REVSTAT – Statistical Journal
Volume 4, Number 3, November 2006, 201–225

COMBINING METHODS IN SUPERVISED CLASSIFICATION: A COMPARATIVE STUDY ON DISCRETE AND CONTINUOUS PROBLEMS

Authors: Isabel Brito
– Institut Curie, Service Bioinformatique, France
isabel.brito@curie.fr

Gilles Celeux
– INRIA Futurs, France
Gilles.Celeux@inria.fr

Ana Sousa Ferreira
– LEAD, FPCE, Universidade de Lisboa, Portugal
asferreira@fpce.ul.pt

Received: December 2005 Accepted: August 2006

Abstract:

- Often in discriminant analysis several models are estimated but based on some validation criterion, a single model is selected. In the purpose of taking profit from several potential models, *classification rules combining models* are considered in this article. More precisely two ways of combining models are considered: a serial combining method and a hierarchical combining method. Serial combining is a convex linear combination of a finite number of models. Hierarchical combining method leads to nested models structured in a binary tree. In this paper, several combining methods resorting from both points of view are presented and their performances are assessed on discrete and continuous classification problems.

Key-Words:

- *Gaussian classification; eigenvalue decomposition; multinomial classification; conditional independence model; convex combining; hierarchical combining.*

AMS Subject Classification:

- 62H30.

1. INTRODUCTION

In multivariate discriminant analysis, each object is assumed to arise from one of K exclusive groups $G_1, ..., G_K$ with prior probabilities $\pi_1, ..., \pi_K$, $\pi_k \geq 0$, $k = 1, ..., K$, $\sum_k \pi_k = 1$. Each object is characterised by a multivariate vector $\mathbf{x}$ of d variables. In this article, all d variables are assumed to be either continuous or discrete. The conditional density that $\mathbf{x}$ belongs to group G_k is denoted by $f_k(\mathbf{x})$. Accordingly to the discrete or continuous case, $f_k(\mathbf{x})$ is a probability or a density probability function which has to be estimated from a n-dimensional training sample $\mathbf{t}$ ($\mathbf{t}_i = (\mathbf{x}_i, z_i)$, $i = 1, ..., n$), where $\mathbf{x}_i$ is the d-dimensional vector measurement for unit i and $z_i \in \{1, ..., K\}$, denotes its group origin. Often, it is convenient to replace z_i with $\mathbf{y}_i$, a K-dimensional binary indicator vector of group membership for unit i: The k-th coordinate of $\mathbf{y}_i$ is 1 if i arises from group G_k and 0 otherwise.

The Bayes classifier assigns an individual vector $\mathbf{x}$ to G_g if

$$\pi_g \, f_g(\mathbf{x}) = \arg\max_k \, \pi_k \, f_k(\mathbf{x}) \; , \qquad k = 1, ..., K \, .$$

Usually, the group conditional probability function $f_k(\mathbf{x})$ is unknown and has to be estimated on the basis of the training sample $\mathbf{t}$. For continuous problems, the parametric paradigm is adopted and these functions are assumed to belong to a family of densities, in particular $f_k(\mathbf{x})$ are assumed to be d-normal with mean vector μ_k and covariance matrix Σ_d.

For discrete problems the most natural model is to assume that the group conditional probabilities $f_k(\mathbf{x})$ where $\mathbf{x} \in \{0, 1\}^d$ are multinomial probabilities. (For simplicity, the discrete variables are supposed to be binary variables.) In this case, the group conditional probabilities are estimated by the observed frequencies in the training set $\mathbf{t}$. Goldstein and Dillon [14] call this model the full multinomial model (FMM). One way to deal with the curse of dimensionality consists of reducing the number of parameters to be estimated. The first-order independence model (FOIM) assumes that the d binary variables are independent in each group G_k ([14]).

In many situations M different classifiers are in competition for the same problem and one of those classifiers is selected, based on some validation criterion. Acting in such a way, leads to reject several classifiers for which the parameters have been estimated. Besides, misclassified objects can be different for the different classifiers. Thus, those classifiers may contain useful information about the supervised classification problem, and this information is lost by selecting a unique classifier. The idea of combining models is present in a growing number of papers, hoping to obtain a more robust and more stable model than any of the competing models ([27], [35], [36], [4], [7], [20], [29] and [25] are examples of such papers).

The aim of this paper is to gather and extend combining methods previously presented ([9], [10], [32] and [34]) and to assess their performances from numerical comparisons on real data set.

In this paper, two ways of combining classifiers, called serial combination method and hierarchical combination method, are considered on the basis of numerical experiments on real data sets. For serial combination, a convex linear combination of M models is considered

$$\sum_m \mathbf{c}^m(\mathbf{x})\,\beta_m\ , \qquad \beta_m \geq 0,\ \ \sum_m \beta_m = 1,\ \ m=1,...,M\ , \tag{1.1}$$

where $\mathbf{c}_m(\mathbf{x})$ indicates the output of model m. Usually, this output is the group conditional probabilities functions $f_k^m(\mathbf{x})$, $k=1,...,K$, or the posterior probabilities $p_k^m(\mathbf{x})$

$$p_k^m(\mathbf{x}) = \frac{\pi_k\, f_k^m(\mathbf{x})}{\sum_g \pi_g f_g^m(\mathbf{x})}\ , \qquad g,k=1,...,K,\quad m=1,...,M\ , \tag{1.2}$$

or sometimes the membership estimation $z^m(\mathbf{x})$. To define the combining coefficients β_m, two strategies are possible: a single coefficient is associated to each model m (β_m is then a scalar) or K coefficients are associated to each model (β_m is then K-dimensional). The latter strategy can be thought of as attractive because it allows to choose a coefficient by model and by group. It means that it would be possible to weight differently the groups in the same combination of models. In fact, many numerical experiments on both real and simulated data ([33] and [10]) showed that this strategy produce awkward combining vectors. Moreover, in discrete problems, the training data sets are most often small in regard to the number of parameters to be estimated, and it is difficult to estimate several combining coefficients per model in a reliable way ([33]). A better strategy is to consider a single coefficient for each model. This strategy produces more stable and more interpretable combined models.

The methods that estimate a single coefficient per model are grouped according two different approaches based on least squares minimisation or on likelihood maximisation. In this work several methods have been considered according both approaches. Those methods are the committee of methods, which is a least squares minimisation technique and the other ones are based on likelihood ratios.

Hierarchical combining is different in spirit. It applies on polychotomous classification problems with $K>2$ groups and leads to nested models. Attention is focused on a method of combining models by a hierarchical coupling method related to an approach of Friedman [13]. This method is reducing the multigroup problem into several two-group problems. The hierarchical combined model is structured into a binary tree where each branch is associated to a model or a combination of models and a dichotomy between groups to be classified ([32], [34] and [9]).

The paper is organized as follows. In Section 2, the models in competition for both continuous and discrete classification problems are presented. In Section 3, the different convex combining strategies are described. Committee of methods and Likelihood ratios combining methods are presented in this section. Section 4 is devoted to the presentation of Hierarchical combining. Section 5 is concerned with the presentation of numerical experiments. The performances of combining models are compared on both discrete and continuous problems. For continuous data problems, serial and hierarchical combining methods are evaluated separately. Thus, when using hierarchical coupling, at each tree level only one model is chosen. For qualitative data problems, when using hierarchical combination at each node of the tree, a serial combination of models can be considered. Two sections, one about computer programs (Section 6) and another with a short discussion (Section 7) ends the paper.

2. CONTINUOUS AND DISCRETE CLASSIFIERS

In continuous supervised classification problems for assessing combining classification methods, the fourteen Gaussian models of EDDA ([3]) have been considered. Defined in the Gaussian setting, each group conditional probability function is supposed to be a d-dimensional Gaussian distribution with vector mean μ_k and covariance matrix Σ_k.

EDDA makes use of the variance matrix eigenvalue decomposition $\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ where $\lambda_k = |\Sigma_k|^{1/d}$, $\mathbf{D}_k$ is the eigenvector matrix of Σ_k and $\mathbf{A}_k$ is a diagonal matrix such that $|\mathbf{A}_k| = 1$, with the normalised eigenvalues of Σ_k on the diagonal in a decreasing order. This decomposition can lead to parsimonious and versatile models. Parameter λ_k denotes the volume of the k-th group, $\mathbf{A}_k$ its shape and $\mathbf{D}_k$ its orientation. Different assumptions on those parameters lead to fourteen models pooled into three families: eight elliptical models, four diagonal models and two spherical models. The eight elliptical models are

$$\left[\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T\right], \quad \left[\lambda_k \mathbf{D}\mathbf{A}\mathbf{D}^T\right], \quad \left[\lambda \mathbf{D}\mathbf{A}_k\mathbf{D}^T\right], \quad \left[\lambda_k \mathbf{D}\mathbf{A}_k\mathbf{D}^T\right],$$
$$\left[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\right], \quad \left[\lambda_k \mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\right], \quad \left[\lambda \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T\right], \quad \left[\lambda_k \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T\right].$$

The absence of subscript k means that the parameter at hand has a fixed value over the groups and its presence that the parameter is free over the groups. For instance, models $[\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T]$ and $[\lambda_k \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T]$ are respectively, the linear discriminant analysis model and the quadratic discriminant analysis model. Assuming that Σ_k are diagonal lead to the simplification $\Sigma_k = \lambda_k \mathbf{B}_k$, where $\mathbf{B}_k$ is a diagonal matrix where $|\mathbf{B}_k| = 1$. The four diagonal models are: $[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{B}_k]$, $[\lambda_k\mathbf{B}_k]$. The spherical models are $[\lambda\mathbf{I}]$, $[\lambda_k\mathbf{I}]$, $\mathbf{I}$ denoting the identity matrix. For each model, parameters $\mathbf{D}_k$, $\mathbf{A}_k$ or $\mathbf{B}_k$ and λ_k are estimated by maximizing the likelihood ([3]).

The output that has been considered for model m, in continuous combining context, is the posterior group probabilities $p_k^m(\mathbf{x})$ ([10]). In the hereunder considered examples, those probabilities have been derived by (1.2), where the prior probabilities π_k have been replaced with n_k/n (n_k is the number of units from G_k in the training set $\mathbf{t}$).

In discrete problems, only two reference models have been considered. They are the full multinomial model (FMM) and the first order independence model (FOIM). Those two models are expected to provide different classifiers in many circumstances. In the full multinomial model (FMM) the conditional probabilities are estimated with the observed frequencies

$$f_k(\mathbf{x}) = \frac{N(\mathbf{x} \mid k)}{n_k} \ , \qquad k=1, ..., K \ , \tag{2.1}$$

where $N(\mathbf{x} \mid k)$ is the number of observations of the training sample, belonging to G_k, for which state $\mathbf{x}$ occurs. This model involves 2^d-1 parameters in each group. Hence, even for moderate d, not all of the parameters are identifiable.

Since data sets are small or very small in regard to the number of probabilities to be estimated, a problem of sparseness is encountered and some of the multinomial cells may have no data in the training sets. Thus smoothing the observed frequencies is desirable. Hand [16] has noticed that the choice of the smoothing method is not very important so that computationally less demanding methods may be used. Thus the observed frequencies are smoothed using a single smoothing parameter λ $(0 < \lambda \leq 1)$ and the conditional densities takes the form (we omit the index k for simplicity)

$$f(\mathbf{x} \mid \lambda) \ = \ \frac{1}{n} \sum_i \lambda^{d-\|\mathbf{x}-\mathbf{x}_i\|} \, (1-\lambda)^{\|\mathbf{x}-\mathbf{x}_i\|} \ , \qquad i=1, ..., n \ . \tag{2.2}$$

When $\lambda = 1.00$ no smoothing is proceeded and the amount of smoothing is increasing as λ decreases to 0. This method will be called KERNEL in the sequel.

The first-order independence model (FOIM) assumes that the d binary variables are independent in each group G_k, $k=1, ..., K$. Then, the group probability function is of the form $\prod_j f(x_j \mid G_k)$, $j=1, ..., d$, and is estimated by

$$f_k^I(\mathbf{x}) \ = \ \prod_j \frac{N(x_j \mid k)}{n_k} \ , \tag{2.3}$$

where $n_k = \sharp G_k$ and $N(x_j \mid k) = \sharp\{y \in G_k \colon y_j = x_j\}$. In this model the number of parameters to be estimated for each group is reduced from $2^d - 1$ to d. This method is simple but may be unrealistic in some situations.

The resulting serial combining classifier is using a single coefficient, producing an intermediate model between the full multinomial model and the first order independence model. Combining methods differ in the way this coefficient is derived.

3. CONVEX COMBINING STRATEGIES

3.1. Committee of methods

A natural way of deriving the coefficients β_m in serial combining is minimising the fitting error using a least squares criterion. The committee of methods introduced by Bishop [4] in the neural computing literature is such an approach. In the committee of methods that will be considered here to get a relevant convex combining of classifiers, the fit of a classifier m is measured with the group classification probabilities, $\mathbf{c}^m(\mathbf{x})$. The committee of models is of the form

$$\mathbf{c}_{\mathrm{COM}}(\mathbf{x}) \;=\; \sum_m \mathbf{c}^m(\mathbf{x})\,\beta_m \;, \tag{3.1}$$

with $\beta_m > 0,\ m = 1, ..., M,$ and $\sum_m \beta_m = 1$. Writing $\mathbf{c}^m(\mathbf{x})$ as

$$\mathbf{c}^m(\mathbf{x}) \;=\; \mathbf{c}(\mathbf{x}) + \mathbf{e}^m(\mathbf{x}) \;, \tag{3.2}$$

where $\mathbf{c}(\mathbf{x})$ is the true group probabilities vector and $\mathbf{e}^m(\mathbf{x})$ represents the vector error of model m, leads to

$$\mathbf{c}_{\mathrm{COM}}(\mathbf{x}) \;=\; \mathbf{c}(\mathbf{x}) + \sum_m \mathbf{e}^m(\mathbf{x})\,\beta_m \;. \tag{3.3}$$

Defining $\mathbf{C}$ the error correlation matrix of the models whose general term is

$$\mathbf{C}_{ml} = E\big[e^m(\mathbf{X})\,e^l(\mathbf{X})\big] \;, \qquad m, l = 1, ...M \,, \tag{3.4}$$

E denoting the expectation under the true distribution of the training dataset, the committee of methods consists of minimizing the error $Er = \sum_m \sum_l \beta_m \beta_l \mathbf{C}_{ml}$ under the constraint that the positive coefficients β are summing to one. Using standard Lagrangian manipulation leads to

$$\beta_m \;=\; \frac{\sum_l (\mathbf{C}^{-1})_{ml}}{\sum_m \sum_l (\mathbf{C}^{-1})_{ml}} \;. \tag{3.5}$$

The correlation error matrix can be estimated by plug-in empirical values

$$\hat{\mathbf{C}}_{ml} \;=\; \frac{1}{n} \sum_i \big(\mathbf{y}_i - \mathbf{c}^m(\mathbf{x}_i)\big)\,\big(\mathbf{y}_i - \mathbf{c}^l(\mathbf{x}_i)\big)^T \,. \tag{3.6}$$

This formula means that in a natural way, the error vector $\mathbf{e}_m(\mathbf{x}_i)$ is estimated with

$$\hat{\mathbf{e}}_m(\mathbf{x}_i) \;=\; \Big(\hat{e}^k_m(\mathbf{x}_i) \;=\; y^k_i - c^k_m(\mathbf{x}_i)\Big) \,. \tag{3.7}$$

3.2. Likelihood ratios

LeBlanc and Tibshirani [20] presented an interesting combination method by likelihood ratios although they did not experiment it. It consists of choosing the combining coefficients as the ratio of the likelihood for model m over the sum of all models likelihoods,

$$\beta_m = \frac{L_m(\theta, \mathbf{x})}{\sum_l L_l(\theta, \mathbf{x})} , \tag{3.8}$$

where, recalling that y_{ik} is the k-th coordinate of the indicator vector giving the label of unit i,

$$L_m(\theta, \mathbf{x}) = \prod_i \prod_k \left[f_k^m(\mathbf{x}_i)\, \pi_k \right]^{y_{ik}} .$$

In the discrete case the single coefficient β is

$$\beta_m = \frac{L_I}{L_I + L_M} , \tag{3.9}$$

where L_I, L_M represents the likelihood for the FOIM and the FMM models, respectively.

Since the likelihood increases with the model complexity, this weighting strategy will favour more complex models. Thus, it could be preferable to propose penalized versions of likelihood ratios.

A natural penalisation is inspired from Akaike Information Criterion (AIC) ([1]). Denoting ν_m the number of independent parameters of model m, the AIC criterion is $\text{AIC} = -2\ln(L_m(\theta, \mathbf{x})) + 2\,\nu_m$ and it leads to the combining coefficients

$$\beta_m = \frac{L_m(\theta, \mathbf{x}) \exp\{-\nu_m\}}{\sum_l L_l(\theta, \mathbf{x}) \exp\{-\nu_l\}} . \tag{3.10}$$

In the discrete case, it takes the form

$$\beta_m = \frac{L_I \exp\{-Kd\}}{L_I \exp\{-Kd\} + L_M \exp\{-K(2^d - 1)\}} , \tag{3.11}$$

because Kd and $K(2^d - 1)$ are respectively the number of independent parameters for the FOIM and the FMM models.

Remark that in the discrete case, it appears that the likelihood ratio strategy derived from AIC leads always to a single coefficient with value one or zero and so this strategy is useless because it leads to a single model, FOIM or FMM (see [34]).

Another possibility, in the Bayesian model averaging spirit ([23] and [29]), is to base the combining weights on integrated likelihood ratios. The integrated or marginal likelihood for model m is

$$L(\mathbf{x} \,|\, m) \;=\; \int L_m(\theta, \mathbf{x})\, p(\theta_m)\, \mathrm{d}\theta_m \;, \tag{3.12}$$

where $p(\theta_m)$ is a prior probability distribution on θ_m.

Unfortunately, in most continuous cases, integral (3.12) is difficult to calculate. Kass and Wasserman [18] and Raftery [29] showed that integrated likelihood can be approximated using BIC criterion of Schwarz ([31]). This approximation leads to the combining coefficient for model m

$$\beta_m \;=\; \frac{L_m(\theta, \mathbf{x})\, n^{-0.5\,\nu_m}}{\sum L_l(\theta, \mathbf{x})\, n^{-0.5\,\nu_l}} \;. \tag{3.13}$$

In the discrete context, it is possible to get exact calculation of integral (3.12). In the non informative Bayesian setting, the prior distribution of FOIM parameters $p(a_j^k)$, $k=1, ..., K$, $j=1, ..., d$, are non informative Jeffreys distribution $\mathrm{B}(1/2, 1/2)$ and prior distribution of FMM parameters $p(b_h^k)$, $k=1, ..., K$, $h=1, ..., s$, where s is the number of states, is a non informative distribution of Jeffreys $\mathrm{D}(1/2, 1/2, ..., 1/2)$. From which, it follows directly that integrated likelihood for FOIM and FMM are

$$L_I(\mathbf{x}) \;=\; \frac{\prod\limits_k \prod\limits_j \mathrm{B}\big(x_k^j + 0.5\, n_k - x_k^j + 0.5\big)}{\mathrm{B}(0.5, 0.5)^{kd}} \;, \tag{3.14}$$

and

$$L_M(\mathbf{x}) \;=\; \frac{\Gamma(s/2)^k \prod\limits_k \prod\limits_h \Gamma\big(0.5 + c_k^h\big)}{\Gamma(1/2)^{ks} \prod\limits_k \Gamma\big(s/2 + n_k\big)} \;, \tag{3.15}$$

where c_k^h is the number of objects of group G_k with state h. And, the resulting combining coefficient β is estimated by

$$\beta \;=\; \frac{L_I(\mathbf{x})}{L_M(\mathbf{x}) + L_I(\mathbf{x})} \;. \tag{3.16}$$

4. HIERARCHICAL COMBINING

When the number of groups K to be discriminated is greater than two, as noted in Friedman [13], it can be advantageous to consider the polychotomous classification problem as a sequence of two group classification problem to get classifiers easier to be estimated and to be interpreted. Friedman proposed to

decompose the K groups in all possible combinations of pairs of groups. For each pair of groups, a classifier is designed. The overall classifier is derived from all the pairwise classifiers by a majority vote.

The strategy we now present is different. A polychotomous problem is decomposed into several dichotomous problems but the dichotomous problems are nested in a hierarchical binary tree. It is the reason why this strategy is called hierarchical coupling. Let $\mathcal{G} = \{G_1, ..., G_K\}$ be the set of groups. Consider a partition of $\mathcal{G}$ in two elements. At this level the best two class partition of groups is designed according to some criterion and the model or combination of models leading to the two class classifier minimizing the cross validated error rate between the two classes is designed. According to the sample size of the learning sample, leave one out or v-fold cross validation is considered. If available, it is also possible to assess the error rate with a test sample.

The procedure is repeated until all the elements in the actual partition are single groups. The combining classifier obtained from this hierarchical coupling procedure can be represented in a hierarchical tree as exemplified in Figure 1.

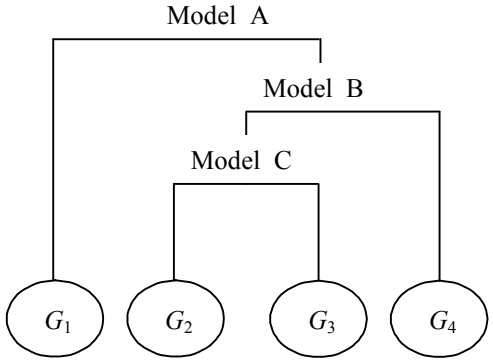


Figure 1: Example of hierarchical combined model for a four group problem.

The classifier depicted in Figure 1 is as follows. When a new observation is presented to the hierarchical classifier it passes through model A that classifies it in G_1 or $G_2 \cup G_3 \cup G_4$. If model A classifies the observation in G_1 the analysis is stopped. Otherwise, the observation passes through model B and the decision is G_4 or $G_2 \cup G3$. If model B does not classify the observation in G4 it passes finally through model C that assigns the observation to G_2 or G_3.

In order to choose, at each level, the best model or combination of models and the best partition, different strategies for continuous and discrete problems are employed.

In continuous data context, it was proceeded as follows:

1. For each possible binary partition all M models are estimated (at the beginning level there are $M(2^{K-1}-1)$ couples (model, partition)).

2. From those couples, the one providing the lowest misclassification error rate (ME) is chosen. In all the experiments, ME is evaluated by leave one out cross validation.

In the discrete case, the hierarchical coupling procedure is somewhat different.

1. At each level of the binary tree, the choice of the two-class decomposition of groups among the $2^{K-1}-1$ possible decomposition is done by minimizing the basic affinity coefficient ([24] and [2]) between the two classes of groups: Denoting $\mathrm{F}_1=\{p_j\}$ and $\mathrm{F}_2=\{q_j\}$, $j=1,...,d$, two discrete distributions defined on the same space, the affinity coefficient between F_1 and F_2 is given by $\rho(\mathrm{F}_1\mathrm{F}_2)=\sum_j \sqrt{p_j}\sqrt{q_j}$. Then the two classes of groups minimizing the affinity coefficient are selected.

2. After the two classes of groups have been chosen, the combining model is chosen by minimizing the error rate evaluated on a test sample or by v-fold cross validation.

Consider the example for a four group problem:

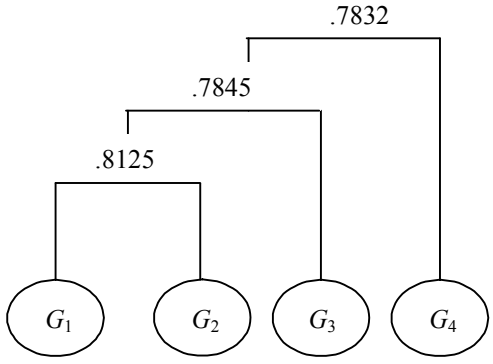


Figure 2: Example of hierarchical combined model for a four group discrete problem with the basic affinity coefficient values displayed.

It can be noticed that hierarchical combining method leads often to simple models at each step. From this point of view, it can lead to easily interpretable and stable decision rules, avoiding unnecessary complicated models.

5. RESULTS ON REAL DATA

In continuous context, combining methods have been applied on benchmark real datasets. Four of them were taken from the Machine Learning Repository of California University [5] (MLR), one from the Oxford University Repository [26] (OR) and another one from [15] (Hab). Table 1 provides a brief description of each dataset and their source.

Table 1: Continuous datasets description.

Dataset	Source	Description	Nb of units	Nb of features	Nb of groups
Bupa	MLR	Presence/absence of liver disorders that might arise from excessive alcohol consumption, measured by blood tests	345	6	2
Crabs	OR	Morphology of two species, blue and orange, by sex, of Australian crabs	200	5	4
Haberman	MLR	Survival of patients who had undergone surgery for breast cancer	306	3	2
Haemo	Hab	Presence of haemophilia on women	75	2	2
Iris	MLR	Measurements on the sepal and petal iris to determine iris specie (the famous Fisher dataset)	150	4	3
Thyroid	MLR	Medical records to predict the type of patients thyroidism	215	5	3

In discrete context, several real and simulated binary datasets were used to evaluate the performance of the considered strategies. Table 2 gives a brief description of each real dataset.

Table 2: Discrete datasets description.

Dataset	Source	Description		Nb of units	Nb of features	Nb of groups
Medical Data	[30]	Presence/absence of four symptoms liver disorders to predict the type of icterus		20	4	2
Psychological Data in older people	[11]	Scores obtained for each older adult in the six dimensions of the Psychological Well-Being Scale taken as binary data into two groups		80	6	2
Psychological Data	[28]	Six binary variables of a psychological test — Rorschach test — in 3 groups with different degrees of alexithymia		34	6	3
Psychological Counselling Career Data	[21]	Students of four licenciature's: Biology (B), Psychology (P), Language and Literature (LL), Engineering (E), described by the Psychological Questionnaire — My Vocational Situation — that is organised in three scales	Vocational Identity (VI) with 6 items	600	6	4
			Occupational Information (OI) with 4 items	600	4	4
			Barriers (B) with 4 items	600	4	4

5.1. Performance of serial combining techniques

The continuous case

Because several of the fourteen EDDA models lead to similar classifiers, combining all of them is useless. The more different models have been determined from the Correspondence Analysis of the fourteen models involved in EDDA described with their posterior densities $p_k^m(\mathbf{x})$ (see Brito [9]). For each dataset, the chosen models are given in Table 3.

Table 3: EDDA models chosen for each dataset by a Correspondence Analysis.

Dataset	Chosen models
Bupa	$[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{I}]$, $[\lambda_k\mathbf{I}]$
Crabs	$[\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T]$, $[\lambda\mathbf{I}]$
Haberman	$[\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T]$, $[\lambda\mathbf{B}_k]$, $[\lambda\mathbf{I}]$
Haemo	$[\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T]$, $[\lambda_k\mathbf{D}\mathbf{A}\mathbf{D}^T]$, $[\lambda\mathbf{I}]$
Iris	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$
Thyroid	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$

Serial combining methods were evaluated by leave-one-out cross validated misclassification error rate (ME). The purpose is to compare combining techniques opposite to single model techniques. In Tables 4 to 5, ME on each database are presented and compared with ME of model chosen with the standard EDDA strategy.

Table 4: Model and ME for each dataset using the committee of methods technique and EDDA.

Dataset	Committee of methods		EDDA	
	Model	ME	Model	ME
Bupa	$.79[\lambda\mathbf{B}] + .21[\lambda\mathbf{I}]$	.3971	$[\lambda\mathbf{B}]$	.4000
Crabs	$[\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T]$	.5000	$[\lambda\mathbf{I}]$	.0500
Haberman	$.4[\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T] + .6[\lambda\mathbf{I}]$	.2549	$[\lambda\mathbf{B}_k]$	.2516
Haemo	$[\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T]$	.1600	$[\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T]$	.1467
Iris	$.82[\lambda\mathbf{B}] + .18[\lambda\mathbf{I}]$	.0400	$[\lambda\mathbf{B}]$	.0400
Thyroid	$.73[\lambda\mathbf{B}] + .27[\lambda\mathbf{I}]$	.0930	$[\lambda\mathbf{B}]$	.0977

For **Bupa** and **Thyroid** datasets, misclassification error rate is slightly better using the committee of methods technique. **Bupa** dataset contains information on the presence or absence of liver disorders caused by excessive alcohol consumption. **Thyroid** dataset resumes medical records in order to predict patient type of thyroidism. In both cases, the diagonal model $[\lambda\mathbf{B}]$ is the model chosen with EDDA method. And, in both cases, the committee of methods technique proposes combining that model to the spherical model $[\lambda\mathbf{I}]$. The resulting shrunk model gives somewhat better predictions than the diagonal model alone.

Haberman dataset describes survival of women who had undergone surgery to remove breast cancer. **Haemo** illustrates the presence or absence of haemophilia on women. For those two datasets, EDDA strategy is slightly better than the application of committee of methods. In the other hand, for **Crabs** dataset which describes the morphology of males and females of two species of Australian crabs and for the famous Fisher dataset **Iris**, both EDDA and committee of methods lead to the same misclassification error. For two of the six examples, the **Crabs** and **Haemo** datasets, the committee of methods technique, lead to a single model, the linear discriminant analysis model, and for all other datasets a combination of models was selected.

Table 5: Model and ME for each dataset using the penalised likelihood ratios technique and EDDA.

Dataset	Penalised likelihood		EDDA	
	Model	ME	Model	ME
Bupa	$\left[\lambda \mathbf{B}_k\right]$	.4000	$\left[\lambda \mathbf{B}\right]$	.4000
Crabs	$\left[\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T\right]$	.5000	$\left[\lambda \mathbf{I}\right]$	.0500
Haberman	$\left[\lambda \mathbf{B}_k\right]$	.2516	$\left[\lambda \mathbf{B}_k\right]$	.2516
Haemo	$.32\left[\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T\right] + .68\left[\lambda_k \mathbf{D}\mathbf{A}\mathbf{D}^T\right]$	.1600	$\left[\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T\right]$	.1467
Iris	$\left[\lambda \mathbf{B}\right]$	.0400	$\left[\lambda \mathbf{B}\right]$	.0400
Thyroid	$\left[\lambda \mathbf{B}\right]$	.0977	$\left[\lambda \mathbf{B}\right]$	.0977

Using the penalised likelihood ratios technique did not produce improved performances on those datasets. The only case where it did not select a single model, for dataset **Haemo**, it provided a slightly higher misclassification error rate.

The discrete case

Since our samples are small the performance of the serial combining methods were evaluated by v-fold cross validation(ME). In Table 6, ME obtained on dataset **Psychological Data in older people** using the committee of methods technique and single models are compared. The performances of the classifiers have been assessed with half-sampling (two-fold cross validation error rate). Group prior probabilities were assumed to be equal, $\pi_k = .5 \ \ (k = 1, 2)$.

Table 6: Estimated error rate (half-sampling) and parameters values for the **Psychological Data in older people**.

	FOIM	FMM	KERNEL	C. MET.	C. MET.
Half-sampling	.30	.41	.32	.25	.25
λ		1.00	.95	1.00	.95
β				.555	.493

The goal of the present study is to explore the impact of playing with pets on psychological well-being among older people [11]. So, the two groups are constituted by 40 aged persons who have pets (group G_1) and 40 aged persons who don't have pets (group G_2).

Remark that this dataset is not very sparse ($2^6 = 64$ states and 80 observations) but, even so, the lowest error rate has been obtained with the committee of methods. The estimation obtained for β, through this strategy, is quite stable, producing a really intermediate model between the full multinomial model and the first order independence model. Also note that this approach seems to be no sensitive to the sparseness problem and so there is no need to smooth of the observed frequencies ($\lambda = 1$). On the basis of this study we can conclude that the involvement of playing with pets among older people can contribute for psychological well-being and thus, perhaps, for a successful ageing.

The numerical experiments performed for the model CMET on simulated binary data showed that good performances can be expected in a setting for which sample sizes are small or very small and population structures are identical in the two classes.

In Table 7, ME using the integrated likelihood ratio techniques and the single models have been compared on dataset **Medical Data**. In that case ME is the five-fold cross validation error rate of compared classifiers. Group prior probabilities were assumed to be equal, $\pi_k = .5$ $(k = 1, 2)$.

Table 7: Estimated error rate with five-fold cross-validation and parameters values for the **Medical Data**.

	FOIM	FMM	KERNEL	INT. LIK.	INT. LIK.
Five-fold cross-vali.	.45	.55	.55	.45	.45
λ		1.00	.95	1.00	.95
β				.832	.985

In this study, the goal is to predict the type of icterus, since it's not easy to make a diagnosis on the basis of liver disorders. Integrated likelihood ratio technique and FOIM provide the same performance for this dataset. The numerical experiments performed for this strategy on simulated binary data have shown that good performances can be expected with this technique in a moderate or large sample setting ([34]). In this small dataset setting (20 patients) it is no surprising that this method does not improve the performance since it involves the evaluation of an additional parameter β.

5.2. Assessing the performance of hierarchical combining

The continuous case

Hierarchical combining concerns only datasets with more than two groups. It has been assessed on **Crabs**, **Iris** and **Thyroid** datasets. All the fourteen models of EDDA were employed to get the hierarchical model. Hierarchical combining and EDDA methods are compared in Table 8.

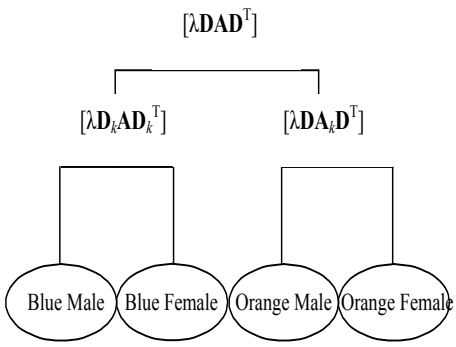
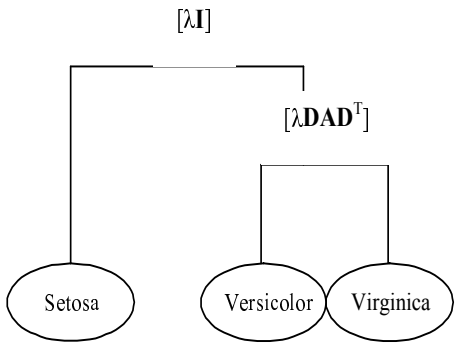
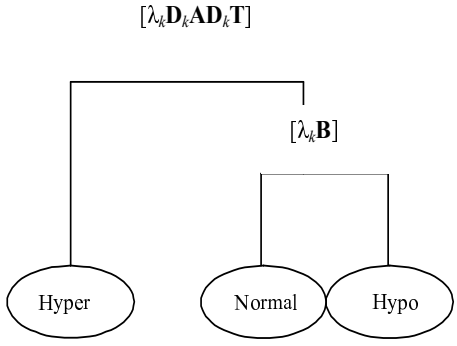
Hierarchical combining concerns only datasets with more than two groups. It has been assessed on **Crabs**, **Iris** and **Thyroid** datasets. All the fourteen models of EDDA were employed to get the hierarchical model. Hierarchical combining and EDDA methods are compared in Table 8. As it can be seen from Table 8, the classification error rates of hierarchical methods and EDDA are quite similar. Here the interest of hierarchical coupling lies essentially in its ability to choose different models at each step of the classification procedure. Thus it can provide more subtle and interpretable results. For instance, for **Iris** dataset, it shows at a glance that the Setosa group can be easily separated from the two other groups with the simplest model $\left[\lambda \mathbf{I}\right]$. On the contrary, for **Thyroid** dataset, it appears that separating the "hyper" group from the other groups needs a more complex model than separating the normal group from the "hypo" group.

Hierarchical coupling model for **Crabs** dataset is also appealing. At the first level, the linear model $\left[\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T\right]$ splits the Blue and Orange species. At the second level, males and females are separated inside each species. For Blue crabs, hierarchical coupling selects an elliptical model allowing for class of males and class of females to have different orientations $\left[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\right]$. For Orange Crabs, an elliptical model $\left[\lambda \mathbf{D}\mathbf{A}_k\mathbf{D}^T\right]$ is preferred which differentiates the shape of males and females classes.

In contrast with EDDA strategy which selects $\left[\lambda \mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\right]$ for separate the four groups, hierarchical coupling is less strict, proposing more adequate models

at the different levels. Only Blue males and females need the $\left[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T\right]$ model to be separated, less complex models being proposed to distinguish other groups.

Table 8: Model and ME for each dataset using the hierarchical coupling technique and EDDA.

Dataset	Hierarchical coupling		EDDA	
	Model	ME	Model	ME
Crabs		.045	$\left[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T\right]$	.045
Iris		.02	$\left[\lambda \mathbf{D} \mathbf{A} \mathbf{D}^T\right]$	.02
Thyroid		.0372	$\left[\lambda_k \mathbf{B}\right]$	.0326

The discrete case

For the **Psychological Data** the misclassification error is assessed by half-sampling. For the **Psychological Counselling Career Data** it is assessed from a test sample. A training sample of 200 students was drawn at random and the rest of the dataset constituted the test sample. Table 9 summarises the results of the four methods for these datasets and the coefficients of the combination obtained in each level of the tree.

The **Psychological Counselling Career Data** set consists of 600 students of the 1^{st} and 2^{nd} forms of four licenciature's degree: Biology (B), Psychology (P), Language and Literature (LL) and Engineering (E). The aim of the study is to know if those four groups of student are different regarding their Career Information.

For the **Psychological Counselling Career Data** the first decomposition chosen by hierarchical coupling for the several scales, suggest that Biology students are different from the other students in what concerns the definition of a clear and stable picture of their goals and interests, Engineering students revealing a distinct need for vocational information from the other students; and the students of odd groups show individually perceived external obstacles or limitations in pursuing occupational goals different from the students of even groups.

Remark that this dataset is not very sparse ($2^6 = 64$ or $2^4 = 16$ states and 200 observations), but again the hierarchical combining method using the integrated likelihood (HIER/IL) or committee of methods (HIER/CM) provides markedly the lowest misclassification error rate. The results of the hierarchical coupling provide markedly the lowest test estimates of the misclassification risk for all scales. However, HIER performs poorly for the Barriers scale.

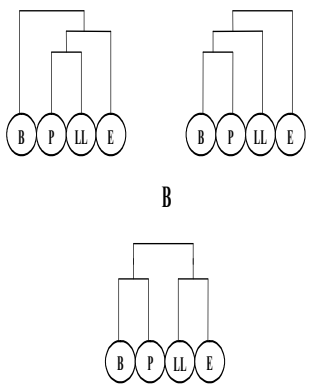
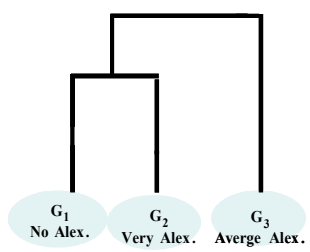
We noted that in some situations, particularly when the groups have very different sizes, usual methods and even the HIER method perform poorly. Moreover, the choice of the decomposition at each level of the tree may be unrealistic. Therefore, new developments on the hierarchical coupling approach are required in such a situation and this is a perspective for future research on this method.

The **Psychological Data** set consists of 34 dermatology's patients divided into three groups — Nonalexithymics Group (G_1), Alexithymics Group (G_2), Intermediate Group (G_3) — according to the value obtained in a psychological test (TAS-20: Twenty Item Toronto Alexithymia Scale) conceived to evaluate the presence of alexithymia[1]. The goal of the study is to evaluate how alexithymia influences personality characteristics (evaluated by another psychological test — Rorschach test).

[1] Alexithymia means "no words to express emotions".

For the **Psychological Data** the first decomposition chosen by hierarchical coupling, suggests that the union of the extremes groups forms a class well-separated from the intermediate group, since these subjects obtained balanced scores. Since the dataset is very sparse ($2^6 = 64$ states and only 17 observations) the hierarchical combining method using committee of methods (HIER/CM) provides the lowest estimated error rate.

Table 9: Model and ME for two datasets using the hierarchical coupling technique.

Dataset	Hierarchical coupling	Model	ME	λ	β		
					1st	2nd	3rd
Psychological Counselling Career Data		VI Scale					
		FOIM	.69	1			
		FMM	.75	1			
		KERNEL	.73	.99			
		HIER/CM	.49	1	.51	.52	.53
		HIER/CM	.49	.99	.47	.47	.48
		HIER/IL	.38	1	.98	.99	1
		HIER/IL	.38	.99	1	1	1
		OI Scale					
		FOIM	.66	1			
		FMM	.66	1			
		KERNEL	.65	.99			
		HIER/CM	.45	1	.50	.51	.52
		HIER/CM	.46	.99	.48	.49	.49
		HIER/IL	.41	1	0	≈ 0	≈ 0
		HIER/IL	.38	.99	0	.02	1
		B Scale					
		FOIM	.66	1			
		FMM	.66	1			
		KERNEL	.65	.99			
		HIER/CM	.50	1	.50	.52	.50
		HIER/CM	.50	.99	.49	.49	.49
		HIER/IL	.52	1	.99	.99	1
		HIER/IL	.52	.99	1	1	1
Psychological Data					1st	2nd	
		FOIM	.53	1			
		FMM	.71	1			
		KERNEL	.65	.99			
		HIER/CM	.29	1	.52	.55	
		HIER/CM	.29	.99	.47	.50	
		HIER/IL	.35	1	.18	.44	
		HIER/IL	.35	.99	.53	.78	

These results are in accordance with the numerical experiments performed for CM and IL strategies on simulated binary data that have shown that good performances can be expected with CM technique in a small or very small sample setting and with IL technique in moderate or large sample setting.

6. COMPUTER PROGRAMS

The efficiency of the combining approaches presented in this paper has been investigated on both real and simulated data. The computer programs realizing these combining approaches were implemented by the authors and are available from them.

The continuous case

All computer programs for the continuous case are written in Matlab® code. The different routines are structured as follows:

- **EDDA** — estimates all EDDA models and the leave-one-out cross validated misclassification error of each model;
- **COMMITTEE** — estimates the serial combined model by a committee of methods strategy;
- **SERIAL** — estimates the serial combined model by a penalized likelihood strategy;
- **HIERARCHICAL** — evaluates the combination of the models for all possible two class of groups. It calculates the leave-one-out cross validated misclassification error of each solution and builds the tree representation.

Run time execution is about five time more important for hierarchical coupling method than for serial combining method. It means that, for most applications, it remains a reasonable method.

The discrete case

The computer programs implemented for the discrete case use FORTRAN® 77 Language according to Microsoft FORTRAN Optimizing Compiler Version 5.0 and they use a structure in three main routines:

- **GESTAO** — determines the group conditional probabilities associated to the full multinomial model (FMM) and to the first-order independence model (FOIM) and their estimative by cross validation;
- **CALFA** — determines the combining coefficient according to the chosen combining strategy;
- **CRULE** — builds the new combining model and determines the error rate evaluated on a test sample.

For the **hierarchical combining**, an additional routine is implemented:

- **HIERQ** — builds the hierarchical binary tree, using the basic affinity coefficient.

After the selection of the two classes of groups have been chosen at each level of the binary tree, the combining model is chosen by minimizing the error rate evaluated on a test sample, using routines **GESTAO**, **CALFA** and **CRULE**.

Finally, it can be noticed that the run time execution for the hierarchical combining is quite similar to that of the serial combining in the $K=3$ group case. Otherwise, when $K>3$, the run time execution for the hierarchical combining triplicate or even more, due to the necessary reorganization of the groups for the evaluation of the basic affinity coefficient for all possible combination of couples of groups. However, the computational time for hierarchical combining remains quite reasonable and cannot be regarded as a drawback of this approach.

7. DISCUSSION

It is worth noticing that the combining methods that were considered in this paper are of different nature than other combining or ensemble methods. For example, Bagging and Boosting methods which are very efficient to improve unstable classifiers are committee-based approaches in which a single classification algorithm is applied to repeatedly modified versions of the data ([7], [8], [12], [17]-chapter 10). On the contrary the combing methods we considered are combining several methods but do not modified the weights of the data. On an other hand, the CRUISE ([19]) and QUEST ([22]) methods are classification tree algorithms different of the hierarchical combining methods we considered because the tree we designed is not a classification tree.

Many combining methods of classification have been considered in different contexts from a practical point of view. The main conclusions of this comparative experimental study are the following. Convex combining appears to be disappointing in the continuous case. In that case, at best, they lead to the same

error rate obtained with the better single model. Moreover, they often prefer a single model to a combination of several models. Convex combining appears to be more efficient to propose a good compromise between FMM and FOIM models in discrete data context. Maybe the reason for this more satisfactory behaviour is that FMM and FOIM are quite different models.

On the contrary hierarchical coupling seems to be a promising technique of combining classification methods when more than two groups are to be classified. In different contexts, hierarchical coupling leads to a substantial improvement of the misclassification error rate and its easily interpretable representation is appealing. It provides original and parsimonious classification rules. An interesting perspective would be to explore all possible hierarchical coupling solutions. This is feasible when the number of groups is less than five. Otherwise, a *branch and bound* algorithm could be considered in order to search for the optimal tree solution in a reasonable time.

Finally, it can be noticed that there is a huge literature on combining models. For instance Bayesian Model Averaging (BMA) (see [23] or [29], among many others) has received a lot of attention. However, the practical implementation of Bayesian Model Averaging is far from being simple especially in the continuous case. Finally, we want to cite the interesting theoretical study of Yang ([37]) which proves that combining models cannot be expected to outperform an optimal single method for large samples.

REFERENCES

[1] Akaike, H. (1974). A new look at statistical model identification, *IEEE Transactions on Automatic Control*, **AU-19**, 716–722.

[2] Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis, *Methods on Operations Research*, **53**, 507–512.

[3] Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition, *Journal of The American Statistical Association*, **91**, 716–722.

[4] Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press London.

[5] Blake, C. and Merz, C. (1998). *UCI repository of machine learning databases*, http://www.ics.uci.edu/~mlearn/MLRepository.html, Dept. of Information and Computer Sciences Irvine University of California.

[6] Breiman, L. (1995). Stacked Regression, *Machine Learning*, **24**, 49–64.

[7] Breiman, L. (1996). Bagging predictors, *Machine Learning*, **26**(2), 123–140.

[8] Breiman, L. (1998). Arcing classifiers, *The Annals of Statistics*, **26**, 801–849.

[9] Brito, I. and Celeux, G. (2000). *Discriminant analysis by hierarchical coupling in EDDA context*. In "Proceedings of the 7th Conference of the International Federation of Classification Societies, IFCS-2000" (J. Jansen, Ed.), Springer-Verlag.

[10] Brito, I. (2002). *Combinaison de modles en analyse discriminante dans un contexte gaussien*, PhD Thesis, Université Joseph Fourier, Grenoble.

[11] Dourado, S.; Mohan, R.; Vieira, A.; Sousa Ferreira, A. and Duarte Silva, M.E. (2003). *Animais de estimação e bem-estar em idosos*. In "Resumos do V Simpósio Nacional de Investigação em Psicologia", Edições Associação Portuguesa de Psicologia.

[12] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**(1), 119–139.

[13] Friedman, J.H. (1996). *Another Approach to Polychotomous Classification*, Technical Report, Stanford University.

[14] Goldstein, M. and Dillon, W. (1978). *Discrete Discriminant Analysis*, Wiley and Sons.

[15] Habbema, J.; Hermans, J. and van den Broek, K. (1974). *A stepwise discriminant analysis program using density estimation*. In "Proceedings of Computational Statistics, Compstat 1974", Physica-Verlag, 101–110.

[16] Hand, D. (1982). *Kernel Discriminant Analysis*, Research Studies Press, Wiley.

[17] Hastie, T.; Tibshirani, R. and Friedman, J. (2000). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag.

[18] Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses with large samples, *Journal of The American Statistical Association*, **90**, 928–934.

[19] Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits, *Journal of The American Statistical Association*, **96**, 589–604.

[20] LeBlanc, M. and Tibshirani, R. (1996). Combining Estimates in Regression and Classification, *Journal of The American Statistical Association*, **91**, 1641–1650.

[21] Lima, M.R. (1998). *Orientação e Desenvolvimento da Carreira em Estudantes Universitários*, PhD Thesis (in Portuguese), Univ. de Lisboa.

[22] Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica*, **7**, 814–840.

[23] Madigan, D.; Raftery, A. and Hoeting, J. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of The American Statistical Association*, **89**, 1535–1549.

[24] Matusita, K. (1955). Decision rules based on distances for problems of fit, two samples and estimation, *Annals of the Institute of Statistical Mathematics*, **26**(4), 631–640.

[25] Merz, C. and Pazzani, M. (1999). A principal component approach to combining regressions estimates, *Machine Learning*, **36**, 9–32.

[26] Oxford University Machine Learning Repository Statlib (1996). http://lib.stat.cmu.edu, Depart. of Statistics at Carnegie Mellon University.

[27] Perrone, M. and Cooper, L. (1973). *When networks disagree: ensemble methods for hybrid neural networks.* In "Artificial neural networks for speech and vision" (R. Mammone, Ed.), Chapman & Hall, 126–142.

[28] Prazeres, N.L. (1996). *Ensaio de um Estudo sobre Alexitimia com o Rorschach e a Escala de Alexitimia de Toronto (TAS-20)*, Master Thesis (in Portuguese), Univ. de Lisboa.

[29] Raftery, A. (1996). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models, *Biometrika*, **83**, 251–266.

[30] Romeder, J. (1973). *Méthodes et Programmes d'Analyse Discriminante*, Dunod, Paris.

[31] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

[32] Sousa Ferreira, A.; Celeux, G. and Bacelar-Nicolau, H. (1999). *Combining Models in Discrete Discriminant Analysis by a Hierarchical Coupling Approach.* In "Applied Stochastic Models and Data Analysis, ASMDA 99" (H. Bacelar-Nicolau, F. Costa Nicolau, J. Janssen, Eds.), INE, 159–164.

[33] Sousa Ferreira, A. (2000). *Combining Models in Discrete Discriminant Analysis*, PhD Thesis (in Portuguese), Univ. Nova de Lisboa.

[34] Sousa Ferreira, A.; Celeux, G. and Bacelar-Nicolau, H. (2000). *Discrete Discriminant Analysis: the Performance of Combining Models by a Hierarchical Coupling Approach.* In "Data Analysis, Classification and Related Methods" (Kiers, Rasson, Groenen, Schader, Eds.), Springer, 181–186.

[35] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society*, **B 36**, 111–147.

[36] Wolpert, D. (1992). Stacked generalization, *Neural Networks*, **5**, 241–259.

[37] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, **92**, 937–950.