Selection of variables in Discrete Discriminant Analysis

Anabela Marques[1], Ana Sousa Ferreira[2], Margarida G.M.S. Cardoso[3]

[1]Barreiro College of Technology, Setúbal Polytechnic, IPS, 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

Summary

In Discrete Discriminant Analysis one often has to deal with dimensionality problems. In fact, even a moderate number of explanatory variables leads to an enormous number of possible states (outcomes) when compared to the number of objects under study, as occurs particularly in the social sciences, humanities and health-related fields. As a consequence, classification or discriminant models may exhibit poor performance due to the large number of parameters to be estimated. In the present paper, we discuss variable selection techniques which aim to address the issue of dimensionality. We specifically perform classification using a combined model approach. In this setting, variable selection is particularly pertinent, enabling the handling of degrees of freedom and reducing computational cost.

Key words: combining models, Discrete Discriminant Analysis, variable selection

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, an n-dimensional sample of multivariate observations is considered $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, where $\mathbf{x}_i$ represents the ith observed state ($i \in \{1, ..., n\}$), described by M discrete variables, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iM})$ (observed state). The class of each observation - one of K exclusive classes $(C_1, C_2, ..., C_K)$ - is assumed to be known.

In general, when dealing with DDA applications, different DDA techniques may yield different classification errors for the same set of observations. In the social sciences, classification problems often exhibit a high number of variables, small or moderate size samples, and also many missing states. In this setting, the use of combined models provides a means to improve the overall performance of classification and also its stability (Ferreira, 2000; Brito, 2002; Brito *et al.*, 2006). However the related dimensionality problems have to be addressed, since there are often a large number of parameters to be estimated and a comparatively small sample available. In this work, four feature selection methods for DDA are discussed, having the aim of identifying the variables that most discriminate between the *a priori* defined classes. Two statistics are considered for this purpose: Chi-Square and Mutual Information. The simple statistics' values rankings provide two criteria. Two alternative selection criteria are based on the Chi-Square's p-values using the Bonferroni Correction and the False Discovery Rate methods (Benjamini and Hochberg, 1995). The reduction in the number of variables is expected to improve the DDA algorithm's efficiency and reduce computational cost.

The DDA approach considered is based on a linear combination of the First-order Independence Model (FOIM) and the Dependence Trees Model (DTM) (Marques *et al.*, 2008).

Classification performance is analyzed using the percentage of correctly classified observations. In addition, the runtime of the DDA algorithm (implemented in R software) is reported.

2. Variable Selection

Although feature selection is a very common theme in the literature on Discriminant Analysis with continuous predictor variables, methods proposed for Discriminant Analysis with discrete predictor variables are quite rare. However, in order to obtain good performances in DDA tasks, dimensionality issues have to be addressed. The selection of the best discriminant variables in a DDA problem is the focus of the present study. Hence we try to find M^* variables, $M^* << M$, leading to better decision rules, using the following methods:

1. Descriptive: the Chi-Square statistic (Q^2) and the Mutual Information index (I) between the M predictor variables and the target classes provide a means to rank the predictors;

2. Inferential: the p-values corresponding to the Chi-Square statistic provide alternative means to rank the same predictors - using the Bonferroni Correction (BON) and the False Discovery Rate (FDR) we obtain two additional rankings of predictors (see e.g. Benjamini and Hochberg, 1995).

When the descriptive indicators are used we report:

1. The minimal feasible solution i.e. the one having the smallest number of predictors which can be treated by the DDA model (note that when we have null mutual information, it is not possible to apply DTM)

2. The solution corresponding to the best DDA performance, i.e. that having the maximum percentage of correctly allocated cases using two-fold cross-validation.

The first descriptive indicator considered is the Chi-Square statistics (Q^2) defined as follows:

$$Q^2(X_m, X_{m'}) = \sum_{i=1}^{L} \sum_{j=1}^{K} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{1}$$

where:
$n_{i.}$ - is the number of observations in the i-th category of X_m.
$n_{.j}$ - is the number of observations in the j-th category of $X_{m'}$.
K - is the number of classes.
L - is the number of categories of the predictor.

and

$$n = \sum_{i=1}^{L} n_{i.} = \sum_{j=1}^{K} n_{.j} = \sum_{i=1}^{L} \sum_{j=1}^{K} n_{ij} \tag{2}$$

The mutual information index (I) is defined as follows:

$$I(X_m, X_{m'}) = \sum_{i=1}^{L} \sum_{j=1}^{K} n_{ij} log \frac{n_{ij}}{n_{i.}n_{.j}} \tag{3}$$

Both $Q^2(X_m, X_{m'})$ and $I(X_m, X_{m'})$ measure the strength of association between X_m and $X_{m'}$. When considering X_m as the predictor and $X_{m'}$ the target classes, these measures provide a means to rank the predictors

according to their discriminant power. In addition, when referring to the use of DTM, an $I(X_m, X_{m'})$ may also be used to measure the association between predictors.

The Chi-Square statistic Q^2 makes it possible to test the association between each predictor and the target classes, following a χ^2 distribution with $(L-1)(K-1)$ degrees of freedom under the null hypothesis (referring to null association) between the predictor and the target class. The implementation of M Chi-Square tests corresponding to the M predictors originates the p-values $p_1, ..., p_m, ..., p_M$.

The Bonferroni Correction (Benjamini and Hochberg, 1995) is a multiple-comparison correction used when several statistical tests are being performed simultaneously. The Bonferroni Correction sets the α value for the entire set of M tests equal to α by taking the α value for each test equal to α/M.

Thus, according to Bonferroni Correction (Benjamini and Hochberg, 1995) we select the predictors which yield

$$P_m \leq \frac{\alpha}{M} \tag{4}$$

The Bonferroni Correction and other traditional multiple comparison procedures are generally too conservative. In order to overcome this limitation, several alternative procedures have been proposed, such as Holm's procedure (Holm, 1979) offering a more flexible trade-off between the test's power and error. The False Discovery Rate (FDR) approach - (Benjamini and Hochberg, 1995) and (Silva, 2010) - also addresses multiple hypothesis testing to correct for multiple comparisons. In a list of statistically significant studies (e.g. studies where the null-hypothesis could be rejected), the FDR procedure is designed to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries") in a less conservative way compared with the Bonferroni Correction. This method relies on the ranked p-values (increasing values) - $p_{1:M}, ..., p_{m:M}, ..., p_{M:M}$ - and selects the predictors satisfying:

$$P_{m:M} \leq \frac{m}{M}\alpha \tag{5}$$

3. Combining Models in DDA

In Discrete Discriminant Analysis the most usual classification rule is based on the Full Multinomial Model (FMM) (Celeux and Mkhadri, 1994) where

the within-class state probability functions are multinomial. When using M binary variables, this model involves the estimation of $2M - 1$ parameters in each class, and so is cumbersome. The First-order Independence Model (FOIM) (Goldstein and Dillon, 1978; Celeux and Mkhadri, 1994) assumes the independence of variables within each class, therefore reducing the number of parameters to be estimated. Using FOIM, the conditional probability of assigning $\mathbf{x}^*$ to class C_k is estimated by:

$$\hat{f}_k\left(\mathbf{x}^* \mid X\right) = \prod_{m=1}^{M} \frac{\#\left\{\mathbf{x}_j \in C_k : x_{jm} = x_m^*\right\}}{n_k}, \ \ j = 1, \ldots, n; \ k = 1, \ldots, K \ \ (6)$$

where n_k represents the C_k class sample dimension.

FOIM, however, can be unrealistic in some situations. One of the alternative models that take into account the interactions between variables is the Dependence Trees Model (DTM), (Celeux and Nakache, 1994; Pearl, 1988).

DTM provides, for each class, an estimate of the conditional probability functions based on the idea proposed by Pearl, 1988. Pearl demonstrated that through knowledge of a graph G, where $X_1, ..., X_M$ represent its M vertices, the probability distribution f^G, associated with the graph can be calculated as the product of the conditional probabilities:

$$f^G(x_1, ..., x_M) = f(x_{r(m)}) \prod_{l(m)=1}^{M-1} f\left(x_m \mid x_{l(m)}\right) \tag{7}$$

where $x_{l(m)}$ represents a variable that is linked to the variable x_m in this graph, arbitrarily choosing one vertex as the root of the graph, $x_{r(p)}$.

The Chow and Liu (Celeux and Nakache, 1994; Pearl, 1988) algorithm is used to construct the graph for each class the length of each graph's edge (referred to the pair of variables $(x_m, x_{m'})$) represents a measure of the association between the same variables, mutual information in particular. After the calculation of the C_2^M mutual information values (see formula (3)), the graph G, with $(M - 1)$ edges, corresponding to the highest total mutual information is selected. For example, take $M = 5$ 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 and the probability distribution of the first-order dependence tree is

$$\begin{aligned} \hat{f}_k\left(\mathbf{x}^*|X\right) = \ & f^{C_k}\left(\mathbf{x}^*|X\right) = \\ = \ & \hat{f}(x_1^*|X)\hat{f}(x_2^*|x_1^*, X)\hat{f}(x_3^*|x_2^*, X)\hat{f}(x_4^*|x_2^*, X)\hat{f}(x_5^*|x_2^*, X) \end{aligned} \tag{8}$$
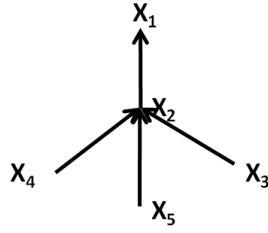
Figure 1. Example of a dependence tree for the case of M=5 variables

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

FOIM is commonly used when independent predictors are considered, while DTM takes into account the relationship between predictors. A combined model using FOIM and DTM may offer some advantages.

Combining models generally aims to obtain more robust and stable results and provide a better data fit (Bishop, 1995; Brito *et al.* 2006). Previous research by Sousa Ferreira (1999, 2000, 2010) revealed good performance for a linear combination of FMM and FOIM in the small case setting, particularly when within-class independent structures or equal correlation structures were considered. These studies also revealed that the (single) coefficient (ranging from 0 to 1) derived for the combination, often tended to heavily weight FOIM while substantially reducing the contribution of FMM, even when considering smoothed frequencies. Based on this empirical conclusion, the replacement of FMM by DTM is considered in the present work. This approach follows on from an earlier proposal, which seems to be promising (Marques *et al.*, 2008). The corresponding conditional probability function is estimated as follows:

$$\hat{P}(\mathbf{x}^* \in C_k | \beta, X) = \beta \hat{P}_{FOIM}(\mathbf{x}^* \in C_k | X) + (1 - \beta)\hat{P}_{DTM}(\mathbf{x}^* \in C_k | X) \quad (9)$$

In order to derive classification rules, the Bayes formula (the posterior probability of an observation - $\mathbf{x}^*$ - being assigned to one of the *a priori* known classes) is used:

$$P(\mathbf{x}^* \in C_k | X, \underline{\pi}) = \frac{\pi_k f_k(\mathbf{x}^*|X)}{\displaystyle\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}^*|X)} \ , \quad k = 1, \ldots, K \qquad (10)$$

where π_k represents the prior probability of class C_k and $f_k(\mathbf{x})$ represents the probability function of $\mathbf{x}$ in the same class. By applying this rule, an observation $\mathbf{x}^*$ 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 sample size of the 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, $\pi_k = \frac{1}{K}$.

4. Data Analysis and Results

This work aims to evaluate the impact of variable selection techniques on DDA results, specifically when using the FOIM and DTM combination (see(9)). The data analysis refers to three real data sets: *Alexithymics*, *Parents* and *Retail*. In these data sets, small and moderate sized samples are considered.

1. *Alexithymics data: 11 variables and 34 individuals*
   This data set consists of 34 dermatology's patients evaluated by a psychological test set (Prazeres, 1996). The whole sample is divided into three classes: Nonalexithymics (C_1), Alexithymics (C_2), Intermediate (C_3) according to the value obtained in a psychological test - the TAS-20 (Twenty Item Toronto Alexithymia Scale). For each patient the value of eleven binary variables of the Rorchach test were available. The Rorschach test is a psychological projective test in which subjects perceptions of inkblots are recorded and analyzed. The predictors are:
   - V1. $CF + C > 0$
   - V2. $CF + C - FC > 0$
   - V3. $V > 0$
   - V4. $C' > 0$
   - V5. $T = 1$
   - V6. $SumSH - SumC > 0$
   - V7. $CombC + SH > 0 - No$
   - V8. $Popular > 8 - No$

V9. $AnalCDI - No$
V10. $Mor > 2 - No$
V11. $"L > 1, 1" - No$

According to the responses given by each subject concerning each inkblot, coders are used to represent the type of answer. For example:

- C - represents chromatic color responses;

- C' - represents achromatic color responses;

- F - is the format element of responses;

- V - represents pure vista responses where shading is interpreted as dimensionality;

- T - represents texture responses;

- SH - represents shading responses;

- Mor - represents morbid contents in responses;

- L - is a ratio that compares the frequency of form responses and will all other answers;

- $Popular$ - represents very frequent responses.

The type of each subject's responses leads to an evaluation of personality characteristics, for example $CF + C > 0$ indicates less affective modulation or CDI represents a difficulty of coping. Results concerning this example are presented in Table 2.

2. *Parents data: 11 variables and 240 individuals*
   This data refers to a study which aims to analyze the relationship between marital satisfaction and coparenting in different stages of the family life cycle (Saraiva, 2010). Coparenting refers to the way in which partners relate to one another as parents and includes cooperation, triangulation and conflict. Cooperation reflects the extent to which couples support and respect each other as parents, triangulation the extent to which parents form an unhealthy alliance with the child and conflict the extent to which parents disagree about the child. The target classes are related to essential stages of family life life - families with children in preschool or primary school (C_1) and families with children in middle school or the 3rdcycle (C_2).
   This data set refers to 240 individuals and considers eleven binary variables.

V1. Gender
V2. Marital Satisfaction global score for mothers
V3. Marital Satisfaction global score for fathers
V4. Coparenting global score for mothers
V5. Coparenting global score for fathers
V6. Cooperation global score for mothers
V7. Triangulation global score for mothers
V8. Conflict global score for mothers
V9. Cooperation global score for fathers
V10.Triangulation global score for fathers
V11.Conflict global score for fathers

Results concerning this example are presented in Table 3.

3. *Retail data: 11 variables and 440 individuals*
   The *Retail Actions* data set refers to 440 clients of a wholesale business. There are two target classes: retail channel (C_1) and Horeca (Hotel, Restaurant and Caf channel) (C_1). Predictors refer to eleven managerial actions that may have an impact on the clients' purchases.

   V1. offering free samples or tastings
   V2. offering discount tickets
   V3. improving the quality of products
   V4. improving packaging
   V5. improving the store layout
   V6. preventing shortages
   V7. offering more competitive prices
   V8. offering a better selection of products and brands
   V9. offering more diversity of products and brands
   V10. presenting more in-store highlights and leaflets
   V11. extending the products assortment

   Answers refer to a binary scale: 1 - probably no; 2 - probably yes (this action will have an impact on my purchases). Results concerning this example are presented in Table 4.

The results of variable selection are presented in Table 1. According to these results the descriptive methods always provide a means to perform feature selection, while the inferential methods evidence limitations. In fact, increasing alpha values does not provide any solutions when using the Bonferroni Correction, while the FDR procedure provides solutions for *Parents* and *Retail* using $\alpha = 29\%$ and $\alpha = 38\%$, respectively.

Table 1. Selected variables for each data set and selection methods

Variable Selection	Data Sets		
	Alexithymics	*Parents*	*Retail*
Q^2 - minimal selection	V1,V3,V5,V6,V9 (M*=5)	V1,V6 (M*=2)	V4,V9 (M*=2)
Q^2 - best selection	V1,V3,V4,V5, V6,V7,V9,V11 (M*=8)	V1,V2,V4,V6, V9 (M*=5)	V2,V4,V8,V9, V11 (M*=5)
I - minimal selection	V1,V3,V6,V9 (M*=4)	V1,V6 (M*=2)	V4,V9 (M*=2)
I - best selection	V1,V2,V3,V6, V7,V9,V10,V11 (M*=8)	V1,V2,V4,V6, V9 (M*=5)	V2,V4,V8,V9, V11 (M*=5)
BON	no selection[1]	no selection[1]	no selection[1]
FDR	no selection[1]	V1,V6 (M*=2)	V2,V5 M*=2

[1]Using inferential methods (BON and FDR) it was not possible to select any set of variables allowing the classification of subjects, even on increasing the α values to 100%.

In Table 1 we represent the minimal selection, i.e. the smallest set of variables that allowed the classification of subjects using the FOIM-DTM combination. We also present the best selection, i.e. the set of variables leading to the best percentage of correctly classified observations.

Classification results based on the selected variables are presented in Tables 2, 3 and 4.

The FOIM-DTM combination coefficients values (β values) appear in the first column of the tables. The next columns concern the percentage of correctly classified observations, using classical two-fold cross-validation: two subsamples split at random are used as "Test" (sequentially) and the average of the corresponding performance measures is presented.

Runtime calculations were obtained using the same computer and the same DDA algorithm implemented in the R software.

The results of the experiments lead us to the following conclusions:

- Computational costs (time of execution) can decrease significantly (e.g. in the *ALEXITHYMICS* results with 11 predictors and 5 predictorsthe time decreases from 20 hours to 46 seconds) while classification accuracy stays approximately the same (e.g. 55.9% to 55.8% in the same *ALEXITHYMICS* experiments).

- The descriptive methods always provide a means to implement the predictor selection, while the inferential methods require specific con-

Table 2. *Alexithymics* Classification Results

β	$\beta * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$				
	All Variables	Q^2		I	
	(11 Var.)	M*=5	M*=8	M*=4	M*=8
0 (DTM)	53.0%	50.0%	47.1%	47.1%	53.0%
0.20	44.1%	50.0%	53.0%	53.0%	58.8%
0.40	41.2%	50.0%	53.0%	47.1%	61.7%
0.50	53.0%	38.2%	**64.7%**	47.1%	**67.6%**
0.60	53.0%	47.1%	58.8%	47.1%	61.7%
0.80	**55.9%**	52.9%	50.0%	47.1%	55.8%
1 (FOIM)	47.0%	55.8%	47.1%	47.1%	47.0%
Runtime	1225.2 min.	0.77 min.	21.47 min.	0.38 min.	21.11 min.

Table 3. *Parents* Classification Results

β	$\beta * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$		
	All Variables (11 Var.)	M*=2 (Q^2, I and FDR)	M*=5 (Q^2 and I)
0 (DTM)	50.8%	57.1%	50.8%
0.20	50.8%	57.1%	50.8%
0.40	52.5%	57.1%	53.4%
0.60	52.0%	57.1%	53.8%
0.80	53.3%	57.1%	55.8%
1 (FOIM)	**53.8%**	57.1%	**58.4%**
Runtime	1713.5 min.	0.24 min.	4.26 min.

Table 4. *Retail* Classification Results

β	$\beta * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$			
	All Variables	M*=2	M*=2	M*=5
	(11 Var.)	(Q^2 and I)	(FDR)	(Q^2 and I)
0 (DTM)	45.1%	60.2%	44.4%	58.6%
0.20	45.9%	60.2%	44.4%	58.6%
0.40	46.6%	60.2%	44.4%	60.2%
0.60	48.1%	60.2%	44.4%	58.6%
0.80	45.9%	60.2%	44.4%	61.7%
1 (FOIM)	**50.4%**	60.2%	**63.9%**	54.1%
Runtime	1483.2 min.	0.44 min.	0.44 min.	7.56 min.

ditions which may not be verified (particularly when the samples are small, as in the *ALEXITHYMICS* case).

- The inferential methods, when applied, yield very high significance levels (38% for *RETAIL* and 29% for *PARENTS*). However, the FDR inferential method yields the best results (best two variable selections attaining the maximum accuracy) for the *RETAIL* data set. For the *PARENTS* data set, the FDR results are similar to the Chi-Square and Mutual Information statistics (the same two predictors being selected).

5. Conclusions and Perspectives

In the present work, we compare the performance of four methods of feature selection for Discrete Discriminant Analysis (DDA) - the aim is to identify the predictors that most discriminate between the a priori defined classes. We specifically use a recent DDA methodological approach, based on a linear combination of the First Order Independence Model (FOIM) and the Dependence Trees Model (DTM), (Marques *et al.*, 2008).

According to the results obtained, we were always able to obtain an admissible selection of variables using the descriptive methods - Chi-Square and Mutual Information between predictors and the target classes providing the features' ranking. As for the inferential methods, the predictors' ranking provided by the Bonferroni correction (BON) and the False Discovery Rate (FDR) procedures, applied to Chi-Square p-values, did not always lead to a selection of acceptable predictors, even when the significance level was increased up to the maximum. However, when BON and FDR provided such a selection, the best classification rates for the FOIM and DTM combined model were attained.

Experimental results also clearly illustrate the impact of variables selection in the DDA model computation time the reduction of computational cost attained is remarkable.

The limitations regarding the inferential methods' performance may be due to the dimensions of the data sets (small and moderate)- this hypothesis should be considered in future work. Future research could also include additional methods for variable selection in DDA.

References

Benjamini Y., Hochberg Y.(1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57 289–300.

Bishop C.M.(1995): Neural Networks for Pattern Recognition. Oxford University Press.

Brito I., Celeux C., Ferreira A.S.(2006): Combining Methods in Supervised Classification: a Comparative Study on Discrete and Continuous Problems. Revstat - Statistical Journal: 4(3): 201–225.

Celeux G., Mkhadri A.(1992): Discrete regularized discriminant analysis. Statistics and Computing 2(3): 143–151.

Celeux G., Nakache J.P.(1994): Analyse Discriminante sur Variables Qualitatives. G. Celeux et J. P. Nakache diteurs, Polytechnica.

Ferreira A.S.(2000): Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas. PhD thesis (in portuguese), Universidade Nova de Lisboa.

Ferreira A.S.(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: 137–145.

Ferreira A.S.(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: 151–156.

Ferreira A.S., Cardoso M.(2010): Evaluation of Results in Discrete Discriminant Analysis. Book of Abstracts of Stochastic Modeling Techniques and Data Analysis International Conference, Skiadas, C. H. (Eds.), Chania, Creta, Grécia, Junho: 94–95.

Ferreira A.S., Celeux G., Bacelar-Nicolau H.(1999): Combining models in discrete discriminant analysis by an hierarchical coupling approach. Proceedings of the IX International Symposium of ASMDA: 159–164.

Ferreira A.S., Celeux G., 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: 181–186.

Ferreira A.S., Celeux G., Bacelar-Nicolau H.(2001): New developments on combining models in Discrete Discriminant Analysis by a Hierarchical Coupling approach. Applied Stochastic Models and Data Analysis- ASMDA; G. Govaert, J. Janssen, N. Limnios (Eds.), UTC: 430–435.

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

Holm S.(1979): A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2): 65–70.

Marques A., Ferreira A.S., Cardoso M. (2008): Uma proposta de combinação de modelos em Análise Discriminante. Estatística - Arte de Explicar o Acaso, in Oliveira I. et al. Editores, Ciência Estatística, Edições S.P.E.: 393–403.

Matusita K. (1955): Decision rules based on distance for problems of fit, two samples and estimation. In Ann. Inst. Stat. Math. 26(4): 631–640.

Pawlak Z.(1982): Rough sets. International Journal of Computer and Information Sci. 11: 341–356.

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

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.

Saraiva F.(2010): Satisfação Conjugal e relação coparental ao longo do ciclo vital da familia. Master Thesis, Univ. Lisbon.

Silva A.P.D.(2010): Classificação supervisionada para dados de elevada dimensão. Livro de resumos das XVII Jornadas de Classificação e Anlise de Dados (JOCLAD 2010), ISCTE Eds. pp. 17.