\eta}}^2 = \frac{1}{N}\sigma^2$$

- Accordingly, the added error of individual classifiers is reduced by a factor N:

$$E_{add}^{ave} = \frac{1}{N} E_{add}$$

Remarks

- We are assuming that the estimation errors of individual classifiers have the same variance σ^2
- For unbiased classifiers, this means that:

$$E_{add} = \frac{1}{s}\sigma^2$$

- I.e., classifiers exhibit equal errors (“balanced” classifiers)
- Classifiers can be imbalanced in the biased case, but Tumer and Ghosh did not analyse explicitly the effect of such “imbalance” on simple averaging performances

Simple averaging of biased classifiers

- Added error of individual classifiers:

$$E_{add}^{m} = \frac{1}{s}\sigma^{2} + \frac{1}{2s}\left(\beta_{i}^{m} - \beta_{j}^{m}\right)^{2}$$

- Added error of the combination of *N* classifiers:

$$E_{add}^{ave} = \frac{1}{Ns}\sigma^{2} + \frac{1}{2s}\left(\overline{\beta}_{i} - \overline{\beta}_{j}\right)^{2}$$

- The variance component is reduced by a factor *N*
- The bias component is not necessarily reduced by *N*

➢ *Averaging is very effective for reducing the variance component, but not for the bias component*

➢ *So, individual classifiers with low biases should be preferred*

Simple averaging of biased classifiers

- Added error of simple averaging can be rewritten as:

$$E_{add}^{ave} \leq \left(\frac{\sigma^2}{N} + \frac{\beta^2}{z^2}\right)$$

- Where β can be regarded as the bias of an individidual classifier, and $z \leq \sqrt{N}$
- If the contributions to the added error of the variance and the bias are of similar magnitude, the actual reduction is given by min(z^2, N)
- If the bias can be kept low, then once again *N* become the reduction factor

Correlated and unbiased classifiers

- Hypothesis:
  - the $\eta_i^m(x)$, $m = 1,\ldots,N$, are identically distributed, but correlated variables
- Added error of individual classifiers:

$$E_{add}^{m} = \frac{1}{s}\sigma^2$$

- Added error of the linear combination of N classifiers:

$$E_{add}^{ave} = E_{add}\left(\frac{1+(N-1)\delta}{N}\right)$$

where

$$\delta = \sum_{i=1}^{L}\delta_i, \quad \delta_i = \frac{1}{N(N-1)}\sum_{m=1}^{N}\sum_{n\neq m} corr\left(\eta_i^m(x), \eta_i^n(x)\right)$$

Correlated and unbiased classifiers

- The reduction factor achieved by simple averaging depends on the correlation between the estimation errors
- Three cases can happen:
  - $\delta > 0$ (positive correlation):
    the reduction factor is less than N
  - $\delta = 0$ (uncorrelated errors):
    the reduction factor is N (as shown previously)
  - $\delta < 0$ (negative correlation):
    the reduction factor is greater than N

➢ Negatively correlated estimation errors allow to achieve a greater improvement than independent errors

Remarks

- The correlation δ is:

$$\delta \geq -\frac{1}{N-1}$$

•As more and more classifiers are used (increasing N), it become very difficult to design uncorrelated classifiers

Correlated and biased classifiers

- Added error of individual classifiers:

$$E_{add}^{m} = \frac{1}{s}\sigma^{2} + \frac{1}{2s}\left(\beta_{i}^{m} - \beta_{j}^{m}\right)^{2}$$

- Added error of the linear combination of N classifiers:

$$E_{add}^{ave} = \frac{1}{s}\sigma^{2}\left(\frac{1+(N-1)\delta}{N}\right) + \frac{1}{2s}\left(\overline{\beta}_{i} - \overline{\beta}_{j}\right)^{2}$$

- As for uncorrelated errors, averaging is effective for reducing the variance component of the added error

Remarks

- Tumer and Ghosh analysis assumes a single decision boundary for each couple of data classes
- So, some conclusions (e.g., for the unbiased and uncorrelated case) can be optimistic
- For a given classification task, different decision boundaries can exhibit different estimation errors
- They did not analyse the effect of classifiers with different errors and pair-wise correlations ("imbalanced" classifiers) on simple averaging performances

➢ **Tumer and Ghosh analysis does not deal with the general case of linear combiners (Weighted Average)**

Experimental Evidences

- SONAR data set (Tumer and Ghosh, 1999)
- Two distinct feature sets and two neural nets (MLP and RBF)

- They showed that using different classifiers trained with different features sets provides low/negative correlated outputs
- Simple averaging of such uncorrelated classifiers reduces the error over the best individual classifiers of about 3%

Multimodal biometrics

(Roli et al., 5th Int. Conf. on Information Fusion, 2002)

- XM2VTS database
  - face images, video sequences, speech recordings
  - 200 training and 25 test clients, 70 test impostors

- Eight classifiers based on different techniques
  - two speech classifiers
  - six face classifiers

Multimodal biometrics application

- Test set error rates of individual classifiers

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

- The four classifier ensembles

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

Multimodal biometrics application

- Test set average error rates of simple averaging vs. BKS

	S.A.	BKS
Ens. 1	6.014	4.909
Ens. 2	5.739	4.246
Ens. 3	4.420	0.474
Ens. 4	5.509	3.487

- For three cases out of four, the difference between simple averaging and BKS lies in the range 1% - 2%

- Simple averaging performs reasonably well also for imbalanced classifiers. Especially, for uncorrelated classifiers with balanced pair-wise correlations !

Extension to Weighted Average

(Roli and Fumera, SPR 2002, MCS 2002)

- N linearly combined classifiers, normalised weights w_k

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

- Hypotheses:
  - the ε_i^k are unbiased (β_i^k=0)
  - $\forall\ m,n\ \eta_i^m$ and η_i^n are correlated, but the correlation coefficient ρ^{mn} does not dependent on the class i
  - η_i^m and η_j^n are uncorrelated for $i \neq j$, $\forall\ m,n$
  - Individual classifiers can have different variances !
- The probabilities estimated by the combiner are:

$$f_i^{ave}(x) = \sum_{k=1}^{N} w_k f_i^k(x) = p_i(x) + \sum_{k=1}^{N} w_k \eta_i^k(x) = p_i(x) + \bar{\eta}_i(x)$$

$$\bar{\eta}_i(x) = \sum_{k=1}^{N} w_k \eta_i^k(x)$$

Added error for Weighted Average

- Roli and Fumera showed that the added error around the boundary between classes *i* and *j* can be expressed as:

$$E_{add}^{ave} = \frac{1}{s}\sum_{k=1}^{N}\sigma_{\eta^k}^2 w_k^2 + \frac{1}{s}\sum_{m=1}^{N}\sum_{n \neq m}\rho^{mn}\sigma_{\eta^m}\sigma_{\eta^n} w_m w_n$$

- Since $E_{add}^k = \frac{1}{s}\sigma_{\eta_k}^2$, it can be rewritten as:

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

Uncorrelated and Unbiased Classifiers

- The expression of the added error reduces to:

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

- The optimal weights are inversely proportional to $E^k{}_{add}$:

$$w_k = \left(\sum_{m=1}^{N} \frac{1}{E_{add}^{m}} \right)^{-1} \frac{1}{E_{add}^{k}}, \quad E_{add}^{ave} = \frac{1}{1/E_{add}^{1} + 1/E_{add}^{2} + \ldots + 1/E_{add}^{N}}$$

- ➢ Simple average ($w_k = 1/N$) is optimal for classifiers with *balanced* (i.e. equal) errors
- ➢ Weighted average is required for *imbalanced* classifiers

Comparison between SA and WA

unbiased and uncorrelated errors

- The difference between the added error of SA and WA (using the optimal weights for WA) is:

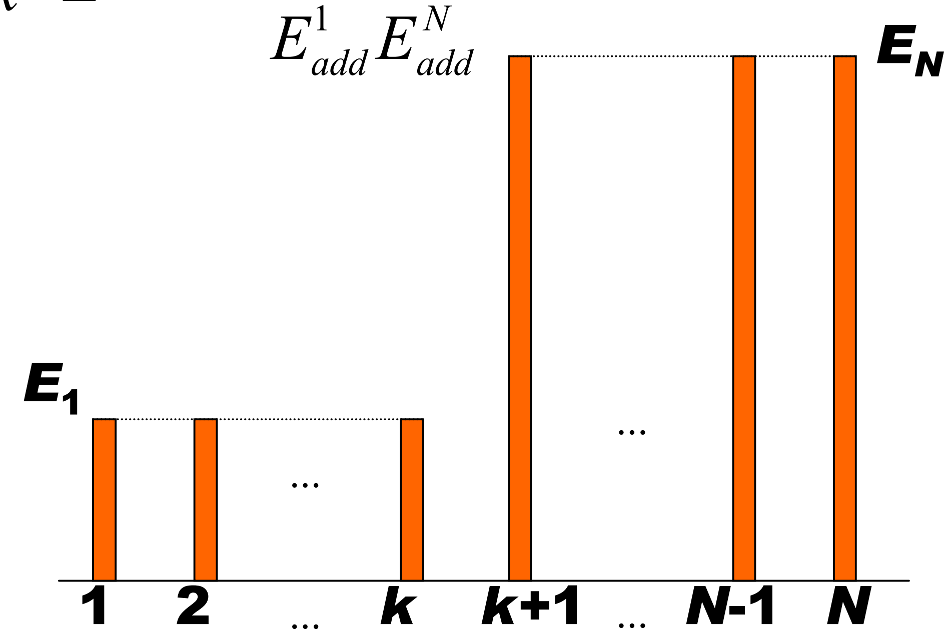
$$E^{SA} - E^{WA} = \frac{1}{N^2}\left(E_{add}^{1} + E_{add}^{2} + \ldots + E_{add}^{N}\right) - \frac{1}{1/E_{add}^{1} + 1/E_{add}^{2} + \ldots + 1/E_{add}^{N}}$$

- What is the “pattern” of classifiers’ errors that maximes the advantage of WA over SA ?
- Is such advantage depending only on the error “range”?

Upper bound of WA over SA

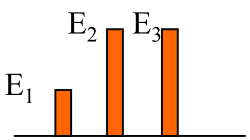
(Roli and Fumera, MCS 2002; manuscript in preparation)

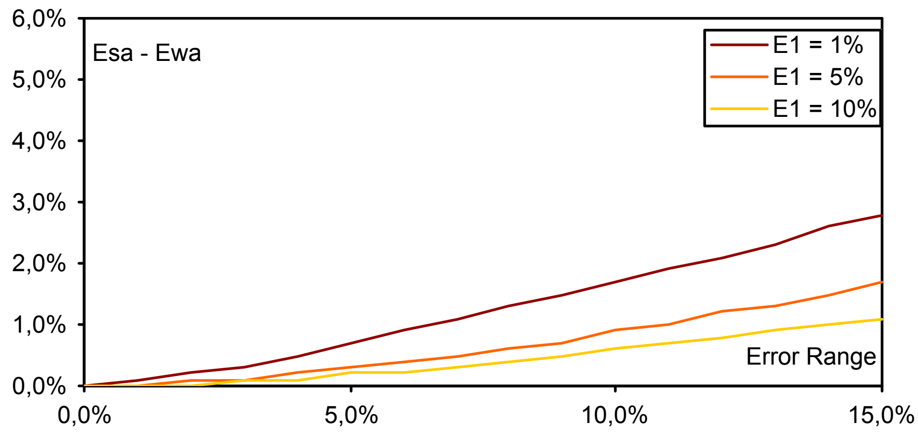
- For a given error range (E^N_{add} - E^1_{add}), the maximum of E^{SA} - E^{WA} is achieved when:
  - k classifiers have errors equal to E^1_{add}
  - N-k classifiers have errors equal to E^N_{add}
  - Where $k^* = \dfrac{NE^1_{add} - N\sqrt{E^N_{add}E^1_{add}}}{E^1_{add}E^N_{add}}$, and $k = \lfloor k^* \rfloor$ or $k = \lceil k^* \rceil$



E^{SA} - E^{WA} vs error range E_N - E_1

An example for N=3 uncorrelated classifiers





➢**Upper bound conditions: E_2= E_3**

•The advantage of WA over SA increases with the error range

•But it remains less than 3%

Weighted averaging of correlated classifiers

- The expected value of the added error is:

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

- For balanced performance and correlation, the optimal weights are $w_k = 1/N$, analogously to the uncorrelated case
- The value of E^{SA}-E^{WA} is affected by errors and correlations imbalance

- What are the conditions on errors and correlations that maximes the advantage of WA over SA ?

Weighted averaging of correlated classifiers

- For correlated classifiers, the optimal weights and the difference E^{SA}-E^{WA} cannot be computed analytically

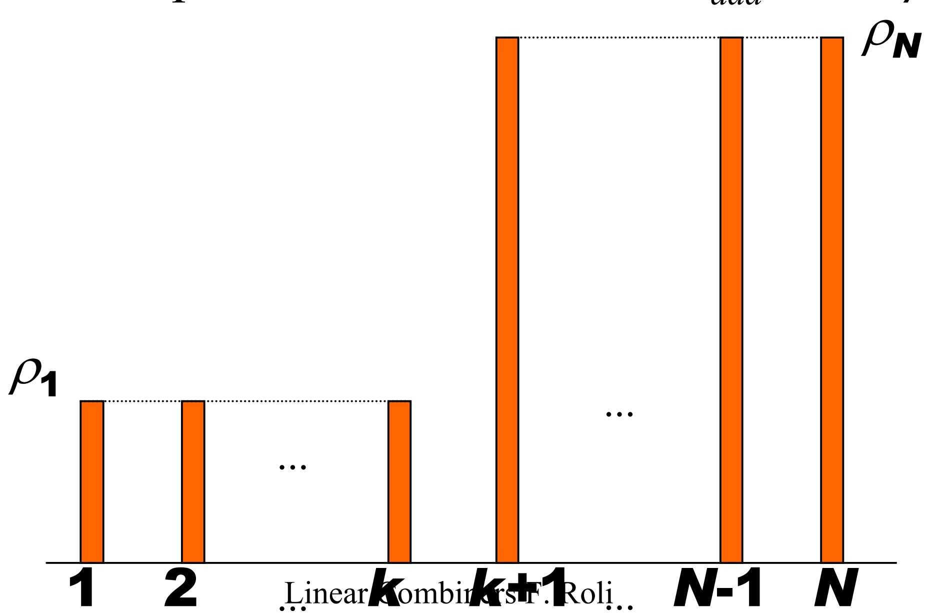
➢ The upped bound conditions for the difference E^{SA}-E^{WA} were searched by numerical analysis

➢ For different values of E^{k}_{add} and ρ^{mn}', the optimal w_k were computed by minimising E^{WA} by exhaustive search. The value of E^{SA}-E^{WA} was also computed by numerical analysis

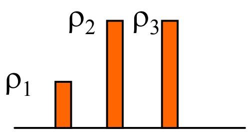
➢ This analysis also showed some effects of errors and correlations imbalance on the difference E^{SA}-E^{WA}

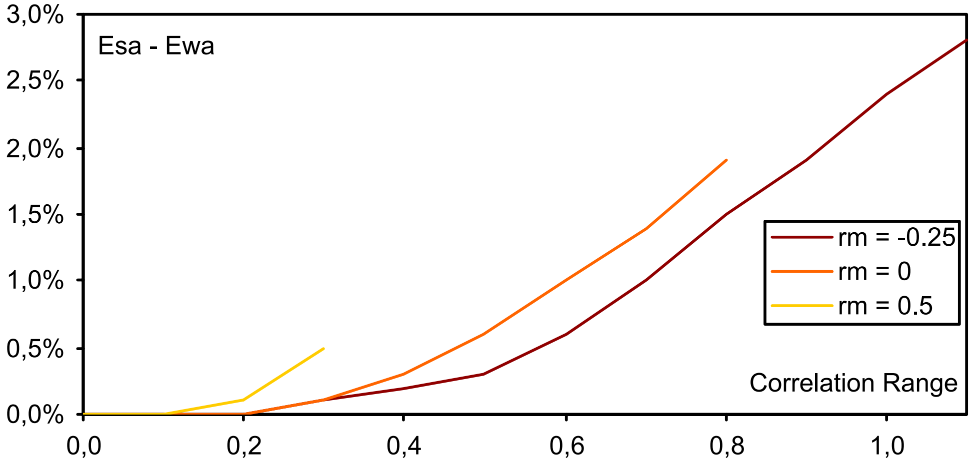
Balanced Errors and Imbalanced Correlations

- Numerical analysis was limited to the cases of N=3 and N=5 classifiers
- For a given correlation range (ρ^N - ρ^1) the maximum of E^{SA} - E^{WA} is achieved when:
  - k classifiers have correlations equal to $\min\{\rho^{mn}\} = \rho^1$
  - N-k classifiers have correlations equal to $\max\{\rho^{mn}\} = \rho^N$
  - The value of k depends on the values of E^k_{add}'s and ρ^{mn}'s



Balanced errors and Imbalanced correlations
An Example (N=3)





$rm=\rho^1$

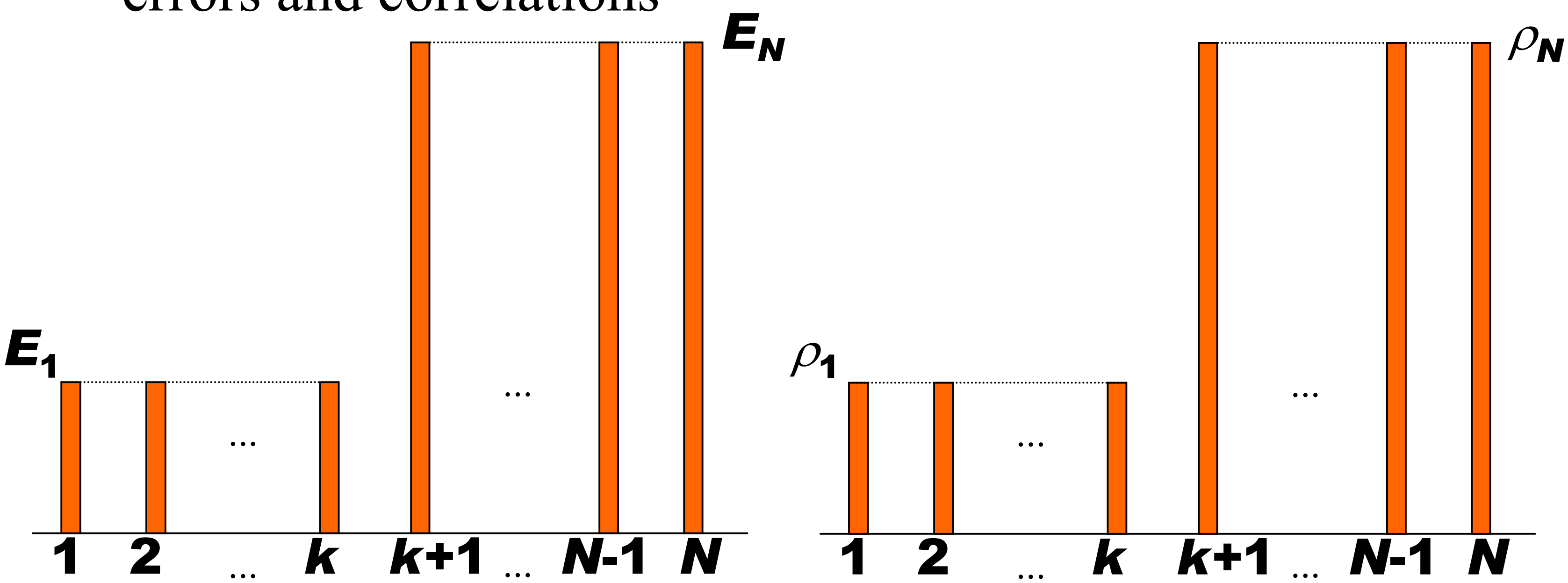
Individual classifiers errors = 10%

- It is worth noting that WA outperforms SA if classifiers have the same accuracy but different pair-wise correlations

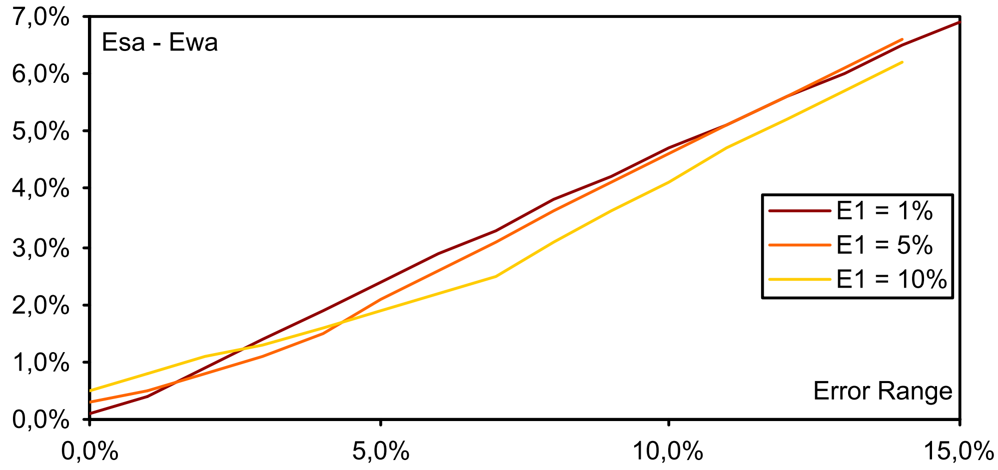
- SA suffers correlations imbalance

Imbalanced errors and correlations

- Numerical analysis showed that the maximum advantage of WA over SA is obtained for this case
- The upped bound conditions for the difference E^{SA}-E^{WA} is the conjunction of the conditions found for imbalanced errors and correlations



Imbalanced errors and Imbalanced correlations
An Example (N=3)



Correlation range: [-0.5; +0.9]

The advantage of WA over SA reaches 7%, while it was less than 3% for uncorrelated classifiers

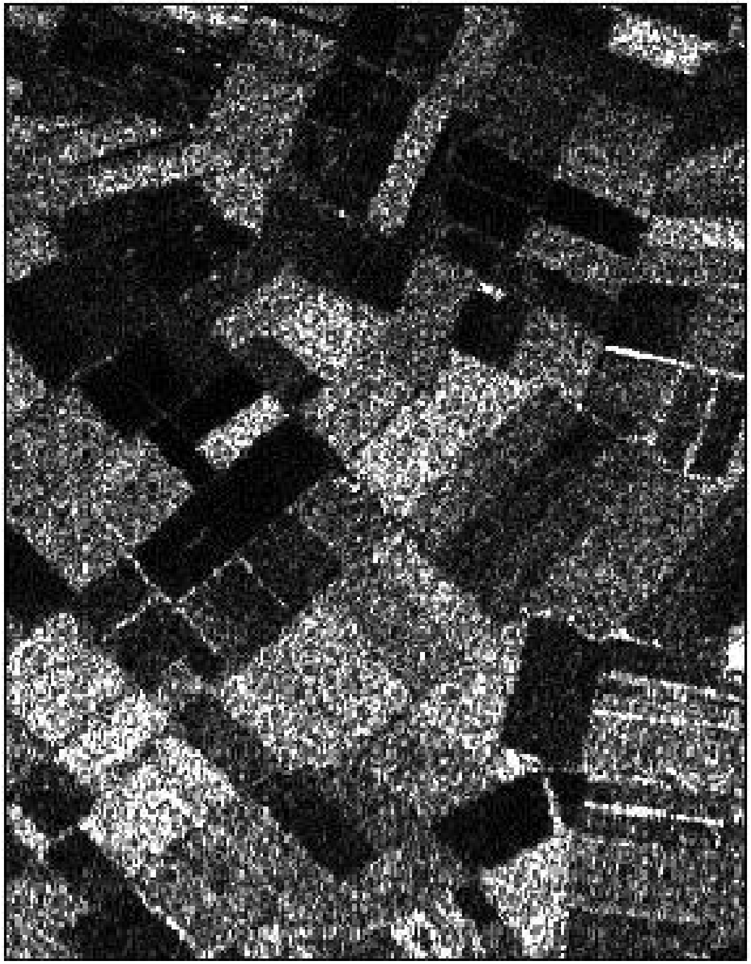
An example of experimental evidence: Remote sensing application

- Feltwell data set

PRL Electronics Annex

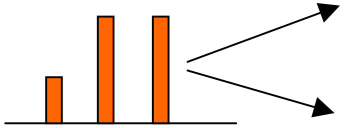
Vol. 21, 2000

  - five agricultural classes
  - fifteen features
    - 6 ATM and 9 SAR channels
  - training set: 5820 pixels
  - test set: 5124 pixels



Remote Sensing Application
(Roli and Fumera, MCS 2002)

- Test set error rates (averaged over ten runs)



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

	combiner error rates			optimal weights		
	E^{sa}	E^{wa}	$\mathbf{E^{sa}\text{-}E^{wa}}$	k-NN	MLP1	MLP2
Ensemble1	10.00	9.37	**0.63**	0.576	0.200	0.224
Ensemble2	12.09	9.69	**2.40**	0.689	**0.080**	0.231
Ensemble3	10.69	9.63	**1.06**	0.681	0.231	**0.088**
Ensemble4	16.81	9.79	**7.02**	0.838	**0.006**	0.156
Ensemble5	12.44	9.73	**2.71**	0.752	**0.103**	0.143

Remarks / Open Issues

- The comparison between WA and SA was focused on the upper bound conditions
- Lower bound conditions are matter of our on-going research
- The advantage of WA over SA for different ensembles can be evaluated if such ensembles have the same error ranges

Open issues:

- Advantage of WA over SA for ensembles with different error ranges
- Quantitative and general measures of imbalance degree

The imbalance concept

- In general, the concept of imbalanced classifiers is hard to be formally defined
- A "pattern" of imbalance that is useful for a fuser can hurt the performances of another fuser

- For linear combiners, this definition of imbalance can be given:

  two classifier ensembles exhibiting the same values of E^{SA}*-*E^{WA} *have the same degrees of imbalance*

Analysis of error-reject trade-off for linear combiners
(Roli et al., ICPR 2002)

- Roli et al. also extended the framework to the analysis of the error-reject trade-off
- We showed that the linear combination can improve the error-reject trade-off of individual classifiers

➢ In particular, we showed that linear combination can reduce the risk "added" to the Bayes one

MCS Workshops Series

http://www.diee.unica.it/mcs

- The series of workshops on Multiple Classifiers Systems, organized by the University of Cagliari and the University of Surrey, are motivated by the acknowledgment of the fundamental role of a common international forum for researchers of the diverse communities.
  - Multiple Classifier Systems 2000, Cagliari, Italy
  - Multiple Classifier Systems 2001, Cambridge, UK
  - Multiple Classifier Systems 2002, Cagliari, Italy
  - Multiple Classifier Systems 2003, Guildford, UK
- Join the discussion list mcs2002-discussion-list@diee.unica.it