Combining Models in Discrete Discriminant Analysis

Anabela Marques[1] Ana Sousa Ferreira[2] Margarida Cardoso[3]

Abstract. When conducting Discrete Discriminant Analysis, alternative models provide different levels of predictive accuracy which has encouraged the research in combined models. This research seems to be specially promising when small or moderate sized samples are considered, which often occurs in practice. In this work we evaluate the performance of a linear combination of two Discrete Discriminant Analysis models: the First-order Independence Model and the Dependence Trees Model. The proposed methodology also uses a Hierarchical Coupling Model when addressing multi-class classification problems, decomposing the multi-class problems into several bi-class problems, using a binary tree structure. The analysis is based both on simulated and real data sets. Results of the proposed approach are compared with those obtained by Random Forests, being generally more accurate. Measures of precision regarding a training set, a test set and cross-validation are presented. The R software is used for the algorithms' implementation.

Keywords Combining models; Dependence Trees model; Discrete Discriminant Analysis; First-order Independence model; Hierarchical Coupling model.

1 Introduction

Discrete Discriminant Analysis (DDA) is a multivariate data analysis technique that aims to classify multivariate observations of discrete variables into one of K *a priori* defined classes.

In DDA a n-dimensional sample of multivariate observations is considered $X = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_n)$, where $\underline{x}_i$ represents the i^{th} observation ($i \in \{1, ..., n\}$), described by P discrete variables, $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{iP})$. The class of each observation - one of K exclusive classes $(C_1, C_2, ..., C_K)$ - is assumed to be known and the

[1]Barreiro School of Technology of Polytechnic of Setúbal, Portugal.
E-mail:anabela.marques@estbarreiro.ips.pt
[2]LEAD, Faculty of Psychology, University of Lisbon, Portugal, CEAUL and UNIDE.
E-mail:asferreira@fp.ul.pt
[3]Dep. of Quantitative Methods of ISCTE - Lisbon University Institute, Portugal and UNIDE.
E-mail:margarida.cardoso@iscte.pt

corresponding prior probabilities are π_k, $k = 1, ..., K$, $\sum_{k=1}^{K} \pi_k = 1$.

DDA has two main goals:

1. To identify the variables that best differentiate the K classes;
2. To assign objects whose class membership is unknown to one of the K classes, by means of a classification rule.

This work is focused on the second goal and we consider objects characterized by binary variables, in the bi-class and in the multi-class case. Note that for P binary variables there are $S = 2^P$ possible states (i.e. $S = 2^P$ possible observable vectors).
To derive the classification rule, based on the referred data, one should determine the posterior probability of an observation. Based on the Bayes formula the posterior probability of an observation - $\underline{x}^*$ - being assigned to one of the *a priori* known classes can be written as follows:

$$P\left(\underline{x}^* \in C_k | X, \underline{\pi}\right) = \frac{\pi_k f_k(\underline{x}^* | X)}{\sum_{k=1}^{K} \pi_k f_k(\underline{x}^* | X)} \ , \quad k = 1, \ldots, K \tag{1}$$

where π_k represents the priori probability of class C_k and $f_k(\underline{x})$ represents the probability function of $\underline{x}$ in the same class. By applying this rule, an observation $\underline{x}^*$ is classified in the class with the maximum posterior probability, thus minimizing the assignment error.
The prior probabilities π_k, often have to be estimated using the sample at hand. When this sample is randomly selected from the population without taking into account the observations class membership, maximum likelihood estimators are used: $\pi_k = \frac{n_k}{n}$, where n_k is the dimension of class C_k. Otherwise, if the sample considered is the union of K independent samples of size n_k, $k = 1, ..., K$, previously selected within each class C_k, equal prior probabilities are considered for all classes, $\pi_k = \frac{1}{K}$. Usually, the states probability function in each class C_k is unknown and must be estimated using the sample observations X.

In DDA, the multinomial model is considered the most natural model where the states probability functions are estimated by the corresponding sample relative frequencies. This is the so called Full Multinomial Model (FMM) that demands a large number of parameters to be estimated (Goldstein and Dillon, 1978).
To overcome this dimensionality problem, several variants of the FMM model have been proposed. In this study, we work with two specific FMM variants - the First-order Independence Model (FOIM) (Goldstein and Dillon, 1978), which assumes that the P discrete variables are independent within each class C_k - and an alternative model that takes into account the dependence between variables - the Dependence Trees Model (DTM) (Celeux and Nakache, 1994).

In real classification problems, the classification errors resulting from different models differ and are often associated with different subjects. Therefore, researchers derive and compare several classification rules and recur to multiple

models. The use of multiple models generally enhances the results accuracy. These models may originate from diverse subsamples drawn from an original dataset: e.g. Breiman (1996) uses the bagging strategy and Friedman (2001) uses the boosting strategy for drawing the successive subsamples. As an alternative approach, when considering a fixed dataset, multiple models may result from different parameterizations of a specific model type (e.g. a tree model with different numbers of levels) or diverse types of models may be considered.
In this context the analyst often selects the classification rule that provides the best classification accuracy. However, the selection of a single classification rule means a high loss of information of the previously estimated models which could be very relevant for classification. In fact, the classification results may be provided by a combination of models overcoming the referred loss of information and enhancing classification results stability and accuracy, e.g. Friedman and Popescu (2008).
Several combined methods can be found in the literature. Recently, (Kotsiantis, 2011), for example, proposed a combined model for classification - Random Subspace using Naïve Bayes (Domingos and Pazzani, 1997) and C4.5 (Quinlan, 1993). Based on 26 well known data sets (with continuous predictors), the author found the results of the proposed method encouraging. However, most studies - (Kotsiantis, 2011) reviews several - refer to Discriminant Analysis in general - DDA studies are rare.
In the present work, we address DDA problems considering a simple linear combination of FOIM and DTM (Marques et al., 2013) and assess its performance in numerical experiments based on real and simulated data sets. In order to deal with multi-class problems, the Hierarchical Coupling Model that decomposes the original multi-class problem in several bi-class problems, using a binary tree structure, is also considered, (Sousa Ferreira et al., 2000).
We compare the performance of the proposed combined model - a non-generative ensemble according to (Re and Valentini, 2011) - with the performance of Random Forests (Breiman, 2001) - a generative ensemble (according to the same authors), that generates sets of base learners acting on the structure of the data set to try to actively improve diversity and accuracy of the base learners. According to (Kotsiantis, 2013, p.278): "Random forests (Breiman, 2001) are one of the best performing methods for constructing Ensembles". In addition, Random Forests tend to perform better when dealing with discrete categorical features (Kotsiantis et al., 2006).
The new DDA approach is presented in the second chapter after introducing the models FOIM and DTM. In the third chapter, the performance of the new model is analyzed, based both on simulated and real data sets, with small and moderate sizes. Finally, conclusions are drawn and perspectives of future work are indicated.

2 Methodological approach

2.1 Discrete Discriminant Analysis

In Discrete Discriminant Analysis the most usual classification rule is based on the Full Multinomial Model (FMM) (Goldstein and Dillon, 1978; Celeux and Nakache, 1994) where the within-classes states probability functions are multi-

nomial. However, for the case where we have P binary variables, this model involves the estimation of 2^{P-1} parameters in each class. Therefore this approach needs to rely on large samples which can be very difficult to obtain in some application domains, such as health sciences and psychology.
As previously referred, the FOIM model assumes the independence of variables within each class therefore reducing the number of parameters to estimate. However, this model may be unrealistic in some situations. Among alternative models that take into account the interactions between variables the Dependence Trees Model (DTM) can be considered, (Celeux and Nakache, 1994).
These models, FOIM and DTM, are described next.

2.2 The First-order Independence Model

The First-order Independence Model - FOIM - (Goldstein and Dillon, 1978; Celeux and Nakache, 1994) is one of the most commonly used DDA models. It assumes that the P discrete variables are independent within each class C_k, reducing to P the number of parameters needed to be estimated for each class C_k.
The condicional probability of assigning $\underline{x}^*$ to class C_k is estimated by:

$$\hat{f}_k\left(\underline{x}^* \mid X\right) = \prod_{p=1}^{P} \frac{\#\left\{\underline{x}_j \in C_k : x_{jp} = x_p^*\right\}}{n_k}, \qquad j = 1, \ldots, n \ ; \ k = 1, \ldots, K \quad (2)$$

where n_k represents the C_k class sample dimension.

2.3 The Dependence Trees Model

The Dependence Trees Model - DTM - (Celeux and Nakache, 1994; Pearl, 1988), takes into account conditional dependence relationships between the predictors. DTM provides for each class an estimate of the conditional probability functions based on the idea proposed by Pearl (1988). Pearl demonstrated that through the knowledge of a graph G, where $X_1, ..., X_P$ represent its P vertices, the probability distribution f^G, associated with this graph, can be calculated as the product of the conditional probabilities:

$$f^G(x_1, ..., x_P) = f(x_{r(p)}) \prod_{l(p)=1}^{P-1} f\left(x_p \mid x_{l(p)}\right) \quad (3)$$

where $x_{l(p)}$ represents a variable that is linked to the variable x_p in this graph, arbitrarily choosing one vertex as the root of the graph, $x_{r(p)}$.
To construct the graph for each class, we rely on the algorithm of Chow and Liu (Celeux and Nakache, 1994; Pearl, 1988), where the length of each edge referred to the pair of variables $(x_p, x_{p'})$ represents a measure of the association between the same variables, mutual information in particular. Mutual information - I - is defined as follows:

$$I(X_p, X_{p'}) = \sum \sum f(x_p, x_{p'}) \log \frac{f(x_p, x_{p'})}{f(x_p) f(x_{p'})} \quad (4)$$

where $f(x_p, x_{p'})$ is estimated using the maximum-likelihood approach.

After the calculation of the C_2^P mutual information values, the graph G, with $P-1$ edges, corresponding to the highest total mutual information is selected. For example, take $P = 5$ variables and if the most important predictor relations are $(X_2, X_1), (X_3, X_2), (X_4, X_2)$ and (X_5, X_2), then Figure 1 represents an example of a dependence tree

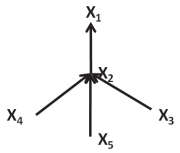


Figura 1: Example of a dependence tree for the case of P=5 variables

and the probability distribution of the first-order dependence tree is

$$\hat{f}_k\left(\underline{x}^*|X\right) = f^{C_k}\left(\underline{x}^*|X\right) = \hat{f}(x_1^*|X)\hat{f}(x_2^*|x_1^*, X)\hat{f}(x_3^*|x_2^*, X)\hat{f}(x_4^*|x_2^*, X)\hat{f}(x_5^*|x_2^*, X) \tag{5}$$

where the marginal and conditional probability functions are determined simply using the observed relative frequencies in sample X.

2.4 Combining Models

The idea of combining different models currently appears in a increasing number of papers, aiming to obtain more robust and stable models - e.g. Leblanc and Tibshirani, 1996; Opitz and Maclin, 1999; Wang et al., 2000; Sousa Ferreira et al., 2004; Brito et al., 2006; Chrysostomou et al., 2008; Kotsiantis, 2011; Marques et al., 2013.

The present study develops from the contribution of Sousa Ferreira (2004) that combines FMM and FOIM, using a single coefficient β, $(0 \leq \beta \leq 1)$ to define a linear combination and explores several strategies to estimate this coefficient, including a regression approach using least squares minimization and likelihood maximization. This approach reveals good performances, with intermediates results between FOIM and FMM, in the small case setting - particularly when data have independent structures in each class, or equal correlation structures. Using an integrated likelihood ratio approach, interesting results are also observed, particularly in the moderate or large case settings and when data have different correlation structures in each class. However, in this FOIM-FMM combination, the coefficient derived often tends to heavily weight FOIM, while reducing substantially the contribution of FMM, even when considering smoothed frequencies. Based on this empirical conclusion, we consider the replacement of FMM, in the combination, by DTM. The corresponding conditional probability function is thus estimated as follows:

$$\hat{P}\left(\underline{x}^* \in C_k|\beta, X\right) = \beta\hat{P}_{FOIM}(\underline{x}^* \in C_k|X) + (1-\beta)\hat{P}_{DTM}(\underline{x}^* \in C_k|X) \tag{6}$$

The performance of the FOIM-DTM linear convex combination is the focus of the present paper. In addition, we consider the performance of the Hierarchical Coupling Model (Sousa Ferreira et al., 2000) integrating this specific combination.

2.5 The Hierarchical Coupling Model

In the multi-class case, the Hierarchical Coupling Model - HIERM - (Sousa Ferreira et al., 2000) may be considered as an alternative to the simple FOIM-DTM convex combination.
HIERM decomposes one multi-class problem into several bi-class problems using a binary tree structure and implements two decisions at each level of the tree:

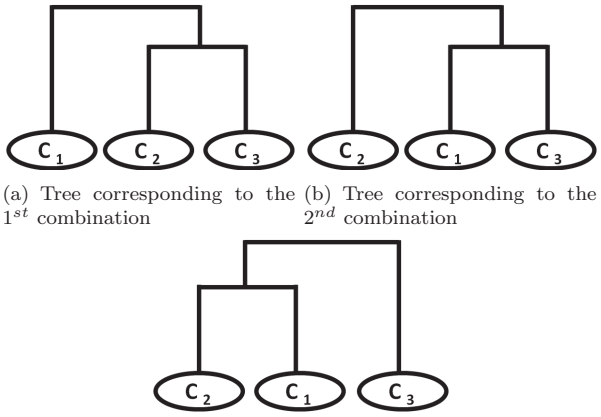
1. Selection of the hierarchical coupling among the $2^{K-1} - 1$ possible class couple;

2. Choice of the model or combining model that gives the best classification rule for the chosen couple.

In the beginning we have K classes corresponding to the samples that we want to reorganize into two classes. So, we propose either to explore all the hierarchical coupling solutions or to select the two new classes that are the most separable. These classes can be selected using the affinity coefficient (Bacelar-Nicolau 1985; Matusita 1955).

$$aff(C_k, C_{k'}) = \sum_{s=1}^{S} \sqrt{\hat{f}(\underline{x}^s \in C_k|X)} \sqrt{\hat{f}(\underline{x}^s \in C_{k'}|X)} \tag{7}$$

For each bi-class problem an intermediate position between FOIM and DTM models may be considered. The process stops when a decomposition of classes leads to a single class.
For example, when having three classes *a priori*, C_1, C_2 and C_3, the following combinations of pairs of classes can be considered: C_1 vs $C_2 \cup C_3$, C_2 vs $C_1 \cup C_3$ and C_3 vs $C_1 \cup C_2$.
Therefore, we can derive the classification rules in these three cases and select the one that yields the smallest misclassification error. Note that in this case ($K = 3$) we only have three tree configurations to consider and so it is possible to explore all the hierarchical coupling solutions (see Figure 2). E.g. in Tree (a), one observation will be first classified into C_1 vs $C_2 \cup C_3$ and if it proceeds for the 2nd level it will be finally classified into C2 or C3 , according to a minimum classification error criterion. However, when the number of classes is large (greater than three) the number of admissible tree configurations becomes larger and more difficult to handle. Then, a criterion to select trees to consider is needed. In the present work we adopt a similarity coefficient based approach and select the best tree using the affinity coefficient described above (Sousa Ferreira, 2010).



(a) Tree corresponding to the 1^{st} combination

(b) Tree corresponding to the 2^{nd} combination

(c) Tree corresponding to the 3^{rd} combination

Figura 2: Binaries trees in the MHIERM model for the K=3 case setting.

2.6 Performance Measures

To evaluate the performance of a classification rule, according to a particular model, one relies on performance measures which derive from classification results as depicted in a confusion matrix - a contingency table that associates actual and predicted classes.
In the binary case - *a priori* classes labeled 0 and 1 - the contingency table is as follows:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \text{Number of } 0\text{´s classified as } 0\text{'s} & \text{Number of } 0\text{´s missclassified as } 1\text{'s} \\ \text{Number of } 1\text{´s missclassified as } 0\text{'s} & \text{Number of } 1\text{´s classified as } 1\text{'s} \end{bmatrix}$$

where $N_0 = a + b$ and $N_1 = c + d$.
In order to find the most appropriate measure of performance several studies have been carried out (Goodman and Kruskal, 1954, 1959; Marzban, 1997; Murphy and Daan, 1985). In Discriminant Analysis the Total Success Rate - TSR measure - is commonly used. It is the average of the group specific success rates estimates weighted by the classes prior probabilities (McLachlan, 1992). And, when the group prior probabilities are estimated by the relative group sizes this measure is called Efficiency (EFF):

$$EFF = \frac{a+d}{N} \tag{8}$$

The EFF measure is simply the proportion of observations correctly classified (based on the diagonal of the confusion matrix) and misses the use of the remaining available information on the confusion matrix. Since this information can benefit the evaluation of performance of the proposed combined models, we should consider an additional evaluation measure. In fact, according to (Paik, 1998), the EFF measure may, sometimes, over-estimate the "true"success rate, particularly when classes' sizes are disproportionate or the success rates within

the classes are very different. Therefore we use an additional measure of performance in the present study - the Phi Statistic (ϕ) or index of mean square contingency, based on all the data in the confusion matrix. (Goodman and Kruskal, 1954)

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{9}$$

where:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{10}$$

n_{ij} - is the number of observations (n.o.) in the contingency table (c. tab)
$n_{i.}$ - is the n.o. in the i^{th} row in the c. tab.
$n_{.j}$ - is the n.o. in the j^{th} column in the c. tab.
n - is the total n. o.

3 Data Analysis and results

In the present work, we use of the FOIM-DTM combination to solve DDA problems. In addition, when multiple classes are considered, we suggest using HI-ERM and also recurring to the FOIM-DTM combination to obtain intermediate classification results in each tree node. Regarding the combination coefficient β, we propose to use a grid of values of $\beta \in [0, 1]$ with increments of 0.1, to weight the contribution of each model.
The Random Forest (RF) algorithm (Breiman, 2001) is used for providing comparative performance evaluation of the proposed DDA approach. The implementation used is in the R package *randomForest*, (Liaw and Wiener, 2013). For each RF we consider 500 trees, based on 500 bootstrap samples. Additionally, for each sample with replacement, we build P RF derived from subsets of features with 1 to P features. Finally, we combine all the RF into one large RF and consider the votes of $500 * P$ trees for classification.
In order to evaluate the performance of the proposed models, we consider both real and simulated data sets.

3.1 Simulated data

We conduct numerical experiments for simulated data using small and moderate sample sizes.

The data is simulated using the Bahadur model, as proposed in Goldstein and Dillon (1978) and in Celeux and Mkhadri (1992). The data sets considered derive from a previous study (Sousa Ferreira 2010; Sousa Ferreira et al. 2001). In order to simulate the predictive binary variables' values, this model defines class conditional probabilities for $C_k, (k = 1, ..., K)$ as

$$P(\underline{x}|C_k) = \prod_p \theta_{kp}^{x_p}(1 - \theta_{kp})^{(1-x_p)}[1 + \sum_{g \neq p} \rho_k(p, g)\mathrm{Z}_{kp}\mathrm{Z}_{kg}] \tag{11}$$

where X_{kp} is a Bernoulli variable with parameter $\theta_{kp} = E(X_{kp}), p = 1, ..., P$ such that

$$\mathrm{Z}_{kp} = \frac{X_{kp} - \theta_{kp}}{[\theta_{kp}(1 - \theta_{kp})]^{1/2}} \qquad and \qquad \rho_k(p, g) = E(\mathrm{Z}_{kp}\mathrm{Z}_{kg}), \tag{12}$$

considering two types of population structures, with $P = 6$ variables for the case of $K = 2$ and $K = 4$ classes. For each structure, data sets generated have 60 observations for each class (small samples) or 200 observations for each class (moderate sample).

Tabela 1: Parameters for simulated Bernoulli variables

K=2	K=4
$\theta_1 = (0.6, 0.4, 0.6, 0.5, 0.5, 0.6)$	$\theta_1 = (0.6, 0.4, 0.6, 0.5, 0.5, 0.6)$
$\theta_2 = (0.5, 0.3, 0.5, 0.4, 0.4, 0.5)$	$\theta_2 = (0.5, 0.3, 0.5, 0.4, 0.4, 0.5)$
	$\theta_3 = (0.6, 0.3, 0.6, 0.4, 0.5, 0.5)$
	$\theta_4 = (0.6, 0.4, 0.6, 0.5, 0.5, 0.6)$

The first structure, denoted IND (Independent), is generated according to FOIM, $(\rho_k(p,p) = 1$ and $\rho_k(p,g) = 0$, if $p \neq g$, $k = 1, ..., K$; $p, g = 1, ..., 6)$ for all classes.

The second one, called DIF (Different), is implemented considering the existence of different relations among the variables, for different classes:

- in the bi-class case $\rho_1(p,p) = 1$ and $\rho_1(p,g) = 0.2$, if $p \neq g$, $p, g = 1, ..., 6$; $\rho_2(p,p) = 1$ and $\rho_2(p,g) = 0.4$, if $p \neq g$, $p, g = 1, ..., 6$;
- in the multiclass case $\rho_k(p,p) = 1$ and $\rho_k(p,g) = 0.1$, if $p \neq g$, $k = 1, 2, 3$; $p, g = 1, ..., 6$; and $\rho_4(p,p) = 1$ and $\rho_4(p,g) = 0.3$, if $p \neq g$, $p, g = 1, ..., 6$.

The prior probabilities are considered equal.

3.2 Real data

We conduct numerical experiments in a very small real data set that refers to 34 dermatological patients with a diagnosis of psoriasis, with chronic evolution, (Prazeres, 1996). The relationship between three classes of patients with different degrees of Alexithymia (referring to difficulty in expressing emotions) and Rorschach test indicators (personality projective test indicators) is explored.
Nowadays, alexithymia is considered a risk factor for the process of somatic and psychological illness. Since it is difficult to identify, due to the absence of obvious mental symptoms, contributions that help to support its identification are relevant.
One of the most commonly used measures of alexithymia is the Toronto Alexithymia Scale (TAS-20). This test is a 20-items (5-point Likert) instrument. Its final score is the sum of the values assigned to the 20 items (Prazeres 1996). According to the test scores, the whole sample is divided into three small classes: Nonalexithymics Class ($C_1, n_1 = 14$), Alexithymics Class ($C_2, n_2 = 13$), Intermediate Class ($C_3, n_3 = 7$).

In this study, the goal is to explore the differences between the classes based on the fact that the alexithymia manifestations often occur after the appearances of an organic disease which, given its emotional significance and seriousness, often reflects in the Rorschach psychological test. This is a psychological test in which subjects' perceptions of inkblots are recorded and analyzed. It consists of a large number of variables measured in different scales, allowing us to know person's personality characteristics and emotional functioning.
In the present study, the characterization of each patient is based on six binary

indicators of the Rorschach test (predictor variables). In this analysis, the characterization of each patient is based on six binary variables (predictor variables) of the Rorschach test indicators (Exner, 2001):

- $CF+C > 0$ - Dichotomization of the variable $CF+C$ based on empirically established value. The value 1 was assigned when the condition is checked and 0 if not checked. $CF + C$ is the sum of chromatic color responses in which the formal element is secondary or absent. It indicates less affective modulation;
- $CF+C-FC > 0$ - Dichotomization of the variable $(CF+C)-FC$. The value 1 was assigned when the condition is checked and 0 if not checked. A positive value in $(CF + C) - FC$ indicates less affective modulation, where FC represents the number of chromatic color responses in which form features are of primary importance;;
- $V > 0$ - In pure vista responses the shading features are interpreted as depth or dimensionality. No form is involved. The value 1 was assigned when the condition is checked and 0 if not checked;
- $C' > 2$ - In pure achromatic color response the response is based on the grey, black or white features of the blot, when they are used as color. No form is involved. The value 1 was assigned when the condition is checked and 0 if not checked;
- $T = 1$ - In pure texture response the shading components of the blot are used to represent a tactual phenomenon, with no consideration to the form features. The value 1 was assigned when $T = 1$ and the value 0 was assigned when $T \neq 1$
- $SumSH-SumC > 0$ - Dichotomization of the variable $SumSH-SumC$, that compares the sum of shading responses plus the achromatic responses with the sum of chromatic color responses. The value 1 was assigned when the condition is checked and 0 if not checked.

The variables involving the chromatic color, achromatic color and shading determinants (C, C', T, V) characterize the emotional functioning.
An increase in T relates to emotional loss (e.g., marital separation). An increase in V relates to feelings of guilt or remorse. Y is related to situational stress. An increase in C' signifies the presence of disturbing negative feelings that result from an inhibition of emotional expression.

Chromatic color responses (FC,CF,C) are related to the release or discharge of emotion and to the extend to which the release is controlled or modulated. Chromatic color responses are expected to be higher than achromatic responses (FC', C'F, C'). When SumC' is greater than SumC the individual is inhibiting the release of emotions and, as a result, is burdened by irritating feelings.

(CF + C)-FC offer information concerning the modulation of emotional discharges. The FC responses relate to well controlled emotional experiences whereas the CF and the C responses relate to less restrained forms of emotional discharge. Adults without psychological problems are expected to yield

higher FC than CF+C.

Since the data were not collected in a mixture model, we could not estimate prior probabilities using relative frequencies, so the prior probabilities are taken to be equal, $\pi_k = \frac{1}{K} = \frac{1}{3}, k = 1, 2, 3$.

3.3 Classification Results

The classification results concerning simulated data sets are presented in tables 2 to 7. The FOIM-DTM combination coefficients values (beta values) appear in the tables' first column, along with the Random Forests combination results. The EFF and ϕ measures reported refer to the training and test samples (for moderate sized samples) or to the training sample and two-fold cross-validation results (for small sized samples).

- Simulated Data Results

Results referred to bi-class problems are presented in Tables 2 and 3. For the large samples (DIF and IND data included) the performance measures agree on the choice of the best model. For the DIF dataset the best results are attained with $\beta = 0.5$ to 0.7 and for the IND dataset the FOIM model yields the best results. For the small samples and the DIF dataset the DTM model attains the best result, while for the IND dataset the best combination regards $\beta = 0.9$.
When four classes are considered (large sample) the performance measures underline the advantage of the proposed combined models: for the DIF dataset the best beta values range from $\beta = 0.2$ to 0.5; for the IND dataset the best result is attained for $\beta = 0.30$ (though there is a tie for the FOIM EFF result). Generally, in the multi-class case, the models performance tends to be very poor when the HIERM approach is not considered. HIERM causes a sharp rise in the classification rates: see Tables 6 and 7 as opposed to Tables 4 and 5.
In general, in the numerical experiments conducted, the proposed approach outperforms Random Forests - it provides consistently better results when referring to small samples and, in conjugation with the HIERM approach for multi-class problems, it is clearly the winner classifier (see Table 9.).

Table 2: Classification performance, sample DIF, 2 Classes.

$\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$						
β	$n = 400$			$n = 120$		
	EFF_{Train}	EFF_{Test}	ϕ_{Test}	EFF_{Train}	EFF_{2-Fold}	ϕ_{2-Fold}
0.00	0.765	0.680	0.363	0.767	**0.792**	**0.607**
0.10	0.765	0.680	0.363	0.767	0.750	0.535
0.20	0.770	**0.685**	0.383	0.767	0.750	0.535
0.30	0.770	**0.685**	0.383	0.767	0.758	0.549
0.40	0.770	**0.685**	0.383	0.767	0.758	0.549
0.50	0.755	**0.685**	**0.390**	0.767	0.758	0.549
0.60	0.755	**0.685**	**0.390**	0.700	0.650	0.300
0.70	0.760	**0.685**	**0.390**	0.683	0.617	0.236
0.80	0.620	0.580	0.160	0.650	0.617	0.232
0.90	0.595	0.575	0.149	0.617	0.584	0.161
1.00	0.560	0.520	0.039	0.583	0.567	0.128
R. Forest	0.780	**0.685**	0.385	0.767	0.775	0.574

Table 3: Classification performance, sample IND, 2 Classes.

$\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$						
β	$n = 400$			$n = 120$		
	EFF_{Train}	EFF_{Test}	ϕ_{Test}	EFF_{Train}	EFF_{2-Fold}	ϕ_{2-Fold}
0.0	0.590	0.600	0.199	0.783	0.533	0.061
0.10	0.590	0.600	0.199	0.783	0.525	0.045
0.20	0.590	0.600	0.199	0.783	0.550	0.094
0.30	0.590	0.600	0.199	0.750	0.533	0.064
0.40	0.590	0.600	0.199	0.750	0.533	0.061
0.50	0.590	0.595	0.189	0.750	0.533	0.061
0.60	0.590	0.595	0.189	0.750	0.558	0.106
0.70	0.580	0.590	0.179	0.717	0.550	0.085
0.80	0.575	0.595	0.189	0.700	0.575	0.130
0.90	0.570	0.605	0.210	0.683	**0.583**	**0.145**
1.0	0.570	**0.610**	**0.220**	0.667	0.567	0.108
R. Forest	0.730	0.560	0.121	0.833	0.542	0.083

Table 4: Classification performance, sample DIF, 4 Classes

$\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$						
β	$n = 800$			$n = 240$		
	EFF_{Train}	EFF_{Test}	ϕ_{Test}	EFF_{Train}	EFF_{2-Fold}	ϕ_{2-Fold}
0.00	0.338	0.278	0.189	0.308	0.242	0.318
0.10	0.358	0.323	0.239	0.308	0.238	0.311
0.20	0.355	0.325	0.245	0.308	0.238	0.311
0.30	0.353	0.325	0.245	0.308	0.238	0.311
0.40	0.353	0.325	0.245	0.308	0.233	**0.367**
0.50	0.353	0.325	0.245	0.308	0.233	0.340
0.60	0.335	0.320	0.218	0.308	0.233	0.345
0.70	0.335	0.320	0.218	0.308	0.238	0.334
0.80	0.320	0.293	0.147	0.317	0.238	0.334
0.90	0.318	0.288	0.136	0.317	0.246	0.259
1.00	0.310	0.290	0.155	0.300	**0.258**	0.254
R. Forest	0.388	**0.332**	**0.264**	0.383	0.204	0.165

Table 5: Classification performance, sample IND, 4 Classes

$\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$						
β	$n = 800$			$n = 240$		
	EFF_{Train}	EFF_{Test}	ϕ_{Test}	EFF_{Train}	EFF_{2-Fold}	ϕ_{2-Fold}
0.00	0.395	0.293	0.236	0.500	0.267	0.246
0.10	0.395	0.293	0.236	0.500	0.267	0.240
0.20	0.400	0.298	0.224	0.492	0.263	0.222
0.30	0.408	0.328	0.260	0.492	0.258	0.219
0.40	0.405	0.323	0.257	0.492	**0.271**	0.225
0.50	0.405	0.315	0.124	0.500	0.263	0.211
0.60	0.393	0.318	0.210	0.492	**0.271**	0.248
0.70	0.370	0.308	0.190	0.483	0.267	0.241
0.80	0.368	0.320	0.214	0.475	0.250	0.255
0.90	0.340	0.315	0.197	0.442	0.250	0.291
1.00	0.310	0.328	0.219	0.408	0.250	**0.296**
R. Forest	0.512	**0.380**	**0.353**	0.625	0.267	0.172

Table 6: Classification performance, sample DIF, 4 Classes

$MHIERM: \quad \beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$						
β	$n = 800$			$n = 240$		
	$C_4 vs. C_1 \cup C_2 \cup C_3$ e $C_1 vs. C_2 \cup C_3$			$C_4 vs. C_1 \cup C_2 \cup C_3$ e $C_3 vs. C_1 \cup C_2$		
	EFF_{Train}	EFF_{Test}	ϕ_{Test}	EFF_{Train}	EFF_{2-Fold}	ϕ_{2-Fold}
0.00	0.710	**0.633**	**1.168**	0.558	**0.458**	0.918
0.10	0.648	0.563	1.043	0.567	0.437	**0.926**
0.20	0.648	0.563	1.043	0.567	0.437	**0.926**
0.30	0.633	0.560	1.037	0.567	0.437	**0.926**
0.40	0.633	0.560	1.037	0.500	0.412	0.861
0.50	0.628	0.555	1.025	0.508	0.412	0.861
0.60	0.625	0.560	1.037	0.517	0.413	0.869
0.70	0.615	0.550	1.016	0.517	0.392	0.847
0.80	0.615	0.583	1.053	0.517	0.396	0.856
0.90	0.605	0.560	1.048	0.500	0.387	0.833
1.00	0.615	0.570	1.073	0.492	0.400	0.857
R. Forest	0.388	0.332	0.264	0.383	0.204	0.165

Table 7: Classification performance, sample IND, 4 Classes

$MHIERM: \quad \beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$						
β	$n = 800$			$n = 240$		
	$C_2 vs. C_1 \cup C_3 \cup C_4$ e $C_4 vs. C_1 \cup C_3$			$C_3 vs. C_1 \cup C_2 \cup C_4$ e $C_1 vs. C_2 \cup C_4$		
	EFF_{Train}	EFF_{Test}	ϕ_{Test}	EFF_{Train}	EFF_{2-Fold}	ϕ_{2-Fold}
0.00	0.595	0.500	0.909	0.717	0.467	0.854
0.10	0.595	0.500	0.909	0.708	0.471	0.862
0.20	0.595	0.500	0.911	0.717	0.483	0.891
0.30	0.615	0.528	0.946	0.717	0.483	0.889
0.40	0.630	0.528	0.946	0.708	0.487	0.879
0.50	0.643	0.530	0.957	0.708	**0.500**	0.924
0.60	0.645	**0.535**	**0.966**	0.700	**0.500**	0.913
0.70	0.618	0.510	0.908	0.700	0.492	0.896
0.80	0.600	0.493	0.860	0.675	0.471	0.901
0.90	0.593	0.505	0.906	0.658	0.488	0.955
1.00	0.553	0.488	0.884	0.617	0.488	**1.000**
R. Forest	0.512	0.380	0.353	0.625	0.267	0.172

- Real Data Results

As in the simulated data results, the HIERM approach clearly improves classification results. The best result in the real data set is attained for $\beta = 0.2$ to 0.4 according to the Phi measure, illustrating the potential of the proposed combination approach to outperform the individual models-components performances. Note that the best binary tree corresponding to the most sep-

arable classes (see Figure 3) corresponds to the smallest affinity coefficient ($aff(C_1, (C_2 \cup C_3)) = 0.435$). The first decomposition chosen by the HIERM model, suggests that the union of the extremes classes forms a well-separated class from the class composed by the intermediate patients, since these subjects obtained balanced scores. Since the data set is very sparse ($2^6 = 64$ states and only 17 observations) the HIERM model provides the lowest estimated misclassification risk.

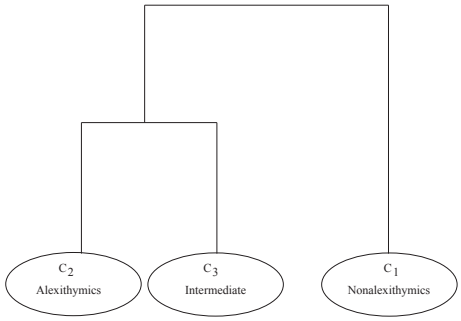


Figura 3: Binary Tree for the Alexithymia data

Table 8: Classification performance, Real Data Results

	$\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$			
β	$\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$		$MHIERM : C_1 \; vs. \; C_2 \cup C_3$	
	EFF_{2-Fold}	ϕ_{2-Fold}	EFF_{2-Fold}	ϕ_{2-Fold}
0.00	**0.471**	0.562	0.412	0.546
0.10	0.412	0.532	**0.500**	0.716
0.20	0.382	0.698	0.470	**0.812**
0.30	0.382	0.698	0.470	**0.812**
0.40	0.353	**0.707**	0.470	**0.812**
0.50	0.382	0.703	0.442	0.630
0.60	0.324	0.570	0.442	0.630
0.70	0.353	0.623	0.442	0.630
0.80	0.353	0.623	0.442	0.630
0.90	0.353	0.623	0.442	0.630
1.00	0.294	0.547	**0.500**	0.527
R. Forest	0.441	0.151	0.441	0.151

Table 9: The Winner classifiers according to EFF and ϕ measures

2 Classes	n	EFF	ϕ	4 Classes	n	EFF	ϕ
DIF	400	RF and FOIM-DTM	FOIM-DTM	DIF	800	H DTM	H DTM
	120	DTM	DTM		240	H DTM	H FOIM-DTM
IND	400	FOIM	FOIM	IND	800	H FOIM-DTM	H FOIM-DTM
	120	FOIM-DTM	FOIM-DTM		240	H FOIM-DTM	H FOIM

4 Conclusions and Perspectives

In the present work we propose using a combination of two classification models - FOIM - First-order Independence Model and DTM - Dependence Trees Model - to overcome the limitations of the individual models, namely in small and moderate sized samples settings. In addition, we propose using the HIERM -Hierarchical Coupling Model approach to address multi-class problems, recurring to a binary tree decomposition scheme.

We conduct a experimental study based on 8 simulated data sets and 1 real data set. We focus on small and moderately sized samples which tend to increase the difficulty of classification problems. Since all features are categorical we perform comparisons with a well known ensemble algorithm recognized to perform well in this setting (Kotsiantis et al., 2006) - the Random Forests ensemble approach (Breiman, 2001).

The results obtained are very encouraging - the performance of the proposed FOIM-DTM combined approach consistently exceeds the Random Forests performance when regarding small data sets. When conjugated with the HIERM approach for multi-class problems, the proposed model outperforms Random Forests in 7 out of the 8 simulated data sets.

In the real data set a very small sample is considered and, in this setting, the HIERM approach outperforms the FOIM-DTM simple combination and Random Forests as well.

We conclude that the FOIM-DTM combination is very flexible, being able to deal with different data correlations structures. In the conditional independent case - IND structure for simulated data - the FOIM naturally tends to yield the best results but the combination FOIM-DTM sometimes emerges as a better than the FOIM alternative, especially in the small sized sample cases. In the conditional non-independent case - DIF structure for simulated data - the DTM naturally tends to emerge although the combination FOIM-DTM sometimes emerges as a better than the DTM alternative, namely in the moderate sized sample cases. For the two-classes problems, the performance measures used generally agree as to the selection of the best solution. For multi-class problems with small sample sizes considered, the performance indicators may disagree. Understanding the disagreement between performance indicators should thus be the subject of future research.

The benefits of the proposed approach should be further investigated using simulated data sets with diverse correlations structures and considering imbalanced data sets too. Also, the use of more real data sets should further evidence the

advantage of the proposed combined approach.

5 References

1. Bacelar-Nicolau H. (1985) 'The Affinity Coefficient in Cluster Analysis'. *Meth. Oper. Res.* Vol. 53, pp.507-512.

2. Breiman L. (1996) 'Bagging Predictors'. *Machine Learning.* Vol. 24, pp.123-140.

3. Breiman L. (2001) 'Random forests'. *Machine learning.* Vol. 45, pp. 5-32.

4. Brito I., Celeux C. and Sousa Ferreira A. (2006) 'Combining Methods in Supervised Classification: a Comparative Study on Discrete and Continuous Problems'. *Revstat - Statistical Journal.* Vol. 4, No.3, pp.201-225.

5. Celeux G. and Mkhadri A. (1992) 'Discrete regularized discriminant analysis'. *Statistics and Computing.* Vol. 2, No.3, pp.143-151.

6. Celeux G. and Nakache J.P. (1994) *Analyse Discriminante sur Variables Qualitatives.* G. Celeux et J. P. Nakache Éditeurs, Polytechnica.

7. Chrysostomou K., Chen S. Y. and Liu X. (2008) 'Combining multiple classifiers for wrapper feature selection'. *International Journal of Data Mining, Modelling and Management*, Vol.1, No.1, pp.91-102.

8. Domingos P. and Pazzani M. (1997) 'On the optimality of the simple Bayesian classifier under zero-one loss'. *Machine learning*, Vol. 29, pp. 103-130.

9. Exner J.E. (2001) *A Rorschach Workbook for the Comprehensive System.* Fifth Edition, Asheville: Rorschach Workshops.

10. Friedman J.H. (2001) 'Greedy Function Approximation: A Gradient Boosting Machine'.*Annals of Statistics.* Vol. 29, pp.1189-1232.

11. Friedman J.H. and Popescu B.E. (2008) 'Predictive Learning Via Rule Ensembles'. *The Annals of Applied Statistics.* Vol. 2, pp.916-954.

12. Goldstein M. and Dillon W.R. (1978) *Discrete Discriminant Analysis.* New York: Wiley.

13. Goodman L. and Kruskal W. (1954) 'Measures of association for cross classifications'. *American Statistical Association Journal.* Vol. 49, pp.723-764.

14. Goodman L. and Kruskal W. (1959) 'Measures of association for cross classifications'. II Further discussion and references. *American Statistical Association Journal.* Vol. 54, pp.123-163.

15. Kotsiantis S.B. (2011) 'A random subspace method that uses different instead of similar models for regression and classification problems'. Int. J. Information and Decision Sciences, Vol. 3, pp. 173-188.

16. Kotsiantis S.B., Zaharakis I.D. and Pintelas P.E. (2006) 'Machine learning: a review of classification and combining techniques'. Artificial Intelligence Review, Vol. 26, pp. 159-190.

17. Leblanc M. and Tibshirani R. (1996) 'Combining Estimates in Regression and Classification'. *Journal of the American Statistical Association*. Vol. 91, pp.1641-1650.

18. Liaw A. and Wiener M. (2013) Package 'randomForest' - Breiman and Cutler's random forests for classification and regression. 4.6-7 ed. Repository CRAN.

19. Maalouf, M. (2011)'Logistic regression in data analysis: an overview', *Int. J. Data Analysis Techniques and Strategies*, Vol. 3, No 3, pp.281-299.

20. Marzban C. (1998) *Scalar Measures of Performance in Rare-event Situations*. Weather and Forecasting, Vol. 13, No.3, pp.753-763.

21. Murphy A.H. and Daan F. (1985) *Forecast Evaluation*. In A. H. Murphy and R. W. Katz (eds), Probability, Statistics and Decision Making in the Atmospheric Sciences. pp. 379-437. Boulder, CO:Westview Press.

22. Matusita K. (1955) 'Decision rules based on distance for problems of fit, two samples and estimation'. *Ann. Inst. Stat. Math*. Vol. 26, No.4, pp.631-640.

23. Marques A., Sousa Ferreira A. and Cardoso M. (2013) 'Variables' selection in Discrete Discriminant Analysis'. *Biometrical Letters*,Vol.50, No.1 (in press).

24. McLachlan G. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY: A Wiley-Interscience Publication.

25. Opitz D. and Maclin R. (1999) 'Popular ensemble methods: an empirical study'. *Artificial Intelligence Research*, Vol. 11, pp.169-198, Morgan Kaufmann.

26. Paik H. (1998) 'The Effect of Prior Probability on Skill in Two-Group Discriminant Analysis'. *Quality and Quantity*. Vol. 32, pp.201-211.

27. Pearl J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Los Altos: Morgan Kaufmann.

28. Peirce C.S. (1884) *The numerical measure of the sucess of predictions*. Science Vol. 4, No. 93, pp.453-454.

29. Prazeres N.L. (1996) 'Ensaio de um Estudo sobre Alexitimia com o Rorschach e a Escala de Alexitimia de Toronto (TAS-20)'. Master Thesis, Univ. Lisbon.

30. Quinlan J. R. (1993) C4.5: programs for machine learning, Morgan Kaufmann.

31. Re M. and Valentini G. (2011) 'Ensemble methods: a review', The CRC Press, LLC.

32. Sousa Ferreira A., Celeux G. and Bacelar-Nicolau H. (2000) *Discrete Discriminant Analysis: The performance of Combining Models by an Hierarchical Coupling Approach*. In Kiers, Rasson, Groenen and Shader, editors, Data Analysis, Classification and Related Methods.Springer, pp.181-186.

33. Sousa Ferreira A. (2004) *Combining models approach in Discrete Discriminant Analysis through a committee of methods. In Classification, Clustering, and Data Mining Applications*; D. Banks, L. House, F. R. McMorris, P. Arabie, W. Gaul (Eds.), Springer, pp. 151-156.

34. Sousa Ferreira A. (2010) A Comparative Study on Discrete Discriminant Analysis through a Hierarchical Coupling Approach. In *Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization*; Hermann Locarek-Junge, Claus Weihs (Eds.), Springer-Verlag, Heidelberg-Berlin, pp.137-145.

35. Wang W., Jones P. and Partridge D. (2000) 'Diversity between neural networks and decision trees for building multiple classifier systems', *Proceedings of the International Workshop on Multiple Classifier Systems*, LNCS, Vol. 1857, Springer, Calgiari, Italy, pp. 240-249.