ISCTE IUL
Instituto Universitário de Lisboa

Escola de Gestão

# Análise Discriminante sobre variáveis qualitativas

Anabela de Fátima Domingues Cardoso Marques

Compilação de artigos especialmente elaborada para obtenção do grau de

Doutor em Métodos Quantitativos

Orientador(a):
Doutora Ana Sousa Ferreira, Professora Auxiliar,
Faculdade de Psicologia de Universidade de Lisboa

Co-orientador(a):
Doutora Margarida G. M. S. Cardoso, Professora Associada,
Departamento de Métodos Quantitativos para Gestão e Economia,
Escola de Gestão, ISCTE – Instituto Universitário de Lisboa

Setembro, 2014

ISCTE IUL
Instituto Universitário de Lisboa

Departamento de Métodos Quantitativos para Gestão e Economia

# Análise Discriminante sobre variáveis qualitativas

Anabela de Fátima Domingues Cardoso Marques

Compilação de artigos especialmente elaborada para obtenção do grau de

Doutor em Métodos Quantitativos

Júri:
Doutora Elizabeth Reis, Professora Catedrática, ISCTE – IUL
Doutor João Gama, Professor Associado, FEUP
Doutora Ana Isabel Lorga da Silva, Professora Associada, FEG – U. Lusófona
Doutora Teresa Oliveira, Professora Auxiliar, U. Aberta
Doutora Catarina Marques, Professora Auxiliar, ISCTE – IUL
Doutora Ana Sousa Ferreira, Professora Auxiliar, FPUL
Doutora Margarida G. M. S. Cardoso, Professora Associada, ISCTE – IUL

Setembro, 2014

*Aos meus filhos*
*Beatriz e Alexandre*

# Agradecimentos

É com um grande prazer que expresso aqui os meus agradecimentos a todas as pessoas que directa ou indirectamente me apoiaram e incentivaram ao longo deste trabalho de investigação.

Em primeiro lugar agradeço às minhas orientadoras, Professora Doutora Ana Sousa Ferreira e Professora Doutora Margarida Cardoso, por toda a disponibilidade e acompanhamento sem os quais não seria possível desenvolver este trabalho com o mesmo rigor e qualidade. A orientação efectuada acompanhada pelo constante incentivo serviu de base ao desenvolvimento de todo o conhecimento que adquiri.

Agradeço ainda à Escola Superior de Tecnologia do Barreiro que financiou parte desta pesquisa, ao Instituto Politécnico de Setúbal que validou a minha candidatura ao programa PROTEC, e em especial, à Professora Otília Dias por acreditar no meu projecto e pelo apoio incondicional, tanto a nível pessoal como profissional.

Gostaria ainda de agradecer às minhas colegas Telma Guerra, Clara Carlos e Raquel Barreira a constante e incansável ajuda com o Latex, bem como o incentivo e positivismo que ofereceram durante a realização deste trabalho.

O meu agradecimento também ao Professor Miguel Moreira, pelo apoio inicial em Matlab, ao Professor Doutor João Maroco e ao meu amigo Nuno Brites, pelo apoio no software R.

À Tânia Greeg, o meu agradecimento no apoio prestado na supervisão da tradução para inglês.

Um agradecimento aos meus alunos, foi graças ao "feedback" transmitido ao longo dos anos que muitas vezes fui buscar energia para continuar este trabalho.

Finalmente, o meu profundo agradecimento à minha família pelo apoio constante e incondicional, em particular, aos meus filhos e marido pela compreensão pelas horas que não estive presente, por acreditarem neste projecto e pela força que sempre me transmitiram, a qual me permitiu chegar ao fim desta etapa da minha vida.

A todos o meu muito OBRIGADO!

# Resumo

Este estudo insere-se no campo da Análise Discriminante Discreta (ADD) propondo uma combinação de modelos, uma vez que se tem verificado que, em geral, a sua aplicação conduz a métodos mais estáveis e robustos. O trabalho que se apresenta é particularmente focado no caso em que se dispõe de classes *a priori* mal separadas e/ou amostras de pequena ou moderada dimensão, situações em que a tarefa de ADD é mais difícil.

Procura-se com esta contribuição, ultrapassar a dificuldade de estimação de um grande número de parâmetros em ADD e encontrar classificadores que melhor se ajustem aos dados em estudo, uma vez que os erros de classificação obtidos por vários modelos não ocorrem sobre os mesmos objetos (Sousa Ferreira, 2000; Brito, 2002 e Brito et al., 2006).

Com este objetivo, propusemos uma combinação de dois modelos com especificidades diferentes, o Modelo de Independência Condicional (Goldstein and Dillon, 1978) e o Modelo Gráfico Decomponível (Celeux and Nakache, 1994; Pearl, 1988).

Tendo-nos deparado, em diversas aplicações do modelo proposto, com um número demasiado elevado de variáveis explicativas face à dimensão da amostra considerada, direcionámos o trabalho na procura de métodos de seleção de variáveis de forma a reduzir a complexidade dos dados a analisar.

Houve, ainda, necessidade de avaliar o impacto de alguns fatores no desempenho dos classificadores propostos, nomeadamente: relação entre as variáveis explicativas intra-classes; grau de separabilidade entre as classes; classes balanceadas ou não balanceadas; número de estados omissos e dimensão da amostra.

**Palavras-Chave:** Análise Discriminante Discreta; Combinação de modelos; Modelo de Independência Condicional; Modelo Gráfico Decomponível; Modelo de Emparelhamento Hierárquico.

**Classificação:** C100; C400

# Abstract

This study falls within the scope of Discrete Discriminant Analysis (DDA) and proposes a combination of models since, overall, its application has been found to lead to more stable and robust methods. The work focuses particularly on the case where there are poorly separated *a priori* classes and/or small or moderate-sized samples which tend to present more difficulties for the DDA task. This contribution sets out to overcome the difficulty of estimating a large amount of DDA parameters and to find classifiers which are better suited to the data under study, given that the classification errors obtained by diverse models do not occur on the same objects (Sousa Ferreira, 2000; Brito, 2002 and Brito et al., 2006).

To this end, we have proposed a combination of two models with different specificities, the First-order Independence Model (Goldstein and Dillon, 1978) and the Dependence Tree Model (Celeux and Nakache, 1994; Pearl, 1988).

In several applications of the proposed model, we were confronted with an excessive number of explanatory variables in relation to the sample size under study. Therefore, our work has been geared towards seeking variable selection methods, so as to reduce the complexity of the data to be analysed. It was also necessary to evaluate the impact of certain factors on the performance of the proposed combined model, namely the relationship among intra-class explanatory variables; the degree of separation between classes; balanced or unbalanced classes; number of missing states and sample size.

**Keyword:** Discrete Discriminant Analysis; Combined models for classification; First-Order Independence Model; Dependence Trees Model; Hierarchical Coupling Model.

**Classification:** C100; C400

# Sumário Executivo

Este estudo insere-se no campo da Análise Discriminante Discreta (ADD), usando uma abordagem pela combinação de modelos, com o objectivo de encontrar classificadores que melhor se ajustem aos dados em estudo, uma vez que os erros de classificação obtidos por vários modelos não ocorrem sobre os mesmos objetos (Sousa Ferreira, 2000; Brito, 2002 e Brito et al., 2006).

Em ADD, existe frequentemente um problema de dimensionalidade dado o grande número de parâmetros a estimar que os modelos mais naturais em classificação discreta exigem, particularmente porque estes métodos são frequentemente aplicados nas áreas das Ciências Sociais e Humanas ou da Medicina onde as amostras são geralmente de pequena dimensão face ao número de variáveis explicativas a analisar. A investigação desenvolvida visa, pois, contribuir para a resolução deste problema de dimensionalidade procurando conduzir assim ao incremento da precisão dos modelos.

O presente trabalho decorreu naturalmente do trabalho desenvolvido por Sousa Ferreira (2000), e propõe um modelo que se define como uma combinação linear convexa dos modelos First-order Independence Model (FOIM) (Goldstein and Dillon, 1978) e Dependence Trees Model (DTM) (Celeux and Nakache, 1994; Pearl, 1988), usando um único coeficiente $\beta$, $(0 \leq \beta \leq 1)$, supondo-se a independência entre as variáveis explicativas dentro de cada classe no primeiro modelo e tendo em conta as interações entre as variáveis explicativas no segundo. Para conhecer o campo privilegiado de aplicação da combinação FOIM-DTM o desempenho do modelo foi avaliado quer sobre dados reais, quer sobre dados simulados.

No início deste estudo vários conjuntos de dados reais foram analisados. Foi determinada, no caso de pequenas amostras, a vantagem das combinações FOIM-DTM face ao algoritmo CART.

No decorrer da investigação tornou-se pertinente considerar métodos de seleção de variáveis de forma a reduzir a complexidade dos dados a analisar. Concluiu-se que a seleção de um pequeno subconjunto de variáveis é capaz de produzir resultados com precisão idêntica ao conjunto inicial de variáveis, reduzindo drasticamente o custo computacional. Um primeiro estudo sobre dados simulados foi realizado sobre 8 conjuntos de dados (com 2 e 4 classes, pequena e moderada dimensão das amostras e graus diversos de interdependência entre as variáveis preditivas) e um conjunto de dados reais. Nele, comparou-se o desempenho de diversas combinações FOIM-DTM com o de Random Forests. Concluiu-se que o desempenho da combinação FOIM-DTM excede consistentemente o desempenho das Random Forests nas amostras de pequena dimensão. Num estudo final, bastante exaustivo, sobre 540 conjuntos de dados simulados, controlaram-se diversos fatores de complexidade associados á tarefa de classificação e o desempenho da combinação FOIM-DTM foi analisado em comparação com os modelos singulares (FOIM e DTM). Verificou-se então que a combinação FOIM-DTM revela efetivamente o seu interesse no caso de amostras de muito pequena ou pequena dimensão e classes *a priori* mal separadas. Foi ainda possível, no mesmo estudo, ordenar os fatores de complexidade de acordo com o seu nível de impacto no desempenho do modelo: separação *a priori* entre classes, rácio entre o n$^{\underline{o}}$ de graus de liberdade e dimensão da amostra, proporção de estados omissos na classe minoritária.

# Glossary

AD - Análise Discriminante
ADD - Análise Discriminante Discreta
ADL - Análise Discriminante Linear
ADQ - Análise Discriminante Quadrática
AFD - Análise Factorial Discriminante
BON - -Bonferroni Correction
CART - Classification and Regression Trees
DA - Discriminant Analysis
DDA - Discrete Discriminant Analysis
DTM - Dependence Tree Model
DFA - Discriminant Factor Analysis
FDR - False Discovery Rate
FMM - Full Multinomial Model
FOIM - First-order Independence Model
HI - Huberty Index
I - Mutual Information Index
HIERM - Hierarchical Coupling Model
LDA - Linear Discriminant Analysis
LR - Logistic Regression
MGD - Modelo Gráfico Decomponível
MHIB - Hybrid Model
MHIER - Modelo de Emparelhamento Hieráquico
MIC - Modelo de Independência Condicional
MMC - Modelo Multinomial Completo
PCA - Principal Component Analysis
QDA - Quadratic Discriminant Analysis
SVM - Support Vector Machines

# Contents

# List of Figures

# List of Tables

# CHAPTER 1

# Introduction

## 1.1 Introduction to Discriminant Analysis

The human being has always been led to group or categorize objects according to the characteristics that distinguish them, and has used methods of varying degrees of complexity to do so. In fact, classification problems frequently occur in the activity of human life. The interest in this area of study is extended to the most diverse areas of activity, and is particularly important in the field of social and human sciences and medicine where, for example, it is possible:

- to classify an e-mail message as spam or not;
- to decide whether to attribute credit to a particular bank client or not;
- to classify a patient in one of a number of defined classes *a priori* corresponding to different diagnoses, thus enabling the initiation of treatment for the disease in question, while awaiting the final results of clinical tests, etc.

Many Classification or Discriminant Analysis methods have been put forward, with a view to resolving classification problems such as those previously illustrated. When referring to classification, we can specify if it is supervised or unsupervised depending on whether the class to which each observed object belongs is known. In unsupervised classification, the class of each object is unknown and the algorithms have to find a structure in the data in order to group them in classes. Within the context of this study, classification is referred to in the sense of supervised classification. In other words, the class to which each of the observed objects belongs is already known.

The first known studies on the problems of classification go back to the 1920s, having emerged in the context of broader studies geared towards recognising human races by means of skull measurements (see Das Gupta, 1973). In 1936, Fisher introduced a definition of the discriminant function as being a combination of $P$ variables which maximises the gap between the average values of two populations, when studying taxonomy-related problems. Hence, the first formulation for a discriminant analysis problem was introduced and a methodology for its resolution was also presented. Later on, Welch (1939) and Wald (1955) presented a Bayesian approach for the classification of two populations and showed that whenever there are underlying multivariate normal distributions with an equal covariance matrix, Fisher's (Fisher, 1936; Welch, 1939; Wald, 1955) linear function leads to an optimal rule, thus minimising *a posteriori* error probability. Given the prevalence of this issue in daily life, and the huge development of data processing during the second half of the twentieth century, a number of researchers, not necessarily from the area of Statistics, have since taken an interest in the subject. Many publications have emerged over the last few decades. particularly with regard to the continuous case, in an attempt to propose new techniques for classifying objects, described by several characteristics, in two or more *a priori* defined classes, so as to obtain classification rules that are better suited to the behaviour of the data.

The classification issue has increasingly been approached and developed by researchers from a diversity of areas, namely Statistics and Machine Learning. Naturally, these approaches use very specific language such as, for example, the term inputs is used in Machine Learning to designate predictors, explanatory or independent variables, commonly used expressions in Statistical literature. Analogously, outputs is the term used for dependent or response variables, etc.

Nevertheless, the classification methods proposed by such various approaches share the same aims. They set out to define rules so that any new observation may be classified into one of the *a priori* defined classes, with greater precision than random decisions and applicable to a broader scope of problems. Therefore, any one of the designations referred to in the literature on classification is generally used by authors from different research areas. The classic statistical approach to classification problems considers discriminant analysis models based on Fisher's linear function, encompassing rather restricting assumptions on the distribution of explanatory variables. A more current approach considers more flexible models without imposing restrictions on the data under study.

The Machine Learning approach uses non-parametric methodologies that automatically learn from a series of examples. Generally, such methodologies call for prior parametrisation, which should be referred to by the analyst as being a result of his/her former experience. Methods such as decision/classification trees, where the classification of an object depends on the sequence of logical steps, are examples of this approach.

Within the context of Machine Learning, methodologies based on representations of artificial neural networks have emerged in analogy with the functioning of the human brain. These methodologies are based on a representation network of several inter-connected neurons. In this case, learning is characterised by the estimation of weights associated with the connections. The first neuronal network was proposed by McCulloch and Pitts (1943) and reproduces the characteristics of a neuron. The back-propagation algorithm is the most common training process of multilayer perceptron networks: the latter was the first learning process in 1970 created by Werbos (1990), although it only became established after its re-discovery by other researchers (Rumelhart et al., 1988).

Some of the most common classification methods are presented below.

## 1.2 Discriminant Analysis Methods

### 1.2.1 Introduction

The problems of the aforementioned supervised classification fall under the scope of Multivariate Statistics, referred to as Discriminant Analysis (DA), which includes methods for classifying new objects into one of the classes defined *a priori*, according to the knowledge of several explanatory variables. DA can, therefore, be seen as a statistical decision-making method that induces the use of probabilistic models to classify new objects (for which the class to which they belong is unknown). However, DA can also have a descriptive objective, when geared towards identifying the variables that best differentiate the *a priori* defined classes. In such cases, geometric models based on Principal Component Analysis or Correspondence Analysis may be used. Usually, the proposed models in DA give priority to the core aim (classification of new objects), despite the concern of a number of authors in finding procedures to simultaneously classify new objects and identify the most discriminative variables.

In a general manner, a Discriminant Analysis problem may be defined in the following way: In an n-dimensional sample, $X = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_n)$ where $\underline{x}_i$ represents the observation $i^{th}$ $(i \in \{1, ..., n\})$, described by P variables, $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{iP})$, knowing the class to which each observation belongs, among the K *a priori* defined classes and mutually exclusive, $(C_1, C_2, ..., C_K)$.

As already mentioned, in 1936 Fisher proposed the first discriminant function definition as a combination of the P explanatory variables which maximizes the gap between the average values of the two classes under study. This method set out to determine the line, in the case of 2 classes *a priori*, or the plane, in the case of three classes, that maximised the gap between each class. In Figure 1.1, by means of the well-known Iris dataset, the linear separability, observable between two of the three presented classes has been illustrated (iris virginica, iris versicolor, iris setosa).

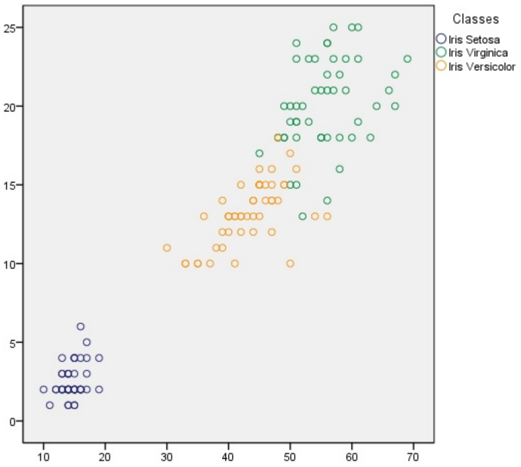


**Figure 1.1:** Iris Data

In DA we are able to distinguish several model types. For example, we may consider a geometric approach, geared towards grouping the initial P variables into homogeneous K classes, with K<P analysing the dispersion of data. Whenever we have a set of quantitative data, this method is referred to as Discriminant Factor Analysis (DFA), defined by analogy with some multivariate dimensionality reduction methods such

as the Principal Component Analysis (PCA), obtaining K-1 new variables as the linear combination of the initial P variables, thus, reducing the dimensionality of the original data, but now in completely separate classes. Whenever there are only two classes, the DFA determines a single factor that minimises variability within classes and maximises variability outside them. This factor, the linear combination of the initial P variables, coincides with Fisher's linear function. Although these types of method are easy to use, they do not enable calculation of the classification probabilities of new objects, nor do they provide necessarily optimal solutions.

With a view to identifying the classification probabilities of a new object, new methods based on Bayes' theorem have emerged, and are referred to as probabilistic methods. These methods make it possible to identify the *a posteriori* classification probability of a new object in one of the defined K classes *a priori*. The *a posteriori* probability is given as:

$$P\left(\underline{x}^* \in C_k | X, \underline{\pi}\right) = \frac{\pi_k f_k(\underline{x}^*|X)}{\sum\limits_{k=1}^{K} \pi_k f_k(\underline{x}^*|X)}\ ,\ \ k = 1, \ldots, K \tag{1.2.1}$$

where $\pi_k$ represents the *a priori* probability of the class $C_k$ and $f_k(\underline{x})$ represents the probability function of $\underline{x}$ for each k class.
Therefore, a new object $\underline{x}^*$, will be classified into the k class of maximum *a posteriori* probability, thus, minimising the classification error rate.

Naturally, in most classification problems, the *a priori* probabilities $\pi_k$ of each class and the functions $f_k(\underline{x})$ are unknown. The various probabilistic methods differ in the way of estimating the probabilities $\pi_k$ and functions $f_k(\underline{x})$ .

The estimation of *a priori* probabilities, $\pi_k$ generally varies according to the type of sampling carried out in order to extract a population sample. In other words, if the sample is randomly collected, without taking the class from which each object has come into consideration, the maximum likelihood estimators are used $\pi_k = \frac{n_k}{n}$, where $n_k$ represents the dimension of class k. On the other hand, if the sample is the result of joint independent k samples sized $n_k$, and randomly selected within each population class, the maximum likelihood estimators cannot be used and these probabilities *a priori* are regarded as equal for all classes, namely $\pi_k = \frac{1}{K}$.
There are several ways of estimating functions $f_k(\underline{x})$ depending also on the type of explanatory variables under study (continuous or discrete).

Fisher's (1936) linear function proposal made no assumption regarding either the distribution of explanatory random variables nor the covariance matrix, despite using the pooled covariance matrix S to estimate the covariance matrix $\Sigma$. According to the developments of Welch (1939) and Wald (1955), usually, whenever we have a classification problem with a set of continuous explanatory P variables, the most common classification rule is based on Normal distribution.

Therefore, when the probability density functions $f_k(\underline{x})$ follow a normal p-dimensional distribution, we may have two distinct situations: homogeneous variance/covariance matrices in the K classes

or different variance/covariance matrices for each class. In the former, Bayes rule leads to a linear classification rule referred to as Linear Discriminant Analysis (LDA) . In the latter situation, however, Bayes rule leads to a quadratic rule, referred to as Quadratic Discriminant Analysis (QDA).
LDA is easy to use but considers assumptions that are too distant from reality, while QDA, more adapted to the reality of most of the phenomena under study, is difficult to apply, since it requires the estimation of many parameters.

Despite the fact that Fisher's linear classifier presents satisfactory results when applied to problems with linearly separable classes, the same may not be said when the data do not present this characteristic. Furthermore, the normality assumption may frequently be very restricting, or even unsuitable, which has led to a search for non-parametric methods to enable the estimation of the probability functions $f_k(\underline{x})$ in each class, thus, overcoming this problem.

Another frequently used method in classification is Logistic Regression (LR) (Lemeshow and Hosmer, 2000). LR follows an approach that may be described as semi-parametric, whereby the *a posteriori* probabilities and not the probability density functions, are estimated for each class. In other words, on the basis of a set of continuous and/or discrete variables, it produces a model that enables the classification of objects in a categorical variable frequently binary {0,1}. For example, in a binary case, the classification probability is estimated in one of two classes ($Y = 0$ or $Y = 1$), in the following manner:

$$P(Y=1) = \frac{\exp(\beta_0 + \underline{\beta}'\underline{x})}{1 + \exp(\beta_0 + \underline{\beta}'\underline{x})} \tag{1.2.2}$$

where the parameters $\underline{\beta}$ are estimated on the basis of a sample, by the maximum likelihood estimators. This function is then linearised from the transformation *Logit*.

As already mentioned, in order to overcome the limitations presented by the models that impose conditions on the distribution of the variables under study, a number of non-parametric methods have recently been suggested. Some of these are Kernel Methods and Nearest-Neighbors Methods or even other types of non-parametric density function estimators, such as those based on maximum likelihood. Since the first two models are the most commonly referred to in the literature, they are described in more detail below.

The Kernel (Shawe-Taylor and Cristianini, 2004) type methods are non-parametric methods to estimate the probability density function $f_k(\underline{x})$, where each observation is considered according to the distance from a central value, the nucleus. In other words, each $\underline{x_i}$ observation is centred and a $h$ parameter is defined, which represents the nearest neighbour of $\underline{x_i}$, thus, taking all the neighbouring points into consideration for estimation.

In general, the likelihood function estimators $f_k(\underline{x})$ may be represented in the following way:

$$\hat{f_k}\left(\underline{x} \mid h\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(\underline{x} - \underline{x_i}) \tag{1.2.3}$$

where $h$ represents the smoothing parameter that defines the proximity between these estimators and the maximum likelihood estimator, whereby $K(x)$ is the non-negative Kernel function which defines the contribution of each point $\underline{x}_i$.

Hence, this method will depend on the kernel function used. It is best to choose a function that will facilitate the processing period, since these methods have been found to require a longer processing period in comparison with the previously presented methods (Silverman, 1986).

The Nearest-Neighbors (Hill, 1967) type methods use the similarity principle, and set out to classify unknown objects into the nearest class of similar objects. Therefore, the application of this method requires a definition of the $r$ number of the nearest neighbour and, consequently, the use of a distance function between pairs of observations. The estimator of the $r$ nearest neighbours $(0 \leq r \leq p-1)$, of the probability function by class is given by:

$$\hat{f}_k\left(\underline{x}^* \mid X\right) = \frac{\#\left\{\underline{x}_i \in C_k : \|\underline{x}_i - \underline{x}^*\| \leq r\right\}}{n_k}, \quad k = 1, \ldots, K \tag{1.2.4}$$

This method is very lengthy in computer terms, since all the distances between a new object and each element of the considered sample have to be calculated for construction of the model.

### 1.2.2 Other Approaches

As already mentioned, significant developments in the field of Computer Science as well as in the increased volume and complexity of the data to be analysed have been observed, which raise new challenges regarding the storage, organisation and analysis of data . This technological advance has led several researchers, from a range of different areas, to search for methods enabling them to extract patterns, tendencies and important information from the data.

So, new learning algorithms became available which are capable of predicting the class of a new object, extracting knowledge from a data set . For example, classification rules may be implemented through: Classification Trees ( e.g. using CART Breiman et al., 1984 or C4.5 Quinlan, 1993), the most frequently mentioned in the literature), Random Forests (Breiman, 2001), Neuronal Networks (e.g using Retropropagation, (Rumelhart, 1986), the most commonly used algorithm for learning based on multi-layered networks) or Support Vector Machines (SVM) (Cortes and Vapnik, 1995).

One of the important issues in the application of these methods is obtaining reliable estimates of the classification errors based on the new data. Therefore, after learning has been completed on the training data, the classification rules should be applied in new cases (test cases) to verify the accuracy of the obtained results. It is important to prevent an overfit to training data so that the algorithms may perform well on test samples. In fact, a good classifier should be capable of the same accuracy when confronted with both training data and new data (test data). Throughout the learning process, the possibility of using a validation set - data that guide learning - is an added bonus. Naturally, in all these methods, a sufficient amount of available data to make up the training , validation and test samples is required. Cross-validation is a common alternative to be used to obtain reliable error estimates when available data is restricted.

*Classification Trees*

Classification trees are non-parametric methods , since they do not require assumptions on the distribution of the variables under study. Trees make it possible to handle a large number of explanatory variables of any nature (qualitative or quantitative) and include techniques for dealing with missing data. The construction of a tree involves recursive partitioning of a data set. This partitioning process begins at the root node (set of learning data) and creates a hierarchical structure which is developed from the root to the leaves. On each level of the tree, decisions are made with regard to the following level, and the tree nodes - data sub-sets - branch out in order to reduce their diversity (in relation to the target variable (classes)). The predictive variables are partitioning instruments which enable divisions in the nodes, for example, on the basis of their possible categorisations. Different methods propose different diversity measures of the target variable, using different branching criteria. Furthermore, stopping rules for ending the partitioning or ramification process, as well as the pruning criteria of tree branches vary according to the methodologies.

*Classification and Regression Trees - CART*

The CART (Breiman et al., 1984) algorithm for the construction of classification trees performs binary recursive partitioning on the data set as a means to finding the most homogeneous data sub-sets regarding the target variable (classes). This process begins at the root node, which includes all the objects of the learning sample. The CART algorithm examines all the explanatory variables (and analyses all the possible values these variables may have) in order to perform the first ramification in two descending nodes. It then selects the variable and the corresponding categorisation that provides the highest diversity decrease of the target in the descending nodes. More specifically, the CART uses the Gini index as a diversity measure. The algorithm continues the binary ramification process until a stopping rule is imposed upon it. Finally, a class is attributed to each terminal node, namely the modal class in the sub-set of observations assigned to this node.

Interpretation of the results obtained by the CART method is simple, which is why it is a very popular method in the areas of social and human sciences and medicine. However, it should be noted that whenever the amount of available observations is low and/or not representative of the patterns in the population, the CART method classification accuracy in the training set may be difficult to replicate in a test sample.

*Random Forests*

The Random Forests algorithm was developed by Breiman (2001) and combines the idea of "bagging"[1]- constructing a forest with various trees on the basis of several samples with replacement of the initial sample - with a random selection of predictive variable sub-sets for ramification in each node. This idea was independently introduced by Ho (1995, 1998) and Amit and Geman (1997). Classification on the basis of the forest or tree committee is finally conducted by means of a voting process (weighted or otherwise) from several trees. In this case, precision estimates are not necessarily based on test samples since the so called "out of bag" estimates are available: in each sample with replacement, the excluded cases are used for testing and the corresponding classification errors are determined. On completion of the forest construction process, all the original sample cases will have been potentially included in an "out of bag"

[1] The "bagging" idea will be developed further on in point 1.3.3, page 15

sample, and finally joined in the final confusion matrix.

The trees (and random forests) deal well with discrete data and accommodate the specificities of their mensuration. As for the neural networks, the modelling of categorical variables generally involves a neuron that corresponds to each category (with the inherent connections), which easily gives rise to dimensionality problems in applications. Finally, it should be noted that the use of SVM in discrete data is naturally hampered by the transformations the method proposes for the variables input.

## 1.3 Discrete Discriminant Analysis

### 1.3.1 Brief Introduction

As already observed, some of the presented classification techniques may be applied to classification problems where the objects are described by quantitative or qualitative variables.
From the methods mentioned in point 1.2, the Kernel (Shawe-Taylor and Cristianini, 2004) type methods and those that use the r Nearest neighbours (Hill, 1967), may be easily generalisable to the discrete case. Logistic regression (Lemeshow and Hosmer, 2000), in accordance with the definition, may also be applied to explanatory qualitative variables.
With the development of technologies, other types of approaches have emerged to address classification problems, namely those described in point 1.2, Classification Trees (CART) (Breiman et al., 1984); Support Vector Machines (Cortes and Vapnik, 1995) and Random Forest (Breiman, 2001). Among these, and due to their nature, the CART algorithm and the Random Forest are models that can be applied to discrete cases. Other models, such as the SVM (Cortes and Vapnik, 1995), in accordance with its definition, implies an increase in the number of variables under study, thus, further complicating the problem of parameter estimation, one of the main problems confronted by researchers in discrete discriminant analysis. However, the specific problems of discrete classification have not been considered in the definition of these models, in other words, where all the explanatory variables are qualitative. It only happens that these models are applicable to the continuous case and to the discrete case.

Our research study falls within the scope of Discrete Discriminant Analysis, which has been far less explored by research than the continuous case. Hence, some models and specific characteristics of discrete classification problems will be presented.
Let us then consider a generic discrete classification problem defined in the following way: In the case of an n-dimensional sample, $X = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_n)$ where $\underline{x}_i$ represents the $i^{th}$ observation $(i \in \{1, ..., n\})$, described by P qualitative variables, $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{iP})$, where the class to which each observation belongs is known, from the K *a priori* defined classes, mutually exclusive, $(C_1, C_2, ..., C_K)$.

In this case, when we have P discrete variables the vector $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{iP})$ represents the $i^{th}$ observation $(i \in \{1, ..., n\})$ and corresponds to one of the observed states. In other words, in the discrete field, we resume the information of a data set by presenting the state matrix and respective observed frequencies.
To exemplify:
Let us assume a problem with two classes and two binary explanatory variables: 0,1. The values observed in this problem can only take on the following values: 00, 01, 10 and 11, which are referred to as observed

states.

Therefore, in general terms, the data set under study is resumed in a state matrix whose dimension will depend on the number of categories of each explanatory variable. In other words, for a sample composed of P binary explanatory variables, the corresponding state matrix will have $2^P$ states to analyse.

The distribution of observed frequencies is presented in Table 1.1, in a sample of 30 observations, described by 4 binary variables for two defined *a priori* classes ($n_1 = 10$ e $n_2 = 20$).

**Table 1.1:** Distribution of the observed frequencies, by state and by class

| State | $(x_1, x_2, x_3, x_4)$ | $C_1$ | | $C_2$ | |
|---|---|---|---|---|---|
| | | Observ. | Freq.Rel. | Observ. | Freq.Rel. |
| 1 | (0,0,0,0) | 0 | 0.000 | 0 | 0.000 |
| 2 | (0,0,0,1) | 4 | 0.400 | 0 | 0.000 |
| 3 | (0,0,1,0) | 1 | 0.100 | 0 | 0.000 |
| 4 | (0,0,1,1) | 1 | 0.100 | 0 | 0.000 |
| 5 | (0,1,0,0) | 0 | 0.000 | 0 | 0.000 |
| 6 | (0,1,0,1) | 2 | 0.200 | 0 | 0.000 |
| 7 | (0,1,1,0) | 0 | 0.000 | 0 | 0.000 |
| 8 | (0,1,1,1) | 2 | 0.200 | 1 | 0.050 |
| 9 | (1,0,0,0) | 0 | 0.000 | 0 | 0.000 |
| 10 | (1,0,0,1) | 0 | 0.000 | 0 | 0.000 |
| 11 | (1,0,1,0) | 0 | 0.000 | 11 | 0.550 |
| 12 | (1,0,1,1) | 0 | 0.000 | 3 | 0.150 |
| 13 | (1,1,0,0) | 0 | 0.000 | 0 | 0.000 |
| 14 | (1,1,0,1) | 0 | 0.000 | 0 | 0.000 |
| 15 | (1,1,1,0) | 0 | 0.000 | 3 | 0.150 |
| 16 | (1,1,1,1) | 0 | 0.000 | 2 | 0.100 |
| | Total | 10 | 1.000 | 20 | 1.000 |

### 1.3.2 DDA Methods

The classification methods differ according to the nature of the explanatory variables, due to the fact that the latter reflect the underlying structure to the data under study. Therefore, methods that take such characteristics into account when dealing with a set of qualitative variables are naturally sought.

The most natural model to represent a problem with qualitative explanatory variables, whether binary or not, is the Full Multinomial Model (FMM )(Goldstein and Dillon, 1978).

As with the continuous case, when handling qualitative variables, there are also reference models that play similar roles to the known methods of Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Despite being the most natural model, the FMM model requires samples of a considerable size to enable estimation of their parameters, and has a similar role in DDA to that of QDA. Since it is not possible in most real situations to satisfy this request, various models have been suggested that stem from the most well-known model in the literature, namely the First-order Independence Model (FOIM), (Goldstein and Dillon, 1978). The FOIM assumes independence among the explanatory variables within each class, which is too unrealistic in many situations. Therefore, the FOIM model represents a reference in DDA similar to that of LDA.

*Full Multinomial Model (FMM )*

The FMM, where the probability functions per class are multinomial probability functions on the set of all possible states, requires however samples of a considerable size to make the estimation of the parameters of the probability functions possible. In fact, the maximum likelihood estimator of the probability of occurrence of each state l, represented by $\underline{x}^*$ observation, in each class k is the relative frequency observed in each class:

$$\hat{f}_k\left(\underline{x}^* \mid X\right) = \frac{\#\left\{\underline{x}_i \in C_k : \underline{x}^* = \underline{x}_i\right\}}{n_k}, \quad k = 1, \ldots, K \tag{1.3.1}$$

Considering the same afore-mentioned example, Table 1.2 presents the values obtained through application of the FMM to the set of all possible states.

**Table 1.2:** Probability estimates of the occurrence of state 1 in class k through the FMM (l=1,...,16 e k=1,2)

| State | $(x_1, x_2, x_3, x_4)$ | $\hat{f}_1\left(\underline{x}^* \mid X\right)$ | $\hat{f}_2\left(\underline{x}^* \mid X\right)$ | Decision |
|---|---|---|---|---|
| | | | | (Class chosen by the model) |
| 1 | (0,0,0,0) | 0.000 | 0.000 | $C_1$ |
| 2 | (0,0,0,1) | 0.400 | 0.000 | $C_1$ |
| 3 | (0,0,1,0) | 0.100 | 0.000 | $C_1$ |
| 4 | (0,0,1,1) | 0.100 | 0.000 | $C_1$ |
| 5 | (0,1,0,0) | 0.000 | 0.000 | $C_1$ |
| 6 | (0,1,0,1) | 0.200 | 0.000 | $C_1$ |
| 7 | (0,1,1,0) | 0.000 | 0.000 | $C_1$ |
| 8 | (0,1,1,1) | 0.200 | 0.050 | $C_1$ |
| 9 | (1,0,0,0) | 0.000 | 0.000 | $C_1$ |
| 10 | (1,0,0,1) | 0.000 | 0.000 | $C_1$ |
| 11 | (1,0,1,0) | 0.000 | 0.550 | $C_2$ |
| 12 | (1,0,1,1) | 0.000 | 0.150 | $C_2$ |
| 13 | (1,1,0,0) | 0.000 | 0.000 | $C_1$ |
| 14 | (1,1,0,1) | 0.000 | 0.000 | $C_1$ |
| 15 | (1,1,1,0) | 0.000 | 0.150 | $C_2$ |
| 16 | (1,1,1,1) | 0.000 | 0.100 | $C_2$ |

Nevertheless, since large samples are necessary in order to estimate the parameters of the FMM model, has become difficult to use it in many practical cases. For example, considering the case of P binary explanatory variables, there will be $2^P$ possible states for analysis, thus, leading to an estimation of $2^P - 1$ parameters. If P=10, then 1024 parameters will have to be estimated. In order to overcome this difficulty, several variants of the FMM have been proposed (Goldstein and Dillon, 1978; Celeux and Nakache, 1994), among which the First-order Independence Model and some models based on the observed frequencies smoothed for each state, according to the application of non-parametric techniques, namely the Kernel Method and the Nearest-Neighbour Method.

*First-order Independence Model (FOIM)*

Application of the First-order Independence Model (FOIM) makes it possible to reduce the number of parameters to be estimated from $2^P - 1$ to $P$ , considering that within each class $C_k$ the explanatory variables are independent from each other. For this model, the conditional probabilities in each class $C_k$, is estimated in the following way:

$$\hat{f}_k\left(\underline{x}^* \mid X\right) = \prod_{p=1}^{P} \frac{\#\left\{\underline{x}_j \in C_k : x_{jp} = x_p^*\right\}}{n_k}, \qquad j = 1, \ldots, n\,;\; k = 1, \ldots, K \tag{1.3.2}$$

where $n_k$ represents the dimension of the class $C_k$.

In Table 1.3, the values of the conditional probabilities estimates through application of the FOIM to the previously presented data are set out.

**Table 1.3:** Probability estimates of the occurrence of state 1 in class k through FOIM (l=1,...,16 e k=1,2)

| State | $(x_1, x_2, x_3, x_4)$ | $\hat{f}_1\left(\underline{x}^* \mid X\right)$ | $\hat{f}_2\left(\underline{x}^* \mid X\right)$ | Decision |
|---|---|---|---|---|
| | | | | (Class chosen by the model) |
| 1 | (0,0,0,0) | 0.036 | 0.000 | $C_1$ |
| 2 | (0,0,0,1) | 0.324 | 0.000 | $C_1$ |
| 3 | (0,0,1,0) | 0.024 | 0.025 | $C_2$ |
| 4 | (0,0,1,1) | 0.216 | 0.011 | $C_1$ |
| 5 | (0,1,0,0) | 0.024 | 0.000 | $C_1$ |
| 6 | (0,1,0,1) | 0.216 | 0.000 | $C_1$ |
| 7 | (0,1,1,0) | 0.016 | 0.011 | $C_1$ |
| 8 | (0,1,1,1) | 0.144 | 0.005 | $C_1$ |
| 9 | (1,0,0,0) | 0.000 | 0.000 | $C_1$ |
| 10 | (1,0,0,1) | 0.000 | 0.000 | $C_1$ |
| 11 | (1,0,1,0) | 0.000 | 0.466 | $C_2$ |
| 12 | (1,0,1,1) | 0.000 | 0.200 | $C_2$ |
| 13 | (1,1,0,0) | 0.000 | 0.000 | $C_1$ |
| 14 | (1,1,0,1) | 0.000 | 0.000 | $C_1$ |
| 15 | (1,1,1,0) | 0.000 | 0.200 | $C_2$ |
| 16 | (1,1,1,1) | 0.000 | 0.086 | $C_2$ |

When the independence assumption between the variables is too unrealistic, classification methods which take into account interactions between explanatory variables can be used, namely the Dependence Trees Model (DTM) (Celeux and Nakache, 1994; Pearl, 1988) and the Bahadur Model (Celeux and Nakache, 1994; Bahadur, 1961).

*Dependence Trees Model (DTM)*

The Dependence Trees Model (DTM), takes into account conditional dependence relationships between the predictors. DTM provides for each class an estimate of the conditional probability function 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:

$$f^G(x_1, ..., x_P) = f(x_{r(p)}) \prod_{l(p)=1}^{P-1} f\left(x_p \mid x_{l(p)}\right) \tag{1.3.3}$$

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, $x_{r(p)}$.
In order 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 refers to the pair of variables $(x_p, x_{p'})$ represents a measure of the association between the same variables, particularly mutual information. Mutual information - I - is defined as follows:

$$I(X_p, X_{p'}) = \sum\sum f(x_p, x_{p'}) \log \frac{f(x_p, x_{p'})}{f(x_p)f(x_{p'})} \tag{1.3.4}$$

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

After calculation of the $C_2^P$ mutual information values, graph G, with $P-1$ edges, corresponding to the highest total mutual information is selected.

For example, the following values are obtained for mutual information and presented in Table 1.4

**Table 1.4:** Mutual Information

| $(X_p, X_{p'})$ | $I(X_p, X_{p'})$ | |
|---|---|---|
| | $C_1$ | $C_2$ |
| (1,2) | 0.000 | 0.063 |
| (1,3) | 0.000 | 0.000 |
| (1,4) | 0.000 | 0.063 |
| (2,3) | 0.014 | 0.000 |
| (2,4) | 0.055 | 0.039 |
| (3,4) | 0.100 | 0.000 |

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

$$f^{C_1}\left(\underline{x}^*|X\right) = \hat{f}(x_2^*|X)\hat{f}(x_4^*|x_3^*, X)\hat{f}(x_4^*|x_2^*, X) \tag{1.3.5}$$

$$f^{C_2}\left(\underline{x}^*|X\right) = \hat{f}(x_1^*|X)\hat{f}(x_2^*|x_1^*, X)\hat{f}(x_4^*|x_1^*, X) \tag{1.3.6}$$

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

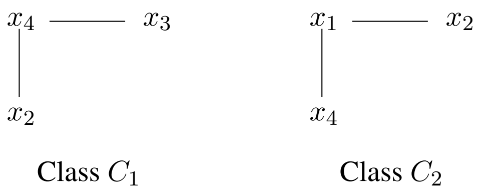


**Figure 1.2:** Example of a dependence tree for the case of P=4 variables

In accordance with the probability distribution of the first-order dependence tree (10) e (11), the value for the $8^{th}$ state $(0, 1, 1, 1)$, is calculated in the following way:

- class $C_1$:

$$f^{C_1}\left(\underline{x}^*|(0,1,1,1)\right) = \hat{f}(x_2 = 1)\hat{f}(x_4 = 1|x_2 = 1)\hat{f}(x_4 = 1|x_3 = 1) = \\ = \frac{4}{10} \times 1 \times \frac{3}{4} = 0.300 \tag{1.3.7}$$

- class $C_2$:

$$f^{C_2}\left(\underline{x}^*|(0,1,1,1)\right) = \hat{f}(x_1 = 0)\hat{f}(x_2 = 1|x_1 = 0)\hat{f}(x_4 = 1|x_1 = 0) = \\ = \frac{1}{20} \times 1 \times 1 = 0.050 \tag{1.3.8}$$

According to these results, a future object, described according to this state, should be classified in class $C_1$.

The values of the conditioned probability estimates through application of the DTM to the previously presented data are as follows:

**Table 1.5:** Probability estimates of occurrence of state 1 in class k by the DTM (l=1,...,16 e k=1,2)

| State | $(x_1, x_2, x_3, x_4)$ | $\hat{f}_1(\underline{x}^* \mid X)$ | $\hat{f}_2(\underline{x}^* \mid X)$ | Decision<br>(Class chosen by the model) |
|---|---|---|---|---|
| 1 | (1,1,1,1) | 0.300 | 0.066 | $C_1$ |
| 2 | (1,1,1,0) | 0.000 | 0.184 | $C_2$ |
| 3 | (1,1,0,1) | 0.400 | 0.066 | $C_1$ |
| 4 | (1,1,0,0) | 0.000 | 0.184 | $C_2$ |
| 5 | (1,0,1,1) | 0.375 | 0.184 | $C_1$ |
| 6 | (1,0,1,0) | 0.025 | 0.516 | $C_2$ |
| 7 | (1,0,0,1) | 0.500 | 0.184 | $C_1$ |
| 8 | (1,0,0,0) | 0.000 | 0.516 | $C_2$ |
| 9 | (0,1,1,1) | 0.300 | 0.050 | $C_1$ |
| 10 | (0,1,1,0) | 0.000 | 0.000 | $C_1$ |
| 11 | (0,1,0,1) | 0.400 | 0.050 | $C_1$ |
| 12 | (0,1,0,0) | 0.000 | 0.000 | $C_1$ |
| 13 | (0,0,1,1) | 0.375 | 0.000 | $C_1$ |
| 14 | (0,0,1,0) | 0.025 | 0.000 | $C_1$ |
| 15 | (0,0,0,1) | 0.000 | 0.000 | $C_1$ |
| 16 | (0,0,0,0) | 0.000 | 0.000 | $C_1$ |

### 1.3.3 Model Combination

Usually, when presented with a classification problem, a number of models are applied in order to select the one that proves to be the most precise. However, this procedure brings about a loss of considerable information obtained by the competing models, which is particularly important when the classification errors of some of these models are found to not occur in the same objects. The model combination approach, instead of selecting a unique model, emerged in DDA as a means to finding a classification rule that could be better adapted to the structure of the data under study. The combination of models tends to frequently improve predictive value.

Over recent years, a large number of publications from various research areas have increasingly presented proposals for combining classification methods in order to improve the models' predictive value. The results already obtained are apparently promising (for example, Wolpert, 1992; Breiman, 1996, 1998; Freund et al., 1996; Friedman et al., 1998; Sousa Ferreira et al., 2000; Friedman, 2001; Milgram et al., 2004; Sabourin, 2004; Brito, 2002; Kotsiantis et al., 2006; Cesa-Bianchi et al., 2006; Friedman and Popescu, 2008; Amershi and Conati, 2009; Janusz, 2010; Kotsiantis, 2011; Re and Valentini, 2011).

Over the years, the several model combination proposals have given rise to a broader range of terminology to designate this type of approach: Blending by Elder and Pregibon (1995), Ensemble of Classifiers by Dietterich (1997), Committee of Experts by Steinberg (1997), Perturb and Combine ($P\&C$) by Breiman (1996) and Combiners by Jain et al. (2000).

Many model combination strategies have been proposed by different researchers, whether by applying several methods to the same data set or by repeatedly using the same method on various data sets. Generally, the final prediction is decided by vote. In this chapter, a number of works in the field of model combination are referred to in chronological order.

In 1992, Wolpert proposed a classifier combination approach with stacking. This proposal consists of applying several algorithms to a data set and then a combined model is applied to attain the final prediction on the basis of the predictions from the previous step. This type of combination is illustrated in Figure 1.3. The use of this classifier combination strategy shows that by moving from one level to another, the combined model seems to learn from the errors in the previous levels and, consequently, improves its performance.

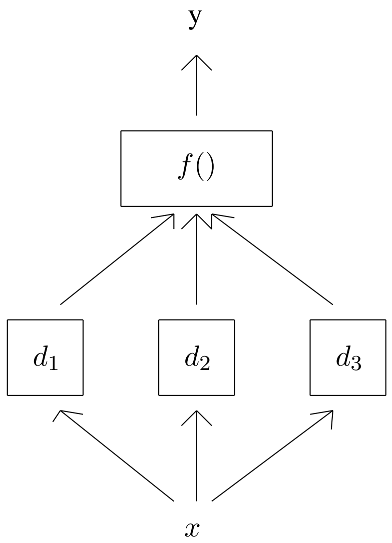


**Figure 1.3:** Illustration of the Stacking Method

where $d_i$ with $i = 1, 2, 3$ represent the values of the predictions found in the application of three different models to the x data set, and $f()$ represents the model that will combine the results obtained in the previous iteration.

This strategy presents certain aspects requiring further analysis, since there are no clear recommendations regarding the number and specificity of the models for the first level nor for the model to be applied in the last level.

Bootstrap Aggregating (bagging) was proposed by Breiman (1996) to improve the stability and precision of the algorithms used in classification, and was found to reduce variability and contribute towards preventing an overfit of the models. This method has appeared in connection with classification tree methods, however, it can be extended to any type of learning model. The bagging strategy builds a set of models based on the creation of equally-sized random samples, with replacement, stemming from the training sample (bootstrap samples). A classification algorithm is applied to each one of these samples and then a final decision is obtained by vote. This method may increase the quality of unstable algorithms such as Decision Trees and Neural Networks, but may also slightly unbalance methods considered to be stable such as the Nearest-neighbours method (Breiman, 1996).

Freund et al. (1996) proposed the boosting strategy, geared towards improving the performance of the classification model. This approach is based on the iterative combination of "weak" classifiers, giving more weight, in each iteration, to the incorrectly classified observations in the previous iteration, thus, giving rise to a "strong" classifier. A classifier is regarded as "weak" if its predictive value is lower than 0.5 (in a binary classification problem and in a balanced case). In other words, it is lower than a random classifier, while a "strong" classifier has a high predictive value, namely higher than random classification. Let us consider a combination of three "weak" classifiers, where there is a learning sample in which three learning sub-samples that randomly divide the original sample are defined. So, method $d_1$ is applied to sample $X_1$ and method $d_1$ to sample $X_2$. All the misclassified $X_1$ objects and all the well classified $X_2$ are now considered, and $d_2$ is applied to this sample. Sample $X_3$ is now taken into consideration and methods $d_1$ and $d_2$ are applied to this sample. The objects, for which the decision $d_1$ and $d_2$ differ, form the $d_3$ learning sample.

In this final stage, the decision for each object is compared by $d_1$ and $d_2$. If the methods agree, this is the final decision, otherwise the decision by $d_3$ is used (Schapire, 1990). demonstrated how this methodology reduced the error rate. The disadvantage of this proposal is that a large sized original sample is necessary so that the following sub-samples are of a non-negligible size. One of the most well-known boosting algorithms is AdaBoost (Adaptative boosting) (Freund et al., 1996) which repeatedly uses the same learning sample, overcoming the need to rely on a large sized sample. However, the classification algorithms used should prevent overfit problems.

Several variants of the AdaBoost model have been proposed. Breiman (1998) regarded it as a variation of the boosting model and referred to it as "arcing" (adaptively resampling and combining).

Within the scope of bagging and boosting methods, the researcher uses a single classification method, making the learning samples vary, believing that the application of this classification method to different sub-samples extracted from the original sample may increase the precision of the results. However, the use of a single classification method may lead to a loss of relevant information for the classification of future objects, since the application of different DA methods to the same observation set has been found to produce different classification errors in the majority of cases. This is due to the fact that each method presents different specificities and, therefore, the behaviour of the sample's distribution should be taken into account. (Sousa Ferreira, 2000; Brito, 2002; Brito et al., 2006).

Several researchers have taken an interest in this study area and have engaged in developing model combination methods for both continuous and discrete cases, although there are still very few studies on the latter.

Among the methods presented in point 1.2, the Random Forest stands out as being the closest method to a model combination approach, given that it stems from a combination of several decision trees using the bagging strategy.

Breiman (1996, 1998) demonstrated that the bagging and arcing strategies improve the performance of a CART model in 11 machine learning databases. Dietterich (2000) proved that the bagging and boosting methods systematically increased the performance of the decision tree algorithm C4.5.

Several researchers have pointed to the advantage of a neural combination network approach (Wolpert, 1992; Opitz and Shavlik, 1996).

Friedman (Friedman et al., 1998; Friedman, 2001) also invested in model combination, using the boosting strategy to withdraw subsequent sub-samples.

In 2000, Sousa Ferreira (Sousa Ferreira, 2000; Sousa Ferreira et al., 1999, 2000) addressed the problem of dimensionality in Discrete Discriminant Analysis (DDA) for small scale samples by following a model combination approach. Among the models proposed by Sousa Ferreira (2000) the Hybrid Model (MHIB) is the most salient, due to its particular way of combining two models, in the case of two *a priori* defined classes, since the objects of one class are classified according to the FOIM model, while those of the other are classified according to the DTM model. The MHIB model proved to be of particular interest in the fields of medicine and social and human sciences, where the study of classification problems with two classes is frequent. Usually, one of the classes consists of a control group and the other class an experimental group. In such situations, independence between the explanatory variables of one class is sometimes, but not so rarely observed, while in the other class relations among the explanatory variables may be found.

In the work developed by Sousa Ferreira (2000), a linear combination between the FOIM model and the FMM model was also proposed, using a single coefficient $\beta$, $(0 \leq \beta \leq 1)$, thus producing an intermediate classification rule between these two models. Later on, the proposed approach was extended to the case where more than two *a priori* defined classes are available.

The performance of this approach was assessed in terms of both real and simulated data, described by qualitative binary variables, and made it possible to ascertain the preferential field of application of the various proposed models, in accordance with the strategy used for estimation of the coefficient $\beta$. The results obtained with this approach proved to be promising in terms of increasing the predictive value of the models.

A performance analysis of the proposed approach made it possible to verify that, despite the promising results, the suggested combination tended strongly towards the FOIM model, reducing the contribution of the FMM model. This finding is what led to the model combination proposal presented in this dissertation.

Brito (Brito, 2002; Brito et al., 2006) proposed a model combination approach in a Gaussian context. Taking into account a number of assumptions regarding the parameters of the Gaussian model (covariance matrix, volume, orientation and form), Brito (Brito, 2002; Brito et al., 2006) considered fourteen models in its combination: eight elliptic, 4 diagonal and two spherical models. Performance assessment of the various proposed combinations was carried out on a number of renowned real databases, such as four at the Machine Learning Repository of California University (MLR), one at the Oxford University Repository (OR) and another at Habema and Van Den Broek (1974).

Milgram et al. (2004) proposed a combination of models with support vectors machines (SVM) and, using recognition data for digital manuscripts (the learning sample consisted of 60.000 cases, 10.000 of which were used as test-samples), showed that the necessary computation time was drastically reduced, while the precision of the SVM methods was maintained. Cesa-Bianchi et al. (2006) showed that the combination of SVM models may be an important tool in Machine Learning, in classification problems in the field of Taxonomy.

Amershi and Conati (2009) also used this approach, combining supervised and unsupervised classification models in the field of education. Moreover, Janusz (2010) studied the combination of multiple classifiers by using a genetic algorithm. Kotsiantis (2011) proposed a combination of the Random Subspace models, using the method of Naïve Bayes (Domingos and Pazzani, 1997) and C4.5 (Quinlan, 1993) and assessed the performance of the new model using 26 databases (with continuous explanatory variables). Kotsiantis (2011) concluded that the results were apparently promising.

Although many of the models put forward in the literature on model combination in classification may be applied to problems with discrete explanatory variables, the studies disseminated in the literature focus on data of a continuous nature, thus, clearly highlighting the pertinence of model combination proposals in DDA.

Hence, this study falls within the scope of model combination in DDA, particularly in the case of poorly *a priori* separated classes and/or small and moderate scale samples, in which a new classification method is proposed, following an approach based on the combination of two well-known models in Discrete Discriminant Analysis: The First-order Independence Model (FOIM) and the Dependence Trees Model (DTM).

The research undertaken in this context tried to analyse the performance of different model combination strategies in DDA, through the use of a single coefficient $\beta$, $(0 \leq \beta \leq 1)$. The aim was to reduce the dimensionality problem, and to find a better classification rule to adapt to the underlying structure of the data, which would lead to good predictive ability and stable results. This option set out to overcome the difficulty in estimating the occurrence probability of unobserved states, as found with the FMM model in the combination proposed by Sousa Ferreira (2000), and, furthermore, to extend its application to explanatory variables that are not necessarily binary.
The conditional probability function for the proposed combination is estimated as follows:

$$\hat{P}\left(\underline{x}^* \in C_k|\beta, X\right) = \beta \hat{P}_{FOIM}(\underline{x}^* \in C_k|X) + (1-\beta)\hat{P}_{DTM}(\underline{x}^* \in C_k|X) \tag{1.3.9}$$

Assessment of the performance of this new classification method was carried out on both real and simulated data, in an attempt to understand its preferential field of application. These studies are presented in the articles by Marques et al. (2008, 2010, 2014a,b) in which some data and results presented in national and international congresses are described, during the course of this research study and appendices.
When more than two *a priori* defined classes are available, the models become even more complex and hamper estimation of the unknown parameters, thus, generally leading to high error rates. The Hierarchical Coupling Model (HIERM), proposed by Sousa Ferreira (2000) is used in order to bridge this difficulty.

*Hierarchical Coupling Model (HIERM)*

The HIERM decomposes one multi-class problem into several bi-class problems using a binary tree structure. At the beginning we have K classes we want to reorganize into several couples of classes.

In other words, the HIERM model transforms a classification problem of multiple classes into multiple binary problems. Therefore, it is necessary to consider the $2^{K-1} - 1$ forms of re-grouping the initial k classes into several couples of classes. In the second stage, either the model combination is applied to every new couple of classes and the best result is chosen, or a criterion is used to select one of these decompositions and the performance of the FOIM-DTM model combination on that couple of classes is assessed. For instance, on each level of the tree a coefficient of similarity may be applied between the two new classes, and the most separate classes may be selected.

The HIERM implies two decisions on each level of the tree:

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

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

For example, 3 defined *a priori* classes may be organized into $2^{3-1}-1=3$ couples of classes in various ways, giving each one a binary tree origin. See Figure 1.4.

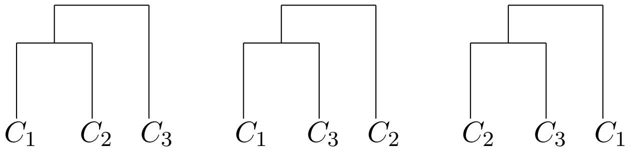


**Figure 1.4:** Binary trees for 3 *a priori* classes, in the HIERM model

In the case of 4 *a priori* classes, the number of available trees increases to 7. Two possible structure examples are presented in Figure 1.5:

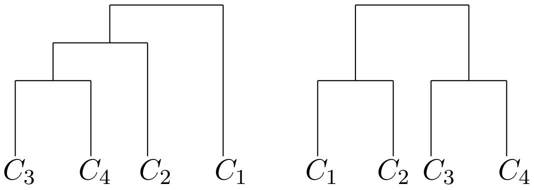


**Figure 1.5:** Example of two binary trees for 4 *a priori* classes, in the HIERM model

Of course, as the number of classes increases, the number of possible class couples to be analysed also increases, thus making this detailed process far too lengthy. Alternatively, a measure to assess the degree of separability between the several class couples may be used, by choosing the two new most separate classes.

The HIERM (Sousa Ferreira, 2000) was applied to the combination proposed in this study whenever more than two defined *a priori* classes were available. In order to calculate the degree of separability between the various class couples, the affinity coefficient $(aff)$ was used between the two discrete probability distributions defined by Matusita (1955) and generalized by Bacelar-Nicolau (1985), where:

$$aff(C_k, C_{k'}) = \sum_{s=1}^{S} \sqrt{\hat{f}(\underline{x}^s \in C_k|X)}\sqrt{\hat{f}(\underline{x}^s \in C_{k'}|X)} \qquad (1.3.10)$$

Performance assessment of the HIERM model with the FOIM-DTM combination was applied to both real and simulated data by Marques et al. (2008, 2010, 2014a,b) and appendices: Cases 1, 2, 4, 5, 6 e 10.

In the real data studies, the problem of having too many explanatory variables in relation to the number of objects under study frequently emerged when the FOIM-DTM combination was applied. This rendered the application of the classification models impracticable, or based on poor performance. It is a dimensionality problem which is quite common in DDA, and often referred to by researchers as "the curse of dimensionality" (Celeux and Nakache, 1994; Brito et al., 2006) which leads to poor performance of the various models.

A number of factors contribute to this discrete classification problem:

- DDA methods are frequently applied to the data of Social and Human Sciences and Medicine, where small scale samples are usually available;
- As already mentioned, a small number of discrete explanatory variables can easily bring about a very high number of states, thus implying that many states go unobserved;
- Samples with a large number of explanatory variables in relation to the number of observed objects are frequently found in DDA;

With a view to overcoming this problem and obtaining reliable estimates for the model parameters, variable selection methodologies have been considered in this study. The use of such methodologies is somewhat unusual in DDA, which is why this has also been examined in the research Marques et al. (2013) and appendices: Cases 6, 7.

### 1.3.4 Validation methods in supervised classification

As already established, there are many possible approaches to defining a classification rule, and there is no model that has a consistently higher performance than that of all the others. Therefore, it is fundamental that their importance is assessed in different contexts, as a means to evaluate the quality of the classification of new objects.

When discussing the evaluation of a discriminant analysis model or supervised classification, the main focus is always the predictive value of the model and not other relevant classification factors, such as the running time, the descriptive value of the model, etc. Nevertheless, in view of the importance of the running time, the performance of the classification of the models with regard to this factor was compared, as presented in the article "Selection of variables in Discrete Discriminant Analysis" (Marques et al., 2013).

Let $\delta(\underline{x}')$ be the term for the classification rule constructed by the application of a certain model to a $\underline{x}'$ learning sample. The most natural measure to assess the performance of $\delta(\underline{x}')$, involves calculating the error rate (ERR) (Celeux, 1990), or conversely, the correct classification rate:

$$ERR = \sum_{i} \sum_{j} Err(\delta(\underline{x}')_{C_i|C_j}) \ , \ \ i \neq j \tag{1.3.11}$$

where $Err(\delta(\underline{x}')_{C_i|C_j})$ represents the missclassified error rate, and the incorrectly classified object is considered to have come from $C_i$ class and is classified by $\delta(\underline{x}')$ in the $C_j$ class, with $1 \leq i, j \leq K$ and $i \neq j$.

Obviously, the true predictive value of a model is unknown, as only data samples are available for the assessment of its performance. However, several methods have been proposed in an attempt to obtain reliable predictive value estimates. These methods stem from a number of performance measures, as correctly as from alternative ways of testing the precision of a model in new observations.

Usually, the results of a discriminant analysis or a supervised classification method are assessed on the basis of counts of misclassified objects. These counts are generally represented in a table referred to as the confusion matrix. The following table represents a confusion matrix for a multiple classification problem, namely in which there are defined K *a priori* classes.

**Table 1.6:** Confusion matrix for a classification problem

| | Predicted classes | | | | | |
|---|---|---|---|---|---|---|
| Original classes | $C_1$ | $C_2$ | ... | $C_j$ | ... | $C_K$ |
| $C_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1j}$ | ... | $n_{1K}$ |
| $C_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2j}$ | ... | $n_{2K}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $C_i$ | $n_{i1}$ | $n_{i2}$ | ... | $n_{ij}$ | ... | $n_{iK}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $C_K$ | $n_{K1}$ | $n_{K2}$ | ... | $n_{Kj}$ | ... | $n_{KK}$ |

The $n_{ij}$ frequency, presented in each cell of table 6 represents the number of objects belonging to class i which were classified by the model in classj. Therefore, the total number of correctly classified objects is given as:

$$\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}\ ,\ \ i=j \tag{1.3.12}$$

and the total number of missclassified objects is given as

$$\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}\ ,\ \ i\neq j \tag{1.3.13}$$

therefore, the rate of correctly-classified objects is given as

$$\frac{\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}}{n}\ ,\ \ i=j \tag{1.3.14}$$

and the total amount of the sample's objects, where:

$$n=\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}=\sum_{i=1}^{L} n_{i.}=\sum_{j=1}^{K} n_{.j} \tag{1.3.15}$$

Some authors have designated this rate as the Accuracy of the model, while other measures may also be defined, such as the Overall Error Rate, given as:

$$\textit{Overall Error Rate}\ =\frac{\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}}{n}\ ,\ \ i\neq j \tag{1.3.16}$$

or even taking into account the predictive value observed in each class, where the rate of correctly-classified objects in class k is given as:

$$\textit{Correct Classification Rate in } C_k = \frac{n_{kk}}{\displaystyle\sum_{j=1}^{K} n_{kj}} \tag{1.3.17}$$

Considering the particular case of two defined *a priori* classes, with:

- correctly-classified rate of Class 1 objects (sometimes referred to as sensitivity):

$$\frac{n_{11}}{n_{11} + n_{12}} \tag{1.3.18}$$

- correctly-classified rate of Class 2 objects (sometimes referred to as specificity):

$$\frac{n_{22}}{n_{21} + n_{22}} \tag{1.3.19}$$

In the same way, the error rate per class may be defined in a classification problem with K classes.
The precision calculation per class has the advantage of demonstrating whether the predictive value is the same in both classes or if it is very high just because the observed precision is very high in only one of the classes.
By assessing the classification rules of the sample in which the same rules were learned, a good fit to the data will naturally be obtained. Therefore, different ways of assessing the predictive value of a model are generally used on new data beyond those that support the learning or training of a model. The most common forms of classification rules are presented below:

- *Resubstitution error* - The term resubstitution error refers to the error rate (or correct classification rate) based on the same sample used in the learning process. Due to the fact that the same sample is used in the validation process, this estimate is over optimistic and may misrepresent the real predictive value of the model.

- *Sample test validation* - The term sample-test validation is employed when a sample is used for the learning process and another sample is used for the estimation of the resulting model's precision. In this case, the initial sample is split into two independent sub-samples, not necessarily equally sized (when small scale samples are available, it is recommended that the sample be split into two equal parts), and the error rate calculated on the basis of the sample that was not used in the construction of the classification rules .

- *V-fold cross validation*- The term V-fold cross validation is used when the initial sample is split into equally-sized $V$ parts and the $V - 1$ parts are used for the construction of classification rules and then evaluated in the remaining sample. Hence, this process gives rise to $V$ iterations. The error rate estimate is obtained by taking an average of the error rates obtained in these $V$ iterations. Typical choices for V are V=2, V=5 or V=10.

- *Leave-one-out Cross- Validation* - The term Leave-one-out Cross-Validation is used when there is a particular case of the cross-validation method. In this case, $V$ is equal to the number of objects of the dataset. The subsequent sub-sets for validation are formed by a single object and the learning set is made up of all the other objects.

- *Bootstrap Validation* - The term Bootstrap Validation is used when the validation is based on a systematic re-sample with replacement. In other words, a random sample with replacement of equal size is created from a data set of n objects. This sample is used as a learning set while the remaining objects form the validation set. This operation is repeated a sufficiently large number of times.

Given that small scale samples were considered in this study, two-fold cross-validation was used. In samples of a reasonable size, sample-testing was used (made up of half of the original observations).

All the previously described measures for evaluating the predictive value of a classification rule have proven to be somewhat inefficient when the defined *a priori* classes are not balanced and, furthermore, even when the classes are balanced, the predictive value of one class is very different to the others. In this case, the evaluation of a rule, using the correct classification or error rates may lead to incorrect conclusions. For example, if one of the classes has around 90% of the observed objects and all these objects are correctly-classified, the idea is conveyed of a highly precise rule (90%), even though all the class two objects may be misclassified. In such cases, it is of particular interest to take into account all the frequencies registered in the confusion matrix and not only those that constitute the secondary or principal diagonal.

Several authors have invested in the search for suitable methods to compare two or more classification models - for example (Sousa Ferreira and Cardoso, 2013; Bostanci and Bostanci, 2013; Gomez and Montero, 2011; Santos and Embrechts, 2009; Demšar, 2006; Dietterich, 1998; Carletta, 1996). Therefore, in addition to the well-known Error Rate (or Correct Classification Rate) , such as in the case of binary problems, the area in percentage under the Receiver Operating Characteristic (ROC) curve, the determination of sensitivity and specificity and the statistics of McNemar's test, used to analyse the frequencies of related samples, have been proposed. In more general terms, in problems with multiple classes, Cohen's Kappa statistic may be referred to, which is an agreement measure between original and predicted classes ( Carletta, 1996; Foody, 2004 or the Wilcoxon test that compares the distribution of the observed results in two related populations. Recently, other performance measures used in external clustering validation have been considered in the assessment of classification methodologies performance (Sousa Ferreira and Cardoso, 2013; Santos and Embrechts, 2009). Nevertheless, it is still difficult to draw clear conclusions on the measures to be used and on what specific contribution they offer to the validation of classification results.

Within the scope of this study, the decision was made to use not only the traditional correct classification rate (or error rate) as performance measures of the proposed model combination, but also the coefficient $\phi$ (Marques et al., 2014a) and the Huberty Index ($HI$) (Marques et al., 2014a):

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{1.3.20}$$

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{1.3.21}$$

$n_{ij}$ - number of observations in the $i^{th}$ row and in the $j^{th}$ column in the contingency table.
$n_{i.}$ - number of observations in the $i^{th}$ row in the contingency table.
$n_{.j}$ - number of observations in the $j^{th}$ column in the contingency table.
$n$ - is the total number of observations.

$$n = \sum_{i=1}^{L} n_{i.} = \sum_{j=1}^{K} n_{.j} = \sum_{i=1}^{L} \sum_{j=1}^{K} n_{ij} \tag{1.3.22}$$

$$HI = \frac{P_c - P_d}{1 - P_d} \tag{1.3.23}$$

where $P_c$ represents the percentage of correctly classified cases and $P_d$ represents the percentage of majority class cases.

### 1.3.5 Selection of variables in supervised classification

When a study is developed on a certain theme, a numerous set of explanatory variables is generally used with a view to characterising the objects under study in a suitable manner. However, some of these variables are frequently redundant bringing no additional information to the model.
In many Discriminant Analysis (DA) applications, only a small sub-set of explanatory variables contain information regarding the class (McLachlan, 1992; Dash and Liu, 1997; Silva, 1999; Cook and Yin, 2001; Rebouças, 2011; Murphy et al., 2010) . Therefore, to consider variables that do not contribute to knowledge on class affectation increases the complexity of the analysis and may, consequently, reduce the performance of the DA model. It is, therefore, natural to include variable selection methods in DA procedures.
In DA, variable selection may be accomplished with two different aims:

- to identify the variables that best differentiate the defined classesa priori;
- to identify the variables that lead to a classification rule with better predictive value than the rule based on the set of all the explanatory variables.

Generally, when we discuss the selection of variables in DA, it is on the basis of the latter aim.
The objective of variable selection is three-fold (Guyon and Elisseeff, 2003): improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.

Variable selection in DA may be accomplished as a process preceding Discriminant Analysis (Filter Method) or, on the contrary, this process can be accomplished by using step by step methods that enable a selection of the variables that offer the best contribution to the precision of a specific classification algorithm (Wraper Method) (Rebouças, 2011; Murphy et al., 2010).

When the selection of variables is accomplished as a process that precedes Discriminant Analysis, univariate methods are used, based on an analysis of the relations between the explanatory variables under study and the dependent variable (classes). The mutual information or coefficient correlations are examples of the type of method that can be used in this approach.

In this case, the P explanatory variables are individually analysed, and those displaying a strong relation with the dependent variable (classes) are selected. Naturally, these methods do not take into account the relations among the various explanatory variables and a sub-set of p variables (p<P) which does not lead to a classification rule with good predictive value may be chosen. In fact, when considered with the others, the variables that are not selected may be important for the construction of a classification rule by a specific algorithm.

In other words, the methods that select the variables as part of the discriminant analysis process, generally provide higher classification precision, but have a high running cost. On the other hand, hybrid methods may be an alternative, since they initially reduce the space of the variables (using the Filter Method), and then use the Wrapper method so that step by step, in combination with the emerging results of a given algorithm, a set of predictive variables providing a good classification may be found.

The selection of variables in DA has been the target of many studies (McLachlan, 1992; Dash and Liu, 1997; Rebouças, 2011; Murphy et al., 2010), in which a number of variable selection techniques are described, namely the step by step methods, inherent to the actual classification models. In classification problems with continuous variables, step by step methods are commonly selected, and which are developed by applying criteria, for example, in the Mahalanobis distance (see McKay and Campbell, 1982; Sousa Ferreira, 1987), unlike the classification problems with discrete variables.

The previously described classification methods: classification trees, Random Forests and SVM and the combination strategy Bagging do not require a previous selection, since this analysis is already a part of their procedure in which the most relevant variables are chosen during the course of the process.

In this study, the problem of dimensionality emerged mainly due to the fact that our work focused on small to moderate samples, a field of DDA in which the dimensionality issue is more pressing. Therefore, the application of a DDA model, such as the proposed combination FOIM-DTM, to real data, small or moderate scale, described by a large number of explanatory variables inevitably leads to low predictive values. This fact geared our research towards the field of variable selection methods, so as to find a minimum number of explanatory variables that suitably characterise the phenomenon under study, and which enable the construction of classification rules within an acceptable period of time. Very little has been studied on variable selection in DDA literature with a view to finding those that will lead to a better classification rule.

Therefore, with a view to finding a sub-set $p^*$, $p^* << P$ of the initial explanatory P variables that produce similar or better results than the initial set, four types of filter selection criteria were adopted in this study:

1. Descriptive: The Chi-Square statistic ($Q^2$) and the Mutual Information Index ($I$) between the P predictor variables and the target classes both provide one criterion for ranking the predictors;

2. Inferential: The p-values corresponding to the Chi-Square test provide alternative criteria for ranking the same predictors. Using the Bonferroni Correction (BON) and the False Discovery Rate (FDR), we obtained two additional predictor rankings (e.g. see (Benjamini and Hochberg, 1995)).

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

$$Q^2(X_m, X_{m'}) = \sum_{i=1}^{L}\sum_{j=1}^{K} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{1.3.24}$$

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

and

$$n = \sum_{i=1}^{L} n_{i.} = \sum_{j=1}^{K} n_{.j} = \sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij} \tag{1.3.25}$$

The Mutual information index ($I$) is defined as follows:

$$I(X_m, X_{m'}) = \sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij} log \frac{n_{ij}}{n_{i.}n_{.j}} \tag{1.3.26}$$

Both $Q^2(Xm, Xm')$ and $I(Xm, Xm')$ measure the strength of the association between $Xm$ and $Xm'$. When considering $Xm$ a predictor and $Xm'$. When considering $Xm$ a predictor and $Xm'$ the target classes, these measures provide means for ranking the predictors according to their discriminant capacity.

The Chi-Square statistic $Q^2$ makes it possible to test the association between each predictor and the target classes, following a $\chi^2$ 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. Then, the Bonferroni Correction, which sets the $\alpha$ value for the entire set of $M$ tests by taking the significance level for each test equal to $\alpha/M$.

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

$$p_m \leq \frac{\alpha}{M} \tag{1.3.27}$$

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 - e.g. Holm's procedure (Holm, 1979) offering a more flexible tradeoff between the test's power and error.

The False Discovery Rate (FDR) approach (Benjamini and Hochberg, 1995;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 to 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 obeying:

$$p_{m:M} \leq \frac{m}{M}\alpha \tag{1.3.28}$$

## 1.4 Thesis Outline

As a result of this research, five articles were written. They will be presented in the next chapter. The first manuscript (Marques et al., 2008), the second (Marques et al., 2010) and the third (Marques et al., 2013) have already been published; the fourth manuscripts (Marques et al., 2014a) is accepted but still finalizing review and the fifth (Marques et al., 2014b) have been submitted and are still under review.

Marques et al. (2008) presented the new combination model in DDA using the HIERM for multiple classes. The proposed model was illustrated on one set of real data and evaluated by resubstitution.

Marques et al. (2010) presented the new DDA approach based on a linear combination of FOIM and DTM. This was applied to classify one set of real data and another with simulated data. This paper has focused on the performance of the new approach in comparison with CART and HIERM, as the data had more than two defined classes *a priori*.

Since this study focuses on small or moderate scale samples, dimensionality problems emerged on a number of occasions, dealing with too many explanatory variables vs. the number of objects under study. This situation motivated the study for the article Marques et al. (2013). In this paper, diverse variable selection techniques were considered to address the issue of dimensionality and their impact on the performance of the new combined classification approach. We concluded that variable selection was particularly pertinent in this setting, enabling the handling of degrees of freedom and significantly reducing the running cost.

In order to understand the preferential field of application of the proposed model, an additional study was conducted - Marques et al. (2014a). It resorted to simulated data sets with two and four classes and controlled the level of correlation between variables within each class. The combined model performance - and also the performance of a Hierarchical Coupling Model when addressing multi-class classification problems - were compared with Random Forests' performance. The obtained results highlighted the pertinence of the proposed model, especially when small samples were considered. A real dataset was used to complete the comparative analysis.

Marques et al. (2014b) evaluated the performance of the proposed FOIM-DTM combination by using simulated datasets. The experimental scenarios considered different factors - class separation, balance, the number of missing states and sample size - and 30 runs were conducted in each scenario. The obtained results enabled further understanding of the performance of the proposed combination, when compared with the single FOIM or DTM methods. In addition, the results were used to build a linear regression model considering performance measures as dependent variables. The obtained model showed a good fit to the data and made it possible to anticipate the performance of the proposed algorithm in a real dataset (based on the corresponding measures of separation, balance, missing states and sample dimension).

In the last part of this thesis, final conclusions and future research issues are presented.

# CHAPTER 2

## Uma proposta de combinação de modelos em Análise Discriminante Discreta

This paper has the following reference:

Marques, A.; Sousa Ferreira, A. and Margarida G. M. S. Cardoso (2008) 'Uma proposta de combinação de modelos em Análise Discriminante Discreta'. *Estatística - Arte de Explicar o Acaso*, in Oliveira, I. et al. Editores, *Ciência Estatística*, Edições S.P.E, 393-403.

**Erratum**

In pag. 9, where is "MHIERM2" should be "MHIER2".

# *Uma proposta de combinação de modelos em Análise Discriminante Discreta*

**Anabela Marques**
*Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal e CEAUL, Projecto de Analise de Dados Multivariados e Modelação* - anabela.marques@estbarreiro.ips.pt

**Ana Sousa Ferreira**
*Universidade de Lisboa, FPCE, Laboratório de Estatística e Análise de Dados e CEAUL, Projecto de Analise de Dados Multivariados e Modelação* - asferreira@fpce.ul.pt

**Margarida Cardoso**
*Departamento de Métodos Quantitativos, UNIDE - Escola de Gestão do ISCTE* - margarida.cardoso@iscte.pt

**Resumo:** A Análise Discriminante (AD) discreta ou sobre variáveis qualitativas é utilizada principalmente em estudos nas áreas das ciências sociais e humanas e da saúde, onde se dispõe frequentemente de grupos *a priori* mal separados e/ou de amostras de pequena dimensão. Nestas condições, o objectivo decisional de afectação dos indivíduos/objectos aos grupos está claramente dificultado. O presente trabalho insere-se no campo da AD sobre variáveis qualitativas, não necessariamente binárias, utilizando uma abordagem de combinação de modelos, para o caso em que se dispõe de mais de dois grupos *a priori*. O objectivo da proposta aqui apresentada é ultrapassar as dificuldades de afectação/classificação presentes em muitas situações práticas. O Modelo de Emparelhamento Hierárquico (HIERM) foi proposto por Sousa Ferreira (Sousa Ferreira (2000), Sousa Ferreira et al. (2000)) no contexto de uma análise discriminante sobre variáveis qualitativas no caso de mais de dois grupos *a priori* e alia a maior simplicidade da estimação do problema de dois grupos, à estabilidade de uma combinação de modelos.
O modelo HIERM:

- decompõe um problema de mais de dois grupos *a priori* em diversos problemas de dois grupos, utilizando uma estrutura de árvore binária;
- em cada nível da árvore, a regra de decisão, baseia-se numa combinação de modelos para o caso de dois grupos *a priori*.

Na abordagem apresentada neste trabalho, a regra de decisão, em cada nível da árvore, baseia-se na combinação dos seguintes modelos: o Modelo de Independência Condicional de ordem um (MIC) que supõe a hipótese de independência entre as variáveis dentro dos grupos e o Modelo Gráfico Decomponível (MGD) (Pearl (1988)) que tem em conta a existência de interacções entre as variáveis.
Para o estudo do desempenho desta nova abordagem estão a ser implementados novos programas estatísticos no software MATLAB que posteriormente serão aplicados a dados reais, comummente utilizados na literatura de AD, sendo ainda comparados com metodologias já conhecidas.

**Palavras–chave:** Análise discriminante, combinação de modelos, modelo de independência condicional, modelo gráfico decomponível, modelo de emparelhamento hierárquico.

**Abstract:** Discrete Discriminant Analysis (DDA) is often used in social sciences, aiming to allocate individuals/objects to *a priori* constituted groups, based on some qualitative attributes referring to the same individuals/objects. The present work regards the use of DDA referred to qualitative attributes which are not necessarily binary. The proposed approach considers discrimination between more than two groups and aims to overcome some difficulties often occurring in practical applications, namely the occurrence of hill-separated groups and/or small size samples. In such situations, the allocation decisions (of individuals/objects) to groups is clearly a difficult task. The Hierarchical Model (HIERM) was proposed by Sousa Ferreira ((2000), (Sousa Ferreira *et al.* (2000) to deal with DDA when more that two *a priori* groups are considered. It is both easy to estimate and capitalizes on the stability yielded by combined models. HIERM:

- Relies on a binary tree structure decomposing a multiple group discriminant problem into several binary problems;
- In each level of the tree the decision rule (for a binary problem) results from a combining model

In the present work the decision rule, in each tree level, is based on First-Order Independence Model (FOIM)(Celeux and Nakache (1994)) which assumes that the P variables are independent in each group, and Dependence Trees Model (DTM) (Celeux and Nakache (1994), Pearl (1988)) which takes into account the interactions between the predictors, bivariate relationships in particular. The proposed algorithm is implemented in MATLAB and is illustrated with a real application. In future research it should be applied to real data commonly used in the AD literature and compared with well known ADD methodologies.

**Keywords:** Discrete Discriminant Analysis, DDA models' combination, First Order Independence Model, Dependence Trees Model, Hierarchical model.

## 1 Introdução

A Análise Discriminante é uma técnica de Análise de Dados Multivariados que pode ser utilizada quando estamos perante um conjunto de n objectos, descritos por P variáveis, provenientes de K grupos definidos *a priori*, mutuamente exclusivos, com o objectivo de:

1. conhecer quais as variáveis, de entre as P variáveis que os descrevem, que melhor diferenciam os K grupos;
2. predizer a pertença de um novo objecto anónimo a um e um só dos K grupos definidos *a priori*, mediante a aplicação de uma regra de decisão que minimize os erros de afectação.

Os K grupos definidos *a priori* podem estar definidos desde o início, aquando da recolha dos dados ou podem ter sido obtidos após a aplicação de outras técnicas estatísticas como por exemplo, por aplicação de técnicas estatísticas como a Análise de Agrupamentos.
Em Análise Discriminante existem métodos que privilegiam mais o objectivo (1). No entanto, a maior parte dos métodos inserem-se mais no campo decisional, ou seja, na situação (2). Mais raramente, alguns métodos conseguem responder aos dois objectivos (1) e (2). De entre os métodos que correspondem à situação (1) podemos dizer que estes foram inspirados na Análise em Componentes Principais ou na Análise de Correspondências, enquanto em (2) nos referimos a métodos probabilísticos (Celeux (1991)).

A regra de decisão mais usual baseia-se na fórmula de Bayes, surgindo naturalmente a forma de estimar a probabilidade *a posteriori* de afectação de um objecto $\underline{x}$, num dos K grupos definidos *a priori*, (Celeux (1991)):

$$P\left(G_k \mid \underline{x}\right) = \frac{p_k P_k(\underline{x})}{\sum\limits_{k=1}^{K} p_k P_k(\underline{x})} \tag{1}$$

onde $p_k$ representam as probabilidades *a priori* do grupo $k$, e $P_k(\underline{x})$ as funções de probabilidade para cada grupo K. Mediante a aplicação desta regra afectamos um novo objecto $\underline{x}$, ao grupo $G_K$ que apresenta a probabilidade *a posteriori* máxima, minimizando assim o erro de afectação.

Este trabalho, insere-se no campo decisional, onde o conjunto de n objectos em estudo é descrito por P variáveis qualitativas, não necessariamente binárias, e provenientes de K grupos definidos *a priori* com K>2.

## 2 Análise Discriminante Discreta (ADD)

Perante um conjunto de dados discretos a regra de decisão mais usual baseia-se no Modelo Multinomial Completo (MMC) (Celeux e Nakache (1994)). No entanto, a sua utilização envolve a estimação de um número muito elevado de parâmetros. Por exemplo, para o caso em que dispomos de P variáveis binárias, teremos de estimar $2^P - 1$ parâmetros. A estimação deste número de parâmetros só se torna viável recorrendo a amostras de elevada dimensão, o que na prática, em algumas áreas, como por exemplo nas ciências da saúde e em psicologia, se tem mostrado difícil de obter. Para ultrapassar este problema da dimensionalidade foram propostas diversas variantes deste modelo (MMC) entre as quais se destaca o Modelo de Independência Condicional (MIC) (Celeux e Nakache (1994)) que assume que dentro de cada um dos grupos definidos *a priori*, $G_k$, as P variáveis são independentes. Neste modelo, a função de probabilidade condicional do grupo $G_k$ é estimada da seguinte forma:

$$\hat{P}\left(G_k \mid \underline{x}\right) = \prod_{p=1}^{P} \frac{\#\left\{\underline{y} \in G_k : y_p = x_p\right\}}{n_k}, \quad \text{para} \;\; p = 1, \ldots, P \tag{2}$$

onde $n_k$ representa a dimensão da amostra para o grupo $G_k$. Este modelo, além de reduzir o número de parâmetros a estimar permite ainda encontrar uma solução para a selecção de variáveis aquando da construção da regra de decisão, escolhendo-as independentemente umas das outras, através do recurso ao Teste do Qui-quadrado de independência entre a variável em causa e a variável que define os grupos *a priori*. Dado que a hipótese de independência entre as P variáveis nem sempre é válida, têm surgido diversos modelos na literatura, em alternativa ao modelo MMC, entre os quais o Modelo Gráfico Decomponível (MGD) (Celeux e Nakache (1994), Pearl (1988)). Este modelo considera as interações entre as variáveis de uma forma fácil de interpretar, dado que se baseia no algoritmo proposto por Chow e Liu (1968) que utiliza uma estrutura em árvore, designada por árvore de dependência, baseando-se na informação mutual. Assim, a função de probabilidade condicional para o grupo $G_k$ é estimada pelo produto das estimativas das probabilidades condicionais, correspondentes aos ramos da árvore seleccionados, que representam as interações mais importantes entre as variáveis. Por exemplo, no caso de termos cinco variáveis explicativas e determinada a informação mutual, se a conclusão fosse que as interações mais importantes eram $(x_2, x_1), (x_3, x_2), (x_4, x_2) e (x_5, x_2)$ teríamos então, como estimativa para a probabilidade condicionada do conjunto das cinco variáveis o seguinte produto:

$$\widehat{P}(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_2) \tag{3}$$

## 3 Combinação de Modelos em ADD

Na década de 90 foram vários os investigadores que começaram por combinar modelos com o intuito de encontrar métodos que se adaptassem melhor ao comportamento dos dados em estudo e que pudessem de alguma forma minimizar o número de parâmetros a estimar, nas mais diversas áreas da Estatística (Wolpert (1992), Raftery (1996), Chipman *et al.*(1998), Smith e Wolpert (1999)). Em AD, pensou-se que em vez de propor novos modelos para reduzir o problema da dimensionalidade, a abordagem pela combinação de modelos conduziria a modelos mais eficientes e estáveis, tanto mais que frequentemente se observa que os erros de afectação obtidos por vários modelos não ocorrem sobre os mesmos objectos (Sousa Ferreira (2000), (Sousa Ferreira *et al.* (2000), Brito (2002), Brito *et al.* (2006)). Sousa Ferreira ((2000), Sousa Ferreira *et al.* (2000)) propôs, um modelo de combinação natural para o caso de dois grupos *a priori* que consiste em utilizar um único coeficiente produzindo um modelo intermédio entre o modelo multinomial completo (MMC) e o modelo de independência condicional (MIC), tendo desenvolvido várias estratégias para estimar esse coeficiente.

Nesse estudo, observou-se que em diversas estratégias utilizadas, o coeficiente da combinação tende a ponderar fortemente o modelo MIC e a reduzir muito a contribuição do modelo MMC, mesmo quando as suas frequências são alisadas. Com base nas conclusões desta investigação, Marques *et al.* (2008) propôs substituir, na referida combinação, o modelo MMC pelo modelo MGD, o qual tem em conta a interacção entre as variáveis em estudo, surgindo assim um novo modelo intermédio entre o modelo MIC e o modelo MGD, sendo a função de probabilidade condicionada estimada da seguinte forma:

$$\hat{P}\left(\underline{x} \mid \beta\right) = \beta \hat{P}_{MIC}(\underline{x}) + (1-\beta)\hat{P}_{MGD}(\underline{x}) \tag{4}$$

Da aplicação deste novo modelo, Marques *et al.* (2008), obteve uma melhoria na taxa de afectação dos n objectos aos K grupos definidos *a priori*. No entanto, estando este estudo ainda numa fase inicial, ainda não é possivel apontar conclusões definitivas.

## 4 Uma variante do Modelo de Emparelhamento Hierárquico

Quando o contexto da análise discriminante sobre variáveis qualitativas se situa no caso de mais de dois grupos a *priori*, pode ser vantajoso, como Friedman (1996) já tinha observado, transformar este caso em diversos problemas de discriminação entre dois grupos, conduzindo a modelos mais fáceis de estimar e de interpretar. Friedman tinha proposto a decomposição dos K grupos em todas as combinações possíveis de pares de grupos. Para cada par estimava a regra de decisão e, no final, cada objecto seria afectado a partir da decisão maioritaria em todos os pares de grupos.

Sousa Ferreira propôs ((2000), Sousa Ferreira *et al.* (2000), Brito *et al.* (2006)) uma abordagem diferente. O modelo intermédio proposto foi generalizado para o caso de mais de dois grupos *a priori* através do Modelo de Emparelhamento Hierárquico (MHIER), o qual decompõe o problema inicial de $K > 2$ grupos, *a priori* , em diversos problemas de dois grupos, utilizando uma estrutura de árvore binária, sendo cada objecto afectado ao grupo associado ao último ramo da árvore onde foi classificado.
O modelo MHIER exige assim, duas decisões em cada nível da árvore:

- A selecção de um par hierárquico entre as $2^{K-1} - 1$ possibilidades de emparelhamento (par de grupos compostos resultante da partição de K grupos);
- Em cada nível da árvore, a selecção do modelo que conduz à melhor regra de decisão.

No primeiro nível da árvore, dispomos de $K$ grupos e pretendemos reorganizálos num par de grupos. Assim, Sousa Ferreira propôs escolher o novo par de grupos compostos que fossem mais separados. Após a selecção do 1º nível da árvore, obtém-se a regra de decisão para este par de grupos. Repete-se, então, a escolha para o segundo nível da árvore entre todos os grupos compostos (formados por dois ou mais grupos iniciais). O processo termina quando a decomposição dos grupos conduz a grupos singulares.

Exemplificando, consideremos um caso de três grupos *a priori*, $G_1$, $G_2$ e $G_3$, teremos então que considerar as seguintes combinações de pares de grupos: $G_1$ versus $G_2 \cup G_3$, $G_2$ versus $G_1 \cup G_3$, $G_3$ versus $G_1 \cup G_2$. Determinada uma medida de proximidade entre estes três pares de grupos será seleccionado para o 1º nível da árvore binária o par com valor da medida de proximidade mínimo. Seguidamente obtém-se a regra de decisão para este par de grupos e repete-se o processo para o 2º nível da árvore (formado exclusivamente por grupos singulares). Neste caso, em que o número de grupos é pequeno, podemos também optar por construir as três árvores binárias correspondentes às combinações de pares de grupos acima referidas, escolhendo no final a que conduz à melhor taxa de afectação.
A regra de decisão obtida através do modelo de emparelhamento hierárquico pode ser representado numa árvore binária como está exemplificado na Figura 1.
O modelo MHIER proposto por Sousa Ferreira (2000) revelou não só ser uma

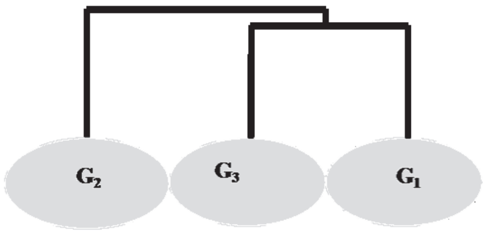


Figura 1: Árvore binária correspondente à 2ª combinação no caso de três grupos.

metodologia promissora para melhorar os erros de má classificação como ainda a sua estrutura em árvore binária conduzir, geralmente, a interpretações particularmente interessantes para os dados em estudo.
Devido a este facto, Marques *et al.*(2008)) utilizou também a abordagem do modelo de emparelhamento hierárquico com a combinação de modelos entre o modelo MIC e o modelo MGD, no caso de K>2 grupos *a priori*. As primeiras avaliações do desempenho deste novo modelo hierárquico, obtidos com dados reais, reforçam os resultados já obtidos por Sousa Ferreira ((2000), Sousa Ferreira *et al.* (2000), Brito *et al.* (2006)). Atendendo a que na maior parte dos casos dispomos de um conjunto de n objectos, descritos por P variáveis indepen-

dentes, não necessariamente binárias, provenientes de K>2 grupos definidos *a priori*, o presente trabalho teve como objectivo reformular a proposta até agora desenvolvida por Marques *et al.* (2008) para variáveis binárias, para este caso mais geral.
Assim, apresentamos de seguida, o pseudo-código de uma nova variante do algoritmo MHIER, implementado com o objectivo de aliar a simplicidade de uma discriminação entre dois grupos com uma nova proposta de combinação de modelos que poderá melhorar a sua capacidade preditiva.

Na concepção deste algoritmo é de extrema importância a natureza dos dados em análise ser qualitativa. Assim sendo, torna-se relevante observar que a forma mais usual de comparar a distribuição de várias populações descritas por variáveis qualitativas consiste em organizar a informação em termos do espaço do resultados associado a essas variáveis (isto é, todos os vectores que são possíveis ser observados com o número de variáveis descritoras em análise) o que nos permite a comparação das populações através das frequências relativas com que cada um desses estados foi observado. Exemplificando, no caso mais simples de duas variáveis binárias e supondo que cada uma delas pode assumir o valor 0 ou 1, teremos, então, quatro estados possíveis :

00, 01, 10, 11.

Para comparar as distribuições de várias populações bastaria comparar a frequência relativa observada de cada estado. Como é sabido em geral esta informação é desconhecida e os diversos modelos de ADD diferenciam-se por proporem técnicas de estimação distintas para estas distribuições desconhecidas.

```
pseudo-código do algoritmo MHIER2
Considere-se a matriz de dados $X = [x_{np}]$ (*n =1...N, número de objectos;
p=1...P, número de variáveis*).
Para cada grupo k (*k=1...K, número de grupos*) calcular:
    o vector de estados $E^k = [m_l]$ (*l=1...lmax, sendo lmax $\leq N$,
    a matriz de frequências de estados $F^k = [f_l^k]$;
(* Construir a árvore binária Ab, base de implementação do modelo hierárquico,
MHIER2 *)
$O_1 = \{\underline{x} : X \in G_1 \cup G_2 \cup ... \cup G_k\}$(*nó raiz de Ab*)
Para j=1...K-1 (* j refere-se a nível da árvore Ab* )
    Para s=1...S(j) (*s número de partições possíveis dos grupos em $O_j$ que
    é $2^{K-1} - 1$ quando referido ao nó raiz*)
        constituir uma partição $\Pi^s$ dos grupos alvo em $O_j$ em 2
        grupos compostos;
        Determinar MP(s) (*medida de proximidade*)entre os 2
        grupos compostos ou efectuar busca exaustiva;
    Identificar os melhores grupos compostos, $\Pi^j$ correspondente ao $Min_s\{MP(s)\}$,
    s=j...S(j) ou usar todos os encontrados na busca exaustiva;
```

```
        Construir ramificação de Ab considerando $\Pi^j$ ou construir as várias
        árvores Ab resultantes da busca exaustiva.
(*Proceder à classificação/discriminação de dados a partir de Ab*)
Para j=1...K-1 (*em cada nível j de Ab vão considerar-se os dois grupos compostos - $G_{s1,j}$ e $G_{s2,j}$ - nele constituídos*) calcular:
        as novas matrizes de frequências de estados dos grupos compostos
        $F^{s1} = [f_l^{s1}]$ e $F^{s2} = [f_l^{s2}]$ ;
        utilizar os critérios MIC, MGD e $(\beta)MIC+(1-\beta)MGD$(*com $\beta = 0.25$;
        $\beta = 0.50$; $\beta = 0.75$ *) para afectação dos dados e construir matrizes de
        classificação correspondentes $A_{(MIC)}$, $A_{(MGD)}$ e $A_{(\beta)MIC+(1-\beta)MGD}$;
Determinar a proporção de casos bem classificados $P_c$ correspondentes a $A_{(MIC)}$, $A_{(MGD)}$ e $A_{(\beta)MIC+(1-\beta)MGD}$;
Afectar cada objecto segundo o último ramo em que ele é afectado/classificado.
Seleccionar o melhor modelo de classificação, correspondendo a $Max\{P_c(A_{(MIC)})\}$, $\{P_c(A_{(MGD)})\}$ e $P_c(A_{(\beta)MIC+(1-\beta)MGD})$, considerando $\beta = 0.25$; $\beta = 0.50$; $\beta = 0.75\}$
```

Para o estudo do desempenho desta nova abordagem forem implementados novos programas estatísticos no software MATLAB, tendo sido aplicados posteriormente a dados reais. Sobre estas análises apresentam-se, a seguir, as conclusões.

## 5 Aplicação

Os dados que se seguem fizeram parte de um trabalho sobre a averiguação das características do stress parental em grupos de pais de crianças que se diferenciam da seguinte forma:

- Grupo $G_1$ - Pais de crianças com doença crónica - Fibrose Quística, n=14;
- Grupo $G_2$ - Pais de crianças com doença crónica - Doença Celíaca, n=13;
- Grupo $G_3$ - Pais de crianças sem problemas físicos ou psicológicos - Grupo de Controle, n=15.

Foram analisadas 42 crianças com as características referidas e registadas as suas respostas no questionário de Índice de Stress Parental (ISP). Este instrumento foi concebido para avaliar a intensidade do stress que ocorre no sistema pais-crianças, quando este é submetido a diversas pressões, e é composto por 108 itens, organizados em dois domínios (Domínio da Criança e Domínio dos Pais), cada um subdividido em várias subescalas. Cada item é medido numa escala de tipo Likert de 5 pontos.

Neste trabalho consideramos apenas, os oito itens correspondentes à subescala Autonomia do Domínio da Criança, cuja nota global revelou, em estudos anteriores, um forte poder discriminativo. Para ilustrar a aplicação do modelo de emparelhamento hierárquico à combinação intermédia entre o modelo MIC e o modelo MGD (MHIERM2) proposto por Marques *et al.* (2008), e atendendo a que apenas temos 3 grupos em estudo, realizámos a análise de forma exaustiva, ou seja, aplicamos o modelo MHIERM2 para os grupos: $G_1 vs.(G_2 + G_3)$, $G_2 vs.(G_1 + G_3)$ e $G_3 vs.(G_1 + G_2)$, tendo obtido os resultados que se apresentam nas Tabelas 1 e 2.

Tabela 1: Taxas de bem afectados para os diversos modelos combinados.

| | $\beta$ MIC $+(1-\beta)$ MGD | | | | |
|---|---|---|---|---|---|
| $\beta$ | 1 | 0.25 | 0.50 | 0.75 | 0 |
| Perc. bem af. | 76.2% | 92.9% | 95.2% | 90.5% | 92.9% |

Tabela 2: Taxas de bem afectados para os diversos modelos MHIERM2.

| | $\beta$ MIC $+(1-\beta)$ MGD | | | | |
|---|---|---|---|---|---|
| $\beta$ | 1 | 0.25 | 0.50 | 0.75 | 0 |
| $G_1 vs.G_2 + G_3$ | 76.2% | 95.2% | 97.6% | 83.3% | 95.2% |
| $G_2 vs.G_1 + G_3$ | 78.5% | 90.5% | 95.2% | 88.0% | 90.5% |
| $G_3 vs.G_1 + G_2$ | 76.2% | 95.2% | 97.6% | 83.3% | 95.2% |

Da análise destes resultados podemos concluir que a combinação intermédia entre o modelo MIC e o modelo MGD fornece óptimos resultados. No entanto, se utilizarmos o Modelo de Emparelhamento Hierárquico para esta combinação intermédia (MHIERM2) ainda conseguimos melhorar os resultados de afectação.

## 6 Conclusões e perspectivas

Como foi referido nas secções anteriores, este trabalho tem vindo a desenvolver o seu campo de aplicação tendo começado por aplicar a abordagem de combinação de modelos proposta, o modelo intermédio entre o modelo MIC e o modelo MGD, no caso de dois grupos *a priori* e variáveis binárias. Posteriormente, generalizou-se ao caso de mais de dois grupos *a priori* através da utilização da ideia do modelo de emparelhamento hierárquico, numa primeira fase com variáveis binárias e neste trabalho com variáveis qualitativas não necessariamente binárias. Há que referir como limitação deste trabalho a análise do desempenho da metodologia proposta já que apenas foi avaliada em dados reais e sem recurso a qualquer amostra holdout. Deste modo, a continuação do trabalho de investigação em torno desta temática irá considerar a implementação de técnicas de validação das taxas de erro, adequadas para as pequenas dimensões de amostras que têm vindo a ser consideradas, como validação cruzada, *leave-one-out* ou *v-fold*. A análise do desempenho dos novos modelos propostos continuará a ser explorada recorrendo a análises comparativas com outros modelos quer sobre dados reais quer sobre dados simulados.

## Referências

[1] Brito, I. (2002). *Combinaison de modèles en analyse discriminante dans un contexte gaussien.* Thèse de Doctorat, Université Joseph Fourier, Grenoble.

[2] Brito, I., Celeux, G. e Sousa Ferreira, A. (2006). *Combining methods in Supervised Classification: a comparative study on discrete and continous problems. REVSTAT - Statistical Journal*, Vol. 4(3), 201-225.

[3] Chipman, H., George, E. e McCullach, R. (1998). *Bayesian CART model search (with discussion). Journal of the American Statistical Association*, 93, 935-960.

[4] Chow, C. K., Liu, C. N. (1968).*Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory IT-14*, 3, 462-467.

[5] Celeux, G. (1991). *Analyse Discriminante sur Variables Continues.* (G. Celeux Éditeur), Collection Didactique: INRIA.

[6] Celeux, G., Nakache, J. P. (1994). *Analyse Discriminante sur Variables Qualitatives.* (G. Celeux e J. P. Nakache Éditeurs), Polytechnica.

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

[8] Marques, A., Sousa Ferreira, A. e Cardoso, M. (2008). *Combinação de Modelos em Análise Discriminante* Em *Livro de Resumos das XV Jornadas de Classificação e Análise de Dados (JOCLAD 2008)*, 70. Setúbal: ESCE-IPS eds.

[9] Marques, A., Sousa Ferreira, A. e Cardoso, M. (2008). *Combining Models in Discrete Discriminant Analysis in the Multiclass Case.* Em *Book of Abstract - COMPSTAT 2008*, (FEUP eds), 270. Porto: INE.

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

[11] Raftery, A. E. (1996). *Approximate Bayes factor and accounting for model uncertainty in generalised linear models. Biometrika*, 83, 251-266.

[12] Smith, P., Wolpert, D. (1999).*Linearly combining density estimators via stacking. Machine Learning*, 36, 59-83.

[13] Sousa Ferreira, A. (2000). *Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas.* Tese de Doutoramento, Universidade Nova de Lisboa.

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

[15] Wolpert, D. (1992). *Stacked Generalization. Neural Networks*, 5, 241-259.

# CHAPTER 3

## Classification and Combining Models

This paper has the following reference:

Marques, A.; Sousa Ferreira, A. and Margarida G. M. S. Cardoso (2010) 'Classification and Combining Models'. In Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010)'. (Published in CD-rom).

**Erratum**

In pag. 2, where is "and DTM - Dependence Trees Model (DTM), Celeux (1994)" should be "and Dependence Trees Model (DTM), Celeux (1994)".

# Classification and Combining Models

**Anabela Marques, Ana Sousa Ferreira and Margarida Cardoso**

ESTBarreiro, Setúbal Polytechnic, Portugal, CEAUL

Email: anabela.marques@estbarreiro.ips.pt

LEAD, Faculty of Psychology, University of Lisbon, Portugal, CEAUL

Email: asferreira@fp.ul.pt

UNIDE, Dep. of Quantitative Methods of ISCTE - Lisbon University Institute, Portugal

Email: margarida.cardoso@iscte.pt

**Abstract:** In the context of Discrete Discriminant Analysis (DDA) the idea of combining models is present in a growing number of papers aiming to obtain more robust and more stable models. This seems to be a promising approach since it is known that different DDA models perform differently on different subjects. Furthermore, the idea of combining models is particularly relevant when the groups are not well separated, which often occurs in practice. Recently, we proposed a new DDA approach which is based on a linear combination of the First-order Independence Model (FOIM) and the Dependence Trees Model (DTM). In the present work we apply this new approach to classify consumers of a Portuguese cultural institution. We specifically focus on the performance of alternative models' combinations assessing the error rate and the Huberty index in a test sample.

We use the R software for the algorithms' implementation and evaluation.

**Keywords:** Combining models, Dependence Trees model, Discrete Discriminant Analysis, First Order Independence model.

## 1. Introduction

Discrete Discriminant Analysis (DDA) is a multivariate data analysis technique that aims to classify and discriminate multivariate observations of discrete variables into *a priori* defined groups (a known data structure or Clustering Analysis results). Considering K exclusive groups $G_1$, $G_2$, …, $G_K$ and a n-dimensional sample of multivariate observations - $X = (x_1, x_2, …, x_n)$ described by P discrete variables - DDA has two main goals:

1. To identify the variables that best differentiate the K groups;

2. To assign objects whose group membership is unknown to one of the K groups, by means of a classification rule.

In this work, we focus in the second goal and we consider objects characterized by qualitative variables (not necessarily binary) belonging to K ≥ 2 *a priori* defined groups. We propose to use the combination of two DDA models: FOIM - First-Order Independence Model and DTM - Dependence Trees Model (DTM), Celeux (1994) - to address classification problem.

In addition, we evaluate HIERM - Hierarchical Coupling Model performance when addressing the multiclass classification problems (Sousa Ferreira *et al*. (2000))

In order to evaluate the performance of the proposed approaches, we consider both simulated data and real data referred to consumers of a Portuguese cultural institution (Centro Cultural de Belém). Furthermore, we compare the obtained results with CART - Classification and Regression Trees (Breiman et al. (1984)) algorithm results.

## 2. Discrete Discriminant Analysis

The most commonly used classification rule is based on the Bayes's Theorem. It enables to determine the *a posteriori* probability of a new object being assigned to one of the *a priori* known groups. The Bayes's rule can be written as follows: if

$$\pi_k P_k(\underline{x}|\mathrm{G_k}) \geq \pi_l P_l(\underline{x}|\mathrm{G_l}) \text{ for l =1, …, K and } l \neq k, \quad (1)$$

then assign $\underline{x}$ to $G_k$. $\pi_l$ represents the *a priori* probability of group l ($G_l$), and $P(\underline{x}|G_l)$ denotes the conditional probability function for the *l*-th group. Usually, the conditional probability functions are unknown and estimated based on the training sample.

For discrete data, the most natural model is to assume that $P(\underline{x}|G_l)$ are multinomial probabilities estimated by the observed frequencies in the training sample, the well known Full Multinomial Model (FMM) (Celeux (1994)). However, since this model involves the estimation of many parameters, there are often related identifiability issues, even for moderate P. One way to deal with this high-dimensionality problem consists of reducing the number of parameters to be estimated recurring to alternative models proposals. One of the most important DDA models is the First-Order Independence Model (FOIM) (Celeux (1994)). It assumes that the P discrete variables are independent within each group $G_k$, the corresponding conditional probability being estimated by:

$$\hat{P}\left(\underline{x}|G_k\right) = \prod_{p=1}^{P} \frac{\#\left\{\underline{y} \in G_k : y_p = x_p\right\}}{n_k} \qquad (2)$$

where $n_k$ represents the $G_k$'s group sample dimension. This method is simple but is not realistic in some situations. Thus, some alternative models have been proposed. The Dependence Trees Model (DTM), Celeux (1994) and Pearl (1988), for example, takes the predictors' relations into account. In this model, one can estimate the conditional probability function, using a dependence tree that represents the most important predictors' relations. In this research, we use the Chow and Liu algorithm (Celeux (1994) and Pearl (1988)) to implement the dependence tree and approximate the conditional probability function.

In this algorithm, the mutual information $I(X_i, X_j)$ between two variables

$$I\left(X_i, X_j\right) = \sum_{X_i} \sum_{X_j} P\left(X_i, X_j\right) log \frac{P(X_i, X_j)}{P(X_i,)P(X_j)} \qquad (3)$$

is used to measure the closeness between two probability distributions. For example, take P = 4 variables and consider the data listed in Table 2. For each pair of variables we obtain the mutual information (see Table 1). Since $I(x_2, x_3)$, $I(x_1, x_2)$ and $I(x_2, x_4)$ correspond to the three largest values the branches of the best dependence tree are $(x_2, x_3)$, $(x_1, x_2)$ and $(x_2, x_4)$ and

$$\hat{P}\left(\underline{x}|G_k\right) = P(x_2)P(x_3|\, x_2)P(x_2|\, x_1)P(x_4|\, x_2) \qquad (4)$$

Table 2 illustrate the differences between the estimates of the 3 referred DDA models corresponding to the data considered. Note that the DTM model estimates are closer to the FMM estimates but there are no null frequencies.

| **($x_i$, $x_j$)** | **I ($x_i$, $x_j$)** |
|---|---|
| ($x_1$, $x_2$) | **0,079434** |
| ($x_1$, $x_3$) | 0,000051 |
| ($x_1$, $x_4$) | 0,005059 |
| ($x_2$, $x_3$) | **0,188994** |
| ($x_2$, $x_4$) | **0,005059** |
| ($x_3$, $x_4$) | -0,024 |

Table 1. Mutual information values

| (x1,x2,x3,x4) values | num. observ./ $G_k$ | $\widehat{P}$ (x1,x2,x3,x4) for group $G_k$ | | |
|---|---|---|---|---|
| | | FMM | FOIM | DTM |
| 0000 | 10 | 0,10 | 0,05 | 0,10 |
| 0001 | 10 | 0,10 | 0,05 | 0,13 |
| 0010 | 5 | 0,05 | 0,06 | 0,03 |
| 0011 | 5 | 0,05 | 0,06 | 0,04 |
| 0100 | 0 | 0,00 | 0,06 | 0,02 |
| 0101 | 0 | 0,00 | 0,06 | 0,02 |
| 0110 | 10 | 0,10 | 0,07 | 0,08 |
| 0111 | 5 | 0,05 | 0,07 | 0,07 |
| 1000 | 5 | 0,05 | 0,06 | 0,04 |
| 1001 | 10 | 0,10 | 0,06 | 0,05 |
| 1010 | 0 | 0,00 | 0,07 | 0,01 |
| 1011 | 0 | 0,00 | 0,07 | 0,02 |
| 1100 | 5 | 0,05 | 0,07 | 0,04 |
| 1101 | 5 | 0,05 | 0,07 | 0,03 |
| 1110 | 15 | 0,15 | 0,08 | 0,18 |
| 1111 | 15 | 0,15 | 0,08 | 0,15 |

Table 2. Conditional probability estimates for group $G_k$

**3. Combining Models in Discrete Discriminant Analysis**

The idea of combining models currently appears in an increasing number of papers. The aim of this strategy is to obtain more robust and stable models. Sousa Ferreira (2000) proposes combining FMM and FOIM to address classification problems with binary predictors, using a single coefficient β ($0 \leq \beta \leq 1$) to weight these models. In spite of yielding good results, the referred approach shows that the resulting FMM weights tend to be frequently negligible, even when the observed frequencies are smoothed (Brito *et al*. (2006)).

In view of these conclusions, Marques *et al*. (2008) propose a new model which has an intermediate position between the FOIM and DTM models:

$$\hat{P}(\underline{x}|\beta) = \beta\hat{P}_{FOIM}(\underline{x}) + (1-\beta)\hat{P}_{DTM}(\underline{x}) \qquad (5)$$

In the present work the combining models' parameter is assigned different values ranging from 0 to 1.

## 4. The Hierarchical Coupling Model

In the multiclass case (K≥2) we can recur to the Hierarchical Coupling Model (HIERM) (Sousa Ferreira *et al*. (2000)) that decomposes the multiclass problem into several biclass problems using a binary tree structure. It implements two decisions at each level:

1. Binary branching criterion for selecting among the possible $2^{K-1}$-1groups combinations;
2. Choice of the model or combining model that gives the best classification rule for the chosen couple.

In the present work we implement branching using the affinity coefficient, Matusita (1955) and Bacelar-Nicolau (1985). Supposing $F_1=\{p_l\}$ and $F_2=\{q_l\}$, $l$=1,…,L are two discrete distributions defined on the same space, the correspondent affinity coefficient is computed as follows:

$$\rho(F_1, F_2) = \sum_{l=1}^{L} \sqrt{p_l}\sqrt{q_l} \qquad (6)$$

The process stops when a decomposition of groups leads to single groups.

For each biclass problem we consider FOIM, DTM or an intermediate position between them.

## 5. Numerical Experiments

We conduct numerical experiments for simulated data and real data using moderate and large samples, respectively. We use test samples to evaluate the alternative models precision. Indicators of precision are the percentage of correctly classified observations ($P_c$) and the Huberty index:

$$\frac{P_{c-}P_d}{1-P_d}$$

where $P_d$ represents the percentage of correctly classified cases using the majority class rule.

### 5.1 Simulated data

The simulated data set considered has 250 observations, 4 groups and 3 binary predictors (see Table 3). To evaluate the proposed models' performance we use precision corresponding to a test (sub)sample: 50% of the original sample. The modal class in the test sample has 32% of the observations which yields the same 32% for percentage of correctly classified observations, according to the majority rule.

| | **Total data set** | | **Training sample** | | **Test sample** | |
|---|---|---|---|---|---|---|
| | **$n_k$** | **%** | **$n_k$** | **%** | **$n_k$** | **%** |
| $G_1$ | 80 | 32% | 40 | 32% | 40 | **32%** |
| $G_2$ | 70 | 28% | 35 | 28% | 35 | 28% |
| $G_3$ | 30 | 12% | 15 | 12% | 15 | 12% |
| $G_4$ | 70 | 28% | 35 | 28% | 35 | 28% |

Table 3. Characterization of simulated data set

The results obatined are presented in Table 4. For this data set the HIERM model and FOIM-DTM combination yeld the best results.

| **Classification Method** | | **% of correctly classified** | **Huberty index** |
|---|---|---|---|
| CART | | 45,6% | 20,00% |
| β*FOIM+ (1-β)*DTM | β = 0 (DTM) | **52,8%** | **30,59%** |
| | β = 0,25 | 47,2% | 22,35% |
| | β = 0,50 | 48,8% | 24,71% |
| | β = 0,75 | 48,8% | 24,71% |
| | β = 1 (FOIM) | 48,8% | 24,71% |
| MHIERM: | β = 0 (DTM) | 45,6% | 20,00% |
| $G_2$+$G_1$ vs $G_3$+$G_4$ | β = 0,25 | 59,2% | 40,00% |
| | β = 0,50 | **60,8%** | **42,35%** |
| β*FOIM+ | β = 0,75 | **60,8%** | **42,35%** |
| (1-β)*DTM | β = 1 (FOIM) | 59,2% | 40,00% |

Table 4. Simulated data set results

### 5.2 Real data

We consider a data set referred to 988 observations originated from questionnaires directed to consumers of Centro Cultural de Belém, a Portuguese cultural institution (Duarte (2009)). Data includes three questions related to the quality of services provided by CCB that this study tries to relate with consumers' education: we specifically use 4 education levels as the target variable. Predictors are: $X_1$-Considering your expectations how do you evaluate CCB products and services?(1=much worse than expected …5=much better than expected); $X_2$- How do you evaluate CCB global quality? (1=very bad quality,…, 5=very good quality); $X_3$: How do you evaluate the price/quality relationship in CCB?(1=very bad…5=very good). The groups distribution is illustrated in Table 5.

| | **Total data set** | | **Training sample** | | **Test sample** | |
|---|---|---|---|---|---|---|
| | $\mathbf{n_k}$ | **%** | $\mathbf{n_k}$ | **%** | $\mathbf{n_k}$ | **%** |
| $G_1$ | 177 | 18% | 115 | 18% | 62 | 18% |
| $G_2$ | 136 | 14% | 88 | 14% | 48 | 14% |
| $G_3$ | 462 | 47% | 300 | 47% | 162 | 47% |
| $G_4$ | 213 | 22% | 138 | 22% | 75 | 22% |

Table 5. Characterization of CCB data set

The results obtained are presented in Table 6. For CCB problem the best solution is achieved by HIERM model and combined FOIM-DTM model.

| Classification Method | | % of correctly classified | Huberty index |
|---|---|---|---|
| CART | | 46,10% | -1,70% |
| β*FOIM+ (1-β)*DTM | β = 0 (DTM) | 45,00% | -3,77% |
| | β = 0,20 | 45,80% | -2,26% |
| | β = 0,40 | 46,40% | -1,13% |
| | β = 0,50 | 47,60% | 1,13% |
| | β = 0,60 | 47,30% | 0,57% |
| | β = 0,80 | **47,80%** | **1,51%** |
| | β = 1 (FOIM) | 47,00% | 0,00% |
| MHIERM: G$_2$ vs G$_1$+G$_3$+G$_4$ β*FOIM+ (1-β)*DTM | β = 0 (DTM) | 47,80% | 1,51% |
| | β = 0,20 | 48,10% | 2,08% |
| | β = 0,40 | 49,30% | 4,34% |
| | β = 0,50 | 49,30% | 4,34% |
| | β = 0,60 | 49,30% | 4,34% |
| | β = 0,80 | 48,40% | 2,64% |
| | β = 1 (FOIM) | **49,90%** | **5,47%** |

Table 6. CCB data set results (test sample)

## 6. Conclusions and perspectives

In the present work we propose using the combination of two DDA models (FOIM and DTM) and use the HIERM algorithm to address classification problems. We compare results obtained with CART results into two data sets: simulated data (250 observations) and real data (988 observations). We use indicators of classification precision obtained in the test set (we consider 50% and 35% of observations for the smaller and larger data set, respectively).

According to the obtained results, the proposed approach performs slightly better than CART, on both simulated and real data. However, the classification precision attained barely attains the precision corresponding to the majority class rule in the real data set. In fact, in this case we are dealing with very sparse data (46% of the multinomial cells have no observed data in any of the groups considered) which turns the classification task very difficult.

In future research, the number of numerical experiments should be increased, both using real and simulated data sets and considering several sample dimensions. The number of variables considered (binary and non-binary) should not originate an excessive number of states (around a thousand) due to the number of parameters that need to be estimated. Alternative strategies to estimate the β parameter, such as least squares regression, likelihood ratio or committee of methods, should also be considered.

## References

1. Bacelar-Nicolau, H., The Affinity Coefficient in Cluster Analysis, in *Meth. Oper. Res.*, **53**: 507-512 (1985).
2. Breiman, L., Friedman, J.H., Olshen, R. A. and Stone, C.J., *Classification and Regression Trees*, Wadsworth, Inc. California (1984).
3. Brito, I., Celeux, C. and Sousa Ferreira, A., Combining Methods in Supervised Classification: a Comparative Study on Discrete and Continuous Problems. *Revstat – Statistical Journal*, Vol. 4(3**)**, 201-225 (2006).
4. Celeux, G. and Nakache, J. P., Analyse Discriminante sur Variables Qualitatives. G. Celeux et J. P. Nakache Éditeurs, *Polytechnica,* (1994).
5. Duarte, A., *A satisfação do consumidor nas instituições culturais. O caso do Centro Cultural de Belém.*Master Thesis. ISCTE, Lisboa (2009)
6. Marques, A.; Sousa Ferreira, A. and Cardoso, M. 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 (2008).
7. Matusita, K., Decision rules based on distance for problems of fit, two samples and estimation. In *Ann. Inst. Stat. Math.*, Vol. 26(4): 631-640, (1955).
8. Pearl J., *Probabilistic reasoning in intelligent systems: Networks of plausible inference.*.Los Altos: Morgan Kaufmann, (1988).
9. Sousa Ferreira, A., *Combinação de Modelos em Análise Discriminante sobre Variáveis Qualitativas*. PhD thesis. Universidade Nova de Lisboa, (2000).
10. Sousa Ferreira, A.; Celeux, G. and Bacelar-Nicolau, H., 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*, pages 181-186. Springer, (2000).

# CHAPTER 4

## Selection of variables in Discrete Discriminant Analysis

This paper has the following reference:
Marques, A.; Sousa Ferreira, A. and Margarida G. M. S. Cardoso (2013) ’Selection of variables in Discrete Discriminant Analysis’. Biometrical Letters. Vol.50, No 1, pp.1-14.

**Erratum**

In pag. 9, where is "There are two target classes: retail channel ($C_1$) and Horeca (Hotel, Restaurant and Caf channel) ($C_1$)" should be "There are two target classes: retail channel ($C_1$) and Horeca (Hotel, Restaurant and Caf channel) ($C_2$)".

In pag.5, where is "$2M - 1$" should be "$2^M - 1$".

DOI:10.2478/bile-2013-0013

*Biometrical Letters*
*Vol. 50 (2013), No. 1, 1-14*

# 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 $i^{th}$ 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 $\chi^2$ 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 $\alpha$ value for the entire set of $M$ tests equal to $\alpha$ by taking the $\alpha$ value for each test equal to $\alpha/M$.

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:

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

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:

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

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

$$\begin{aligned} \hat{f}_k\left(\mathbf{x}^* | X\right) = & \; f^{C_k}\left(\mathbf{x}^* | X\right) = \\ = & \; \hat{f}(x_1^* | X) \hat{f}(x_2^* | x_1^*, X) \hat{f}(x_3^* | x_2^*, X) \hat{f}(x_4^* | x_2^*, X) \hat{f}(x_5^* | x_2^*, X) \end{aligned} \quad (8)$$

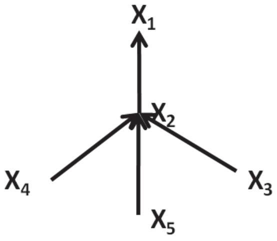


**Figure 1.** Example of a dependence tree for the case of M=5 variables

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:

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

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:

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

where $\pi_k$ 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 $\pi_k$, 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 $\alpha$ 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 ($\beta$ 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

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | |
|---|---|---|---|---|---|
| $\beta$ | 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

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | |
|---|---|---|---|
| $\beta$ | 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

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | |
|---|---|---|---|---|
| $\beta$ | 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.

# CHAPTER 5

## Combining Models in Discrete Discriminant Analysis

This paper has the following reference:

Marques, A.; Sousa Ferreira, A. and Margarida G. M. S. Cardoso (2014) ’Combining Models in Discrete Discriminant Analysis’, International Journal of Data Analysis Techniques and Strategies. (to appear)

# Combining Models in Discrete Discriminant Analysis

## Anabela Marques

Barreiro College of Technology, Setúbal Polytechnic,
Rua Américo da Silva Marinho - Lavradio, 2839-001 Barreiro, Portugal
E-mail: anabela.marques@estbarreiro.ips.pt

## Ana Sousa Ferreira

LEAD, Faculty of Psychology, University of Lisbon and UNIDE,
Alameda da Universidade, 1649-013 LISBOA, Portugal
E-mail: asferreira@fp.ul.pt

## Margarida G. M. S. Cardoso

UNIDE and Department of Quantitative Methods of ISCTE - Lisbon University Institute, Avenida das Forças Armadas, 1649-026 Lisboa, Portugal
E-mail: margarida.cardoso@iscte.pt

**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:** Discrete Discriminant Analysis; Combining models; Dependence Trees model; First-Order Independence model; Hierarchical Coupling model; Random Forest.

**Reference** to this paper should be made as follows: Marques, A.; Sousa Ferreira, A. and Margarida G. M. S. Cardoso (xxxx) 'Combining Models in Discrete Discriminant Analysis', *International Journal of Data Analysis Techniques and Strategies*, Vol. x, No. x, pp.xxx–xxx.

**Biographical notes:** Anabela Marques is an Adjunct Professor in the Barreiro College of Technology, Setúbal Polytechnic, Portugal. She is a Phd student and holds a degree in Statistics and Operational Research and Master in Probability and Statistics (both in FC-University of Lisbon). Her research domains include Multivariate Data Analysis, Supervised Classification, Teaching Statistics.

Ana M. P. N. Sousa Ferreira is an Assistant Professor at the Faculty of Psychology of the University of Lisbon, Portugal. She holds a degree in Mathematics - Scientific Branch of Statistics, Operational Research and Computing (FC-University of Lisbon), a Master in Probability and Statistics and a PhD in Mathematics in the Specialization of Statistics (FCT-New University of Lisbon). Multivariate Data Analysis is her general area of research and Supervised Classification, Teaching Statistics and Data Analysis of Large Scale Surveys in Education constitute the main research domains, in particular Psychology and Education stand out as areas of applications.

Margarida G. M. S. Cardoso is an Associate Professor at the Business School of ISCTE- Lisbon University Institute, Portugal. She holds a degree in Mathematics (FC-University of Lisbon), a Master in Operations Research and Systems Engineering and a Phd in Systems Engineering (IST-University of Lisbon). Her research domains include Data Analysis using Multivariate Statistics and Data Mining - Clustering and Classification techniques in particular. Marketing Research stands out as an area of application.

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

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:

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

where $\pi_k$ represents the *prior* 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 $\pi_k$, 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 resorting to multiple models, enhancing 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 and 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 multinomial. 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:

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

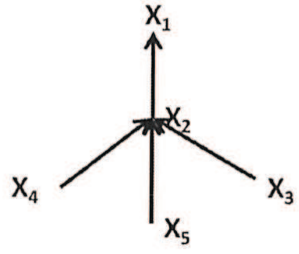
where $x_{l(p)}$ represents a variable that is linked to the variable $x_p$ in this graph, arbitrarily choosing one vertex as the root of the graph, $x_{r(p)}$.
To construct the graph for each class, we rely on the algorithm of Chow and Liu (Celeux and Nakache, 1994; Pearl, 1988), where the length of 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:

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

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

**Figure 1** Example of a dependence tree for the case of P=5 variables



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

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

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 (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 $\beta$, $(0 \leq \beta \leq 1)$ 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 performance, with intermediate 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:

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

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:

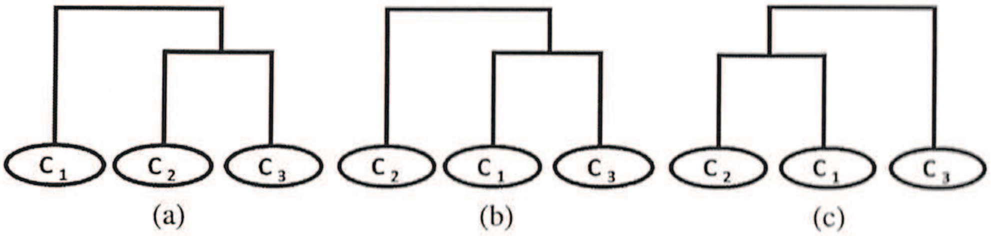
1. Selection of the hierarchical coupling among the $2^{K-1}-1$ possible class couples;
2. Choice of the model or combining model that gives the best classification rule for the chosen couple.

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).

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

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 3 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 (Sousa Ferreira, 2010).

**Figure 2** Binaries trees in the HIERM 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:

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

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, 1998; 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 model, 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 ($\phi$) 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 the FOIM-DTM combination to solve DDA problems. In addition, when multiple classes are considered, we suggest using HIERM and also recurring to the FOIM-DTM combination to obtain intermediate classification results in each tree node. Regarding the combination coefficient $\beta$, 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 $P$ Random Forests derived from subsets of features with dimension ranging from 1 to $P$, for branching. Finally, we combine all the 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 previous studies (Sousa Ferreira, 2010; Sousa Ferreira et al., 2000). In order to simulate the predictive binary variables' values, this model defines class conditional probabilities for $C_k, (k = 1, ..., K)$ as

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

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

$$Z_{kp} = \frac{X_{kp} - \theta_{kp}}{[\theta_{kp}(1 - \theta_{kp})]^{1/2}} \quad \textit{and} \quad \rho_k(p,g) = E(Z_{kp}Z_{kg}), \tag{12}$$

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).
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 $p \neq g,\ p, g = 1, ..., 6$.

**Table 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 *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)(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.
  $(CF+C)-FC$ offers 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 $CF+C$;
- $V>0$ - 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 $(FC', C'F, C')$.

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 $\phi$ 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 (moderate 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.

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $n = 400$ | | | $n = 120$ | | |
| | $EFF_{Train}$ | $EFF_{Test}$ | $\phi_{Test}$ | $EFF_{Train}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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.

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $n = 400$ | | | $n = 120$ | | |
| | $EFF_{Train}$ | $EFF_{Test}$ | $\phi_{Test}$ | $EFF_{Train}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $n = 800$ | | | $n = 240$ | | |
| | $EFF_{Train}$ | $EFF_{Test}$ | $\phi_{Test}$ | $EFF_{Train}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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

| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $n = 800$ | | | $n = 240$ | | |
| | $EFF_{Train}$ | $EFF_{Test}$ | $\phi_{Test}$ | $EFF_{Train}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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

| | $HIERM: \ \beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $n = 800$ | | | $n = 240$ | | |
| | $C_4 vs. C_1 \cup C_2 \cup C_3$ e $C_1 vs. C_2 \cup C_3$ | | | $C_4 vs. C_1 \cup C_2 \cup C_3$ e $C_3 vs. C_1 \cup C_2$ | | |
| | $EFF_{Train}$ | $EFF_{Test}$ | $\phi_{Test}$ | $EFF_{Train}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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

| | $HIERM: \ \beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $n = 800$ | | | $n = 240$ | | |
| | $C_2 vs. C_1 \cup C_3 \cup C_4$ e $C_4 vs. C_1 \cup C_3$ | | | $C_3 vs. C_1 \cup C_2 \cup C_4$ e $C_1 vs. C_2 \cup C_4$ | | |
| | $EFF_{Train}$ | $EFF_{Test}$ | $\phi_{Test}$ | $EFF_{Train}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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 separable 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.

**Figure 3** Binary Tree for the Alexithymia data

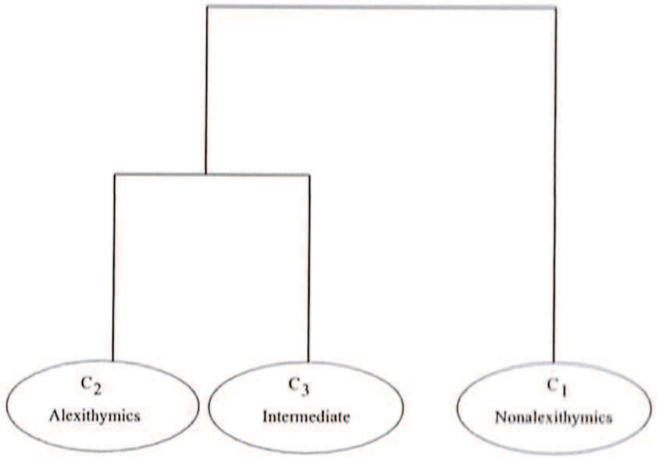


**Table 8** Classification performance: real data

| $\beta$ | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | |
|---|---|---|---|---|
| | $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | $HIERM : C_1\ vs.\ C_2 \cup C_3$ | |
| | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ | $EFF_{2-Fold}$ | $\phi_{2-Fold}$ |
| 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 $\phi$ measures

| 2 Classes | n | EFF | $\phi$ | 4 Classes | n | EFF | $\phi$ |
|---|---|---|---|---|---|---|---|
| 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 unbalanced data sets too. Also, the use of more real data sets should further evidence the advantage of the proposed combined approach.

## References

Bacelar-Nicolau H. (1985) 'The Affinity Coefficient in Cluster Analysis'. *Methods Operations Research*. Vol. 53, pp.507–512.

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

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

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.

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

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

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.

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.

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

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

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

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

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

Goodman L. and Kruskal W. (1959) 'Measures of association for cross classifications. II: Further discussion and references'. *American Statistical Association Journal*. Vol. 54, No. 285, pp.123–163.

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.

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.

Kotsiantis S.B. (2013) 'Decision trees: a recent overview'. *Artificial Intelligence Review*. Vol. 39, No. 4, pp.261–283.

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

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

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

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

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.

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.

Marques A., Sousa Ferreira A. and Cardoso M. (2013) 'Selection of variables in Discrete Discriminant Analysis'. *Biometrical Letters*,Vol. 50, No. 1, pp.1–14.

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

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

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

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

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

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.

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

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

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.

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.

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.

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.

# CHAPTER 6

## Performance of combined models on binary discrete classification

This paper has the following reference:

Marques, A.; Sousa Ferreira, A. and Margarida G. M. S. Cardoso (2014) ’Performance of combined models on binary discrete classification’. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. (under review)

# Performance of combined models on binary discrete classification

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

**Abstract**: Diverse Discrete Discriminant Analysis (DDA) models perform differently on different samples. This fact has encouraged research in combined models which seems specially promising when the *a priori* classes are not well separated or when small or moderate sized samples are considered, which often occurs in practice. In this work, we evaluate the performance of a convex combination of two DDA models: the First-Order Independence Model (FOIM) and the Dependence Trees Model (DTM). We use simulated data sets with two classes and consider diverse data complexity factors which may influence the combined model's performance - the classes' separation, balance and number of missing states, as well as sample size and also the number of parameters to be estimated in DDA. We resort to cross-validation to evaluate the precision of classification.

The results obtained illustrate the advantage of the proposed combination when compared to FOIM and DTM: it yields the best results, specially when very small samples are considered. The experimental study conducted also provided the ranking of the data complexity factors, according to their relative impact on classification performance, resorting to a regression model. It lead to the conclusion that classes' separation is the most influent factor on classification performance. The ratio between the number of degrees of freedom and sample size, along with the proportion of missing states in the majority class, also have significant impacts on classification performance. An additional attainment of this study, also deriving from the estimated regression model, is the ability to successfully predict the precision of classification on real data set based on the data complexity factors.

**Keywords.** Discrete Discriminant Analysis, Separability, Classification performance, Combined models for classification.

## 1 Introduction

Some researchers have tried to understand the relationship between the data characteristics and the performance of classifiers. For example, Ho and Basu (2002), studied the case of two class problems and described the nature of classification difficulty. They enumerated diverse measures of a classification problem complexity and adopted a typology considering: 1) overlap of individual features, 2) measures of separability of classes and 3) measures of geometry. Sotoca, Sanchez, and Mollineda (2005) used those measures and add 4) statistical measures (e.g number of binary attributes, number of classes, entropy of

[1] Barreiro College of Tecnology, Setúbal Polytechnic, IPS, Portugal. E-mail:anabela.marques@estbarreiro.ips.pt

[2] Faculty of Psychology, University of Lisbon, Portugal, and UNIDE. E-mail:asferreira@psicologia.ulisboa.pt.

[3] UNIDE and Department of Quantitative Methods, ISCTE - University Institute of Lisbon, Portugal. E-mail:margarida.cardoso@iscte.pt

classes, mean absolute correlation coefficients between two features etc) when conducting a meta analysis of classifiers. Finch and Schneider (2007), considered the weight of classes, and three factors related to continuous predictors. Macia, Bernadó-Mansilla, and Orriols-Puig (2008), also used measures of geometry to characterize the complexity of data sets and studied binary classification. These authors considered several scenarios for synthetic continuous data, controlling the numbers of instances and the number of attributes and focused on the length of the class boundary to assess complexity of the data set.

Studies referring to the performance of classification based on nominal predictors are very rare in the literature. In this work we conduct numerical experiments to evaluate the performance of binary classifiers in Discrete Discriminant Analysis (DDA), aspiring to contribute to filling this gap in the literature. For this end, we set different scenarios using simulated data sets considering diverse data complexity factors. The generated data sets are meant to provide means to compare the performance of single and combined DDA models and to provide new insights concerning the impact of data complexity factors on discrete classification performance. In particular, we focus on DDA in very small, small and moderate sized samples, which turn classification tasks harder - Ho and Basu (2002) - and, we believe, discrete classification tasks even harder.

# 2 Methodology

## 2.1 A combined model for classification

In the present work we address Discrete Discriminant Analysis (DDA) tasks - to classify and discriminate multivariate observations of discrete variables into *a priori* defined classes - using a combined model proposed by Marques et al. (2013).

Generally, in supervised classification, several models are estimated and a unique classifier is selected based on some validation criterion. However, the discarded classifiers usually contain important information about the classification problem which is lost by selecting a single classifier (Brito et al., 2006). In addition, often it is observed that misclassified objects are different for different models. This fact has recently encouraged a large number of publications, from several areas of research, focused on the combination of classification models (e.g. Wolpert, 1992; Breiman, 1996, 1998; Freund and Shapire, 1996; Friedman et al., 1998; Sousa Ferreira et al., 2000; Friedman, 2001; Milgram, Sabourin and Cheriet, 2004; Brito, 2002; Kotsiantis et al., 2006; Cesa-Bianchi et al., 2006; Friedman and Popescu, 2008; Amershi and Conati, 2009; Janusz, 2010; Kotsiantis, 2011; Re and Valentini, 2011).

In the scientific literature the combining approach appears designated by several terms as, for instance, Blending by Elder and Pregibon (1995), Ensemble of Classifiers by Dietterich (1997), Committee of Experts by Steinberg (1997), Perturb and Combine (P&C) by Breiman (1996) and Combiners by Jain, Duin and Mao (2000). Nevertheless, all authors focused in a quite simple idea: train one model in several samples from the same data set or train several models from the same data and combine their output predictions usually using a voting process. Examples of the first strategy are Bagging (Breiman, 1996) using bootstrap samples of the training data set, Boosting (Freund & Schapire, 1996) weighting more heavily cases misclassified by decision tree models or Arcing (Breiman, 1998) weighting random subsamples of the training data set. On the other hand, training diverse types of models, can achieve uncorrelated output predictions and thus reduce the misclassification error rate (Abbot, 1999; Amershi and Conati, 2009Brito, 2002; Brito et al., 2006; Cesa-Bianchi et al., 2006; Janusz, 2010; Kotsiantis, 2011; Sousa Ferreira, 2000, 2004). Although many of the combined models for classification proposed in the literature can be applied to problems with discrete explanatory variables, studies in the literature heavily focus on continuous data. Therefore, we dedicate our research to combining models in DDA, a natural approach which usually increases classification performance (Sousa Ferreira, 2000, 2004, and 2010).

The model proposed is a linear convex combination of the First-Order Independence Model (FOIM) and the Dependence Trees Model (DTM):

- The First-Order Independence Model (FOIM) (Goldstein and Dillon, 1978) assumes that the P discrete variables are independent in each class $C_k$, $k = 1, ..., K$.
- The Dependence Trees Model (DTM) (Pearl, 1988; Celeux & Nakache, 1994) is an alternative model that takes the predictors relationships into account.

The corresponding conditional probability function is estimated as follows:

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

with $(0 \leq \beta \leq 1)$ .
Where $X = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_n)$, $\underline{x}_i$ represents the $i^{th}$ object ($i \in \{1, ..., n\}$), described by P discrete variables, $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{iP})$ (observed state), K exclusive classes $(C_1, C_2, ..., C_K)$ and a n-dimensional sample. For modelling purposes prior probabilities are considered equal.
The R software is used for the algorithm's implementation.

## 2.2 Data complexity and the performance of classifiers

The performance of classifiers can be influenced by several factors: classes separation, balance (Prati, Batista, & Monard, 2004; Macia, Bernadó-Mansilla, & Orriols-Puig, 2008; Ho & Basu, 2002), sample size (Raudys & Jain, 1991) and also (in the specific DDA domain), the number of missing states - e.g. (Sousa Ferreira, 2004, 2010). Some studies have addressed the relationships between more than one factor, namely when continuous predictors are considered - e.g. (Prati, Batista, & Monard, 2004) refer to overlapping and balance and conclude that the lack of separation between classes tends to surpass the importance of unbalanced classes in what regards the difficulty of binary classification tasks. Pinches (1980), points out the relevance of sample size and comments on the impact of unequal sample sizes per class. Raudys and Jain (1991) consider the relationship between sample size and the number of missing states and also underline the intrinsic relationship between the sample size and the number of predictors as a determinant of classification complexity. Macia, Bernadó-Mansilla, and Orriols-Puig (2008) resort to the generation of synthetic data sets to evaluate data complexity and find that the length of the classes' boundary is a dominant factor in assessing the complexity of the data set.
In the present work several scenarios are set for generating data to evaluate the impact of data characteristics in the performance of a discrete binary classifier. First, for a fixed number of predictors (four), we consider very small, small and moderate sized samples. The second experimental factor is the degree of classes' separation which is measured by the affinity coefficient ($A$) (Matusita, 1955; Bacelar-Nicolau, 1985). This coefficient is computed as follows:

$$A(\underline{f}, \underline{f}') = \sum_{l=1}^{L} \sqrt{f_l}\sqrt{f'_l} \tag{2}$$

where $\underline{f} = (f_1, ..., f_L)$ and $\underline{f}' = (f'_1, ..., f'_L)$, are two discrete distributions defined on the same states' space ($f_l$ stands for the relative frequency of the $l^{th}$ state, $l = 1, ..., L$).
The third experimental factor considered is balance - the weight of the majority class is used as its measure. The number of missing states is included as an additional complexity factor. This factor is not pre-specified but is determined for the simulated data sets generated under the experimental scenarios

(defined by the previously referred factors).

To evaluate the DDA results obtained with the combined model we report the percentage of correctly classified observations ($P_c$) and the Huberty Index ($HI$)

$$HI = \frac{P_c - P_d}{1 - P_d} \tag{3}$$

where $P_d$ represents the percentage of observations corresponding to the majority class and $P_c$ is the percentage of correctly classified cases. The Huberty index is intended to provide a fair comparison between the performance of both balanced and unbalanced cases since it quantifies the percentage of improvement in classification performance taking into account the majority class rule as a default classification rule. Two-fold results are reported for both measures of performance.

Finally, we attempt to model the relationship between the combined classifier performance and the complexity data factors considered in this work. For this end, we resort to simulated data and use regression on the combined model's performance. The percentage of correctly classified observations (two-fold result) is the response variable considered (note that since the weight of the majority class is included as a predictor, the Huberty Index can be discarded at this stage). The estimated linear regression model will be judged according to its fit to data and its predictive efficacy tested in one real data set.

# 3 Data analysis and results

## 3.1 Simulated Data

The performance of the FOIM, DTM and combined FOIM-DTM discrete classifiers is evaluated based on simulated data within diverse experimental scenarios. First, we focus on binary classification. Then we consider 4 binary predictors, a reasonable number taking into account we want to address classification on small sized samples. Having set this general scenario, we specify the following complexity factors: 1) separability - thresholds for the affinity coefficient values are above 0.7 for poorly separated classes, between 0.2 and 0.7 for moderately separated classes and under 0.2 for well separated classes; 2) sample size - $n = 60$, $n = 120$ and $n = 400$ samples sizes are considered; 3) balance - unbalanced classes refer to different sample proportions - $(1:2)$, $(1:3)$ and $(1:3)$ for $n = 60$, $n = 120$ and $n = 400$, respectively. The average of missing states (the fourth experimental factor) is finally quantified for each simulated data set.

The multinomial distribution parameters, along with the complexity factors' characteristics regarding the data sets considered are presented in Table 1 and Table 2.

For each of the eighteen resulting scenarios we generate 30 data sets. Based on the 540 generated data sets we aim at understanding the comparative advantage of the combined DDA model. In addition, we will be able to use a regression model in order to evaluate the relative impact of each factor - separability, balance, sample size, number of estimated parameters and number of missing states - on the performance of binary discrete classification.

Table 1: Synthetic datasets parameters: the 4 binary predictors' probabilities

| Separability | $C_1$ | $C_2$ |
|---|---|---|
| poor | $(0.5, 0.5; 0.5, 0.5; 0.5, 0.5; 0.5, 0.5)$ | $(0.5, 0.5; 0.5, 0.5); 0.5, 0.5; 0.5, 0.5)$ |
| moderate | $(0.4, 0.6; 0.6, 0.4; 0.4, 0.6; 0.6, 0.4)$ | $(0.7, 0.3; 0.3, 0.7; 0.7, 0.3; 0.3, 0.7)$ |
| good | $(0.1, 0.9; 0.7, 0.3; 0.2, 0.8; 0.6, 0.4)$ | $(0.9, 0.1; 0.3, 0.7; 0.8, 0.2; 0.1, 0.9)$ |

Table 2: Average numbers of missing states (30 runs in each scenario)

| Separation | $n = 60$ | | | $n = 120$ | | | $n = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | Total | $C_1$ | $C_2$ | Total | $C_1$ | $C_2$ | Total |
| balanced | | | | | | | | | |
| poor | 2.30 | 2.30 | 4.63 | 0.23 | 0.37 | 0.60 | 0.00 | 0.00 | 0.00 |
| moderate | 3.00 | 4.57 | 7.57 | 0.70 | 2.37 | 3.07 | 0.00 | 0.47 | 0.47 |
| good | 7.23 | 8.83 | 16.07 | 4.73 | 6.97 | 11.70 | 1.93 | 3.40 | 5.33 |
| unbalanced | | | | | | | | | |
| poor | 4.67 | 1.40 | 6.07 | 2.40 | 0.07 | 2.47 | 0.03 | 0.00 | 0.03 |
| moderate | 5.37 | 3.43 | 8.80 | 3.40 | 1.53 | 4.93 | 0.20 | 0.17 | 0.37 |
| good | 7.80 | 6.60 | 14.40 | 6.97 | 3.83 | 10.80 | 3.70 | 1.30 | 5.00 |

## 3.2 Real Data

A real data set is considered to compare the effective FOIM-DTM performance with the estimated performance based on the complexity factors considered, using an estimated regression model. It is based on the *Congressional Voting Records Data Set* in the UCI Machine Learning Repository - see Bache and Lichman (2013) - which includes votes for each of the U.S. House of Representatives Congressmen on 16 *key votes* identified by the Congressional Quaterly Almanac (CQA), 1984. In this data set classification is meant to discriminate between democrats (DEM) and republicans (REP). The 16 predictors (*key votes*) are binary variables indicating: 1- *yes*; 2- *no*. In this work we only consider individuals providing complete answers and finally select the four most discriminant predictors - we use the Cramer's V statistic measuring the association between each predictor and the classes to identify the most promising variables. In table 3, the final data set considered is described.

Table 3: Congressional voting records (reduced) data set

| Predictors' | category | DEM ($C_1$) | REP($C_2$) |
|---|---|---|---|
| V4. adoption-of-the-budget-resolution | 1-yes | 85.5% | 15.7% |
| | 2-no | 14.5% | 84.3% |
| V5. physician-fee-freeze | 1-yes | 4.8% | 99.1% |
| | 2-no | 95.2% | 0.9% |
| V6. el-salvador-aid | 1-yes | 20.2% | 95.4% |
| | 2-no | 79.8% | 4.6% |
| V13. education-spending | 1-yes | 12.9% | 85.2% |
| | 2-no | 87.1% | 14.8% |
| Total | 232 | 124 | 108 |

## 3.3 Results

### 3.3.1 Descriptive results

The descriptive results referring to the performance of the combined FOIM-DTM classifier are presented in this section. They refer to 30 classifications runs in each scenario.

The performance of the combined classifier FOIM-DTM in the 540 synthetic data sets is summarized on Table 4. Detailed results are provided in Table 5 and Table 6. When very small samples are considered the proposed combined classification algorithm is a clear winner - it outperforms FOIM and DTM in the 180 corresponding data sets. When $n = 120$ (small sized sample) FOIM and DTM are also able to deliver the best classification results for the balanced data sets. For $n = 400$ (moderate sized sample) the general winner classifier is FOIM, although the proposed combination may outperform FOIM in an unbalanced setting with poorly and moderately separated classes.

In general, unbalanced data sets correspond to harder classification tasks - see Huberty index values in Table 5 and Table 6. Also, there is a clear increase in classification performance associated with an increase in separation. Specifically, for the unbalanced data sets with poorly separated classes, the default classification precision overcomes the precision of the proposed algorithm. The performance results obtained are generally consistent (over the 30 runs in each scenario)- see the coefficient of variation values. However, the Huberty index may exhibit high variability when confronted with difficult classification tasks i.e. generally when poorly separated classes are considered and also when unbalanced and moderately separated classes are considered.

Table 4: Average $\beta$ coefficient referring to the best classifier (30 runs)

| Separation | $n = 60$ | $n = 120$ | $n = 400$ |
|---|---|---|---|
| balanced | | | |
| poor | 0.7 | 0 (DTM) | 1 (FOIM) |
| moderate | 0.8 | 0.9 | 1 (FOIM) |
| good | 0.9 | 1 (FOIM) | 1 (FOIM) |
| unbalanced | | | |
| poor | 0.6 | 0.9 | 0.5 and 0.9 |
| moderate | 0.6 | 0.8 | 0.6 and 0.7 |
| good | 0.9 | 0.9 | 1 (FOIM) |

Table 5: Average classifier's performance for balanced datasets (30 runs)

| | | $n = 60$ | | $n = 120$ | | $n = 400$ | |
|---|---|---|---|---|---|---|---|
| Separation | | mean | var. coef. | mean | var. coef. | mean | var. coef. |
| poor | $P_c$ | 56.6% | 0.12 | 53.8% | 0.07 | 51.7% | 0.06 |
| | $HI$ | 13.1% | 1.06 | 7.6% | 1.00 | 3.4% | 1.80 |
| moderate | $P_c$ | 72.7% | 0.11 | 72.7% | 0.07 | 74.3% | 0.03 |
| | $HI$ | 45.3% | 0.34 | 45.4% | 0.22 | 48.5% | 0.09 |
| good | $P_c$ | 92.7% | 0.04 | 92.7% | 0.28 | 92.8% | 0.02 |
| | $HI$ | 85.4% | 0.09 | 85.4% | 0.06 | 85.7% | 0.04 |

Table 6: Average classifier's performance for unbalanced datasets (30 runs)

| | | $n = 60$ $(1:2)$ | | $n = 120$ $(1:3)$ | | $n = 400$ $(1:3)$ | |
|---|---|---|---|---|---|---|---|
| Separation | | mean | var. coef. | mean | var. coef. | mean | var. coef. |
| poor | $P_c$ | 57.2% | 0.13 | 54.9% | 0.09 | 54.9% | 0.06 |
| | $HI$ | -28.5% | -0.79 | -80.3% | -0.25 | -80.5% | -0.16 |
| moderate | $P_c$ | 74.1% | 0.10 | 77.1% | 0.06 | 74.6% | 0.03 |
| | $HI$ | 22.2% | 1.04 | 8.2% | 2.21 | -1.7% | -5.94 |
| good | $P_c$ | 90.8% | 0.04 | 92.1% | 0.03 | 91.9% | 0.01 |
| | $HI$ | 72.4% | 0.16 | 68.4% | 0.18 | 67.4% | 0.07 |

### 3.3.2 Regression on performance

The performance results obtained in the numerical experiments conducted enable us to estimate a regression model in order to:

1. predict the $P_c$ measure of performance based on the data characteristics (/complexity factors);

2. understand the relative impact of each experimental complexity factor on performance.

To implement the regression we specifically consider the following measures of the experimental complexity factors: the affinity coefficient value - $A_{ff}$ - is used to measure the classes' separation; the weight of the majority class - $W_{mc}$ - is used to measure balance; dimensionality is measured by the ratio - $P_{df}$ - between the "number of degrees of freedom"and sample size, i.e. $P_{df} = (n - (P \times 2 + 1))/n$ (note that $P = 4$ is the number of predictors and we have to estimate parameters referred to two classes); finally, the proportions of missing states in each class - $P_{msc_1}$ and $P_{msc_2}$ - are considered.
A generalization of the Tobit regression model is used and the MLE estimated coefficients are obtained using the censReg package (Henningsen, 2010). The estimated regression model is presented in table 7. Additional columns in the right refer to standardized variables - these results are meant to help better evaluating the relative importance of predictors.
According to the non standardized as well as the standardized models, the three complexity factors having the larger impact on classification precision (by decreasing order) are: separation, ratio between the degrees of freedom and sample size and proportion of missing states in the minority class. The weight of majority class, the proportion of missing states in the majority class have a weaker impact on performance. In fact, according to the standardized coefficients ranking (an alternative modelling approach), the impact of the last factor is non-significant.
As expected, the larger the proportion of degrees of freedom the easiest the classification task is. The remaining factors have a negative impact on performance. The squared correlation between observed and estimated $P_c$ values is 0.95 evidencing a good fit to data.

Table 7: ML Estimated regression coefficients

| | Coef. | p-value | Coef. (Std.) | p-value (Std.) |
|---|---|---|---|---|
| constant | 0.923 | 0.000 | 0.240 | 0.000 |
| $A_{ff}$ | -0.692 | 0.000 | -1.017 | 0.000 |
| $P_{df}$ | 0.258 | 0.001 | 0.207 | 0.000 |
| $P_{msc_1}$ | -0.599 | 0.000 | -0.161 | 0.000 |
| $W_{mc}$ | 0.099 | 0.000 | 0.039 | 0.039 |
| $P_{msc_2}$ | -0.196 | 0.001 | -0.039 | 0.358 |

When applying the estimated regression model to the real data set (reduced *Congressional Voting Records*) we may anticipate the percentage of correctly classified observations based on its characteristics: affinity coefficient 0.195; proportion of missing states on the majority class 0.125; proportion of missing states on the minority class 0.281; ratio between degrees of freedom and sample size 0.961; and balance 0.534. In fact, before performing classification we could foresee $\widehat{P_c} = 95.9\%$ based on the estimated regression model (see coefficients in Table 7) and, according to the classification results obtained with the combined model FOIM-DTM on this data set, the actual percentage of correctly classified observations is $P_C = 95.7\%$.

## 4 Conclusions and Perspectives

In the present work, we evaluate the performance of a combined model - a convex combination of FOIM and DTM - for binary discrete classification. We set 18 scenarios for generating simulated data sets with 4 binary predictors controlling for factors considered relevant for classification precision. These factors include three degrees of classes' separability, classes' weights (balanced or not) and sample dimension ($n = 60$, $n = 120$, $n = 400$). In addition, the number of missing states is quantified in each scenario.

The differentiated scenarios provided very different classification performances. According to the obtained results, the combined method achieves the best results for small sample cases (whether balanced or unbalanced) and performance improves with the increase of classes' separability, as expected. The worst performances are registered for unbalanced and poorly separated classes - the combined model is unable to surpass default classification precision (the lowest Huberty Index value is -80.5%). Within the balanced scenario, when moderately separated classes are considered, the increase of the sample dimension increases the classification ability of the single FOIM model. For unbalanced data sets, the proposed combination generally achieves the best results obtained.

Based on experimental data - 30 classification runs in each scenario - a regression model is estimated which provides new insights regarding the relative impact of experimental factors on binary discrete classification precision. Separability turns out to be the most important experimental factor - the more weakly separated the classes are (the higher the affinity coefficient) the weaker the classification performance is. The proportion of the number of degrees of freedom vs. sample size is the second most important factor, with a positive impact on performance. The third one is the proportion of missing states in the minority class and it has a negative impact on performance, as expected.

The estimated regression model exhibited a good fit to synthetic data and also enabled to anticipate the performance of the proposed FOIM-DTM algorithm on a real data set - a data set extracted from the *Congressional Voting Records Data Set* in the UCI Machine Learning Repository. In this data set, the difference between the estimated and the actual measure of performance (percentage of correctly classified observations) is 0.002.

To our knowledge, this type of study is the first conducted for evaluating DDA performance. In future research, additional measures of complexity of discrete classification problems may be considered - e.g an alternative measure of the degree of classes' separability (other than the affinity coefficient). Also, some of the experimental factors that were taken into account may vary their categories, and their interaction may be further analyzed.

## References

Abbott, D. W. (1999). Combining models to improve classifier accuracy and robustness. In Proceedings of Second International Conference on Information Fusion, Fusion'99 (Vol. 1, pp. 289-295).

Amershi, S., & Conati, C. (2009). Combining Unsupervised and Supervised Classification to Build User Models for Exploratory. JEDM-Journal of Educational Data Mining, 1(1), 18-71.

Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis. Methods of operations research, 53, 507-512.

Bache, K., & Lichman, M. (2013). UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (1998). Half & Half bagging and hard boundary points. Technical Report, Statistics Department, University of California.

Brito, I. (2002). Combinaison de modèles en analyse discriminante dans un contexte gaussien (Doctoral dissertation, Grenoble 1).

Brito, I., Celeux, G., & Sousa Ferreira, A. (2006). Combining methods in supervised classification: A comparative study on discrete and continuous problems. REVSTAT - Statistical Journal, 4(3), 201-225.

Celeux, G., & Nakache, J. P. (1994). Analyse discriminante sur variables qualitatives. Politechnica.

Cesa-Bianchi, N., Claudio G., & Luca Z.(2006). Hierarchical classification: combining Bayes with SVM. Proceedings of the 23rd international conference on Machine learning. ACM.

Dietterich, T. G. (1997). Machine-learning research. AI magazine, 18(4), 97.

Elder, J.F., & Pregibon, D. (1995). A Statistical Perspective on Knowledge Discovery in Databases. Advances in Knowledge Discovery and Data Mining. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Editors. AAAI/MIT Press.

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3(2), 47-57.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In ICML (Vol. 96, pp. 148-156).

Friedman, J. H., Hastie, T. & Tibsharani, R. (1998). Additive Logistic Regression: A Statistical View of Boosting. Technical Report, Stanford University.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189-1232.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. The Annals of Applied Statistics, 916-954.

Goldstein, M., & Dillon, W. R. (1978). Discrete discriminant analysis. New York: Wiley.

Henningsen, A. (2010). Estimating Censored Regression Models in R using the censReg Package. R package vignettes.

Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. Pattern Analysis and Machine Intelligence, IEEE Transactions on,24(3), 289-300.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(1), 4-37.

Janusz, A. (2010). Combining multiple classification or regression models using genetic algorithms. Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. Artificial Intelligence Review, 35(3), 223-240.

Macia, N., Bernadó-Mansilla, E., & Orriols-Puig, A. (2008, December). Preliminary approach on synthetic data sets generation based on class separability measure. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on (pp. 1-4). IEEE.

Marques, A., Sousa Ferreira, A., & Cardoso, M. G. (2013). Selection of variables in Discrete Discriminant Analysis. Biometrical Letters, 50(1), 1-14.

Matusita, K. (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. The Annals of Mathematical Statistics, 631-640.

Milgram, J., Cheriet, M., & Sabourin, R. (2004, October). Speeding up the decision making of support vector classifiers. In Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on (pp. 57-62). IEEE.

Milgram, J., Mohamed C., & Robert S. (2004).Speeding up the decision making of support vector classifiers. Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on. IEEE.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.

Pinches, G. E. (1980). Factors influencing classification results from multiple discriminant analysis. Journal of Business Research, 8(4), 429-456.

Prati, R. C., Batista, G. E., & Monard, M. C. (2004). Class imbalances versus class overlapping: an analysis of a learning system behavior. In MICAI 2004: Advances in Artificial Intelligence (pp. 312-321). Springer Berlin Heidelberg.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on pattern analysis and machine intelligence, 13(3), 252-264.

Re, M., & Valentini, G. (2011). Ensemble methods: a review.

Sousa Ferreira, A., Celeux, G., & Bacelar-Nicolau, H. (2000). Discrete Discriminant Analysis: The Performance of Combining Models by a Hierarchical Coupling Approach. In Data Analysis, Classification, and Related Methods (pp. 181-186). Springer Berlin Heidelberg.

Sousa Ferreira, A. (2004). Combining Models in Discrete Discriminant Analysis Through a Committee of Methods. In Classification, Clustering, and Data Mining Applications (pp. 151-156). Springer Berlin Heidelberg.

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.)(pp. 137-145). Springer-Verlag, Berlin-Heidelberg.

Sotoca, J. M., Sanchez, J. S., & Mollineda, R. A. (2005). A review of data complexity measures and their applicability to pattern classification problems.Actas del III Taller Nacional de Mineria de Datos y Aprendizaje.TAMIDA, 77-83.

Steinberg D. (1997), CART Users Manual, Salford Systems

Wolpert, D. H. (1992). Stacked generalization. Neural networks, 5(2), 241-259.

# CHAPTER 7

# Conclusions and Perspectives

In DA, instead of proposing new models to find a classification rule to minimize the misclassification error, a number of researchers have opted to combine models, taking advantage of the specificities of each, with a view to finding classifiers that adjust better to the data under study, thus, leading to more precise, stable and robust models.

As already observed, the dimensionality problem, often referred to by researchers as "the curse of dimensionality" (Celeux and Nakache, 1994) frequently emerges in DDA and leads to the weak performance of various models. This problem stems from the large amount of parameters that need to be estimated in the most natural models in DDA, such as the Full Multinomial Model (FMM). Furthermore, in the fields of Social or Human Sciences and Medicine, the available samples are small to moderate in size.

On the other hand, although there is an abundance of research in the area of model combination for continuous classification problems, which has subsequently led to an explosion of publications over recent years, there is still very little on discrete classification problems.

Therefore, and drawing from a study conducted by Sousa Ferreira (2000), a model combination in DDA has been proposed in this study for small or moderate sized samples. The proposed model is defined as a convex linear combination of the First-order Independence Model (FOIM) and the Dependence Trees Model (DTM), assuming independence among the explanatory variables within each class in the former, and taking into account the interactions among the explanatory variables in the latter.

The DTM model was chosen to integrate this model combination due to an understanding that while being capable of leading to predictions that are not correlated with those of the FOIM model, it can reduce the misclassification rate. In fact, the combination of different classifiers is currently a very popular field of research (Abbott, 1999; Amershi and Conati, 2009; Brito, 2002; Brito et al., 2006; Cesa-Bianchi et al., 2006; Janusz, 2010; Kotsiantis, 2011; Sousa Ferreira, 2000, 2004, 2010).

On the other hand, despite not having explored this advantage within the scope of the current study, this model defines a conditional probability function for each defined *a priori* class, providing information, per class, on the most important interactions among the explanatory variables. In the near future, an evaluation of this advantage is anticipated for an analysis of the results, in the specific case of real data.

The combination of the FOIM-DTM models, using a single coefficient $\beta$, $(0 \leq \beta \leq 1)$, leads to an intermediary model, in the case of K *a priori* classes defined between the FOIM and DTM models in discrete and not necessarily binary data. When there are more than two classes, the Hierarchical Coupling Model (Sousa Ferreira, 2000 can also be used, by transforming a problem of K classes into several problems of two classes, and by applying the FOIM-DTM combination at each level of the tree. Use of the HIERM model has enabled considerable improvement of the combined model's performance. The performance study of the proposed FOIM-DTM model was tested on a number of real datasets and later, in an attempt to ascertain its main field of application, it was also tested on simulated data. Some of the studies conducted on real and simulated data are briefly presented in the Appendix, in which the main data characteristics and the performance of the proposed model are described. With regard to the studies conducted with sets of real data, the following information may be highlighted:

- In the comparative studies conducted with the decision tree model implemented by the CART (Appendix: Cases 3, 4 and 5) algorithm, the FOIM-DTM combination systematically presents a better correct classification rate.
- In the comparative study with the combination FOIM-FMM (Appendix: Case 4), the FOIM-DTM combination presents results closer to those obtained with the FOIM-FMM combination, but are not necessarily better. The accomplishment of further comparative studies between the two FOIM-DTM and FOIM-FMM combinations is anticipated .
- In the case of problems with more than two defined *a priori* classes, when the HIERM model is applied to the FOIM-DTM combination (Appendix: Cases 2, 3, 4, 5 and 6), in addition to increasing the value of the correct classification rate, it also highlights the contribution of the combination itself.
- In the studies that were carried out (Appendix: Cases 1, 2, 3, 4, 5, 6 and 7), when compared with the single models, the FOIM-DTM combination often presents a higher performance rate, although it displays a certain degree of instability.

The first studies carried out with real datasets did not lead to a clear conclusion as to the contribution of the FOIM-DTM combination for DDA, since the results were not consistently higher than those obtained with the single models or with the previously proposed FOIM-FMM combination. This issue triggered the research conducted within the scope of the present dissertation, and led the study to assess both the importance of variable selection in DDA and the importance of understanding the relations between the data complexity factors and the models' performance. Moreover, it shed light upon the need to use simulated data in order to ascertain the main field of application of the FOIM-DTM combination.

Indeed, by constructing a model combination, the complexity of the model is increased, as even by using a simple convex linear combination, the number of parameters to be estimated is augmented. Moreover, the dimensionality problem frequently emerges, mainly due to the fact that the proposed model sets out to contribute to classification problems for small and moderate-sized samples. This knowledge served as the basis for the research conducted on the study of the variable selection methods that would enable the choice of a set of variables leading to similar or better results than the initial explanatory variable set. It was possible to verify through this study (see: Marques et al., 2013; Appendix: Cases 6 and 7), how the descriptive variable selection methods lead to an interesting choice in the run-time/correct classification rate relation.

It was also found that the use of inferential methods for selecting variables is not always possible, since the significance level $\alpha$ tends to assume rather high values. Nevertheless, whenever it is possible to apply inferential variable selection methods, they prove to be effective (Cases 6 and 7), showing that they can lead to a better performance level than with a set of all the explanatory variables or equivalent. The BON proved to be more efficient than the FDR method in terms of managing to maintain good performance of the model, by reducing the number of explanatory variables under study. Nevertheless, the FDR method displayed the ability to drastically reduce running time.

As already mentioned, concern was also shown towards ascertaining the main field of application of the FOIM-DTM combination. Therefore, more recently, assessment of the performance of this combined model was geared towards comparing datasets with different structures and comparing the performance of the model in these various structures.

At an initial stage, the study focused on sample dimension control and the intensity of the relations among variables within each class (Marques et al., 2014a; Appendix: Case 8). Secondly, an attempt was made to study other complexity factors that might influence the performance of the afore-mentioned model such as: separation between classes; balanced or unbalanced classes ; number of unobserved (missing) states; number of parameters to be estimated and sample size.

At an initial stage (Appendix: Case 8), in the case of a moderate-sized independent structure, the FOIM model is the dominant model that leads to better performance. For moderate-sized samples with a correlation structure, the DTM model displays its contribution in the combined model, whereby the FOIM-DTM combination proves to lead to better performance. For small-sized samples, of both an independent and correlation structure, the FOIM-DTM presents the best results. When the structures are independent, the FOIM model tends to obtain the best results, however for small-sized samples the FOIM-DTM combination surpasses this model. When the structure assumes some degree of correlation, the DTM model emerges naturally, however, the FOIM-DTM combination appears as an alternative to DTM for moderate-sized samples.

In Marques et al. (2014a) the study was broadened to the case of 4 classes and the results of the FOIM-DTM combination were compared with the results obtained by application of the Random Forests to the same dataset. In this study, the FOIM-DTM combination approach is shown to almost systematically surpass the performance of Random Forests.

In the second stage, the binary case was explored for two and three *a priori* defined classes. It is the non-binary case that will be analysed in greater depth in future studies.

In Marques et al. (2014b) and Appendix: Case 9, the influence of four factors in the performance of the FOIM-DTM was observed: the separation between classes measured by the affinity coefficient that varies between [0-1], the closer this coefficient is to zero, the stronger the separation; samples of balanced and unbalanced classes were assessed with both equal and different-sized samples; the amount of missing data was assessed by observation of the number of non-observed states in each class.

In relation to the accomplished work (Appendix: Case 9), the interest of the proposed method seems to stand out whenever there are just a few factors reversely influencing the performance of the model. In other words, if the classification problem has a very high level of difficulty (poorly separated classes, small-sized sample and a large number of unobserved states), owing to its simplicity, the FOIM model displays the best performance. However, if the level of difficulty is not quite so high (moderately or correctly separated classes, very small-sized sample and large amount of unobserved states) or (moderately separated classes, small-sized sample and some unobserved states) or (badly or moderately separated classes, moderate-sized sample and entirely observed states) the FOIM-DTM combination leads to the best result.

It was also possible to conclude that (Marques et al., 2014b) the advantage of the combined model when compared to single models tends to become noticeable in the case of small or very small-sized samples. It also provided an order of the complexity factors in accordance with their level of impact on the model's performance: separation between classes, ratio between the number of degrees of freedom and sample size, proportion of missing states in the minority class.

It should be noted that, in this study, other measures were sometimes used to assess the performance of the models in addition to the traditional correct classification rate. The Huberty Index, particularly important in unbalanced cases, and the $\phi$ statistic are examples of such measures, although the interest of these alternatives and their relation with the correct classification rate remain unclear, requiring continued evaluation in the near future.

To sum up, the research developed in this study enabled us to verify that the interest of the FOIM-DTM combination is effectively revealed in small or very small-sized samples and poorly separated *a priori* classes.

When the performance of the combination is analysed on the basis of existing relations among the explanatory variables within each class, the FOIM model proves to be the most suitable for independence structures and small or moderate-sized samples. For samples with related explanatory variables, the combination emerges as an alternative to the single models, especially when the available samples are small.

On the other hand, this study has also given an important contribution to DDA with two issues that have yet to be studied in greater depth in discrete data classification: variable selection and the study of model performance using non-traditional methods, such as the $\phi$ statistic and Huberty Index.

In the future, continued assessment will be conducted on the performance of the combined model in non-binary discrete data, and this assessment will be further studied in the case of $K \neq 2$ and in DDA variable selection methods.

# CHAPTER 8

# Appendices

## CASE 1[a]

***Data description:***

Data collected from Goldstein and Dillon (1978) concerning purchasing preferences for types of stores that sell audio equipment:

- $N = 412$ subjects;
- $K = 2$;
- $C_1$ - Customers for Great Department Stores with 154 subjects;
- $C_2$ - Clients Specialty Shops these devices with 258 subjects;
- $P = 4$.

The explanatory variables are four dichotomous variables (1 = yes; 0 = no):

- Variable 1: Did you search information with your relatives?
- Variable 2: Did you ask for products information?
- Variable 3: Had you previous experiences in purchasing audio equipment?
- Variable 4: Did you receive information about products from catalogs?

In this study, several combinations FOIM-DTM were trained, using values for the $\beta$ coefficient ranging from 0.05 to 0.95 with successive increments of 0.05.

***Results:***

Table 1: Percentage of correctly classified cases

| $\beta * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$ | | | | | | |
|---|---|---|---|---|---|---|
| $\beta$ | $\beta = 0$ | $\beta = 0.5$ | $0.10 \leq \beta \leq 0.35$ | $\beta = 0.4$ | $0.45 \leq \beta \leq 0.95$ | $\beta = 1$ |
| % of correctly classified cases (% CC) | 69.4% | 69.4% | 68.7% | 68.2% | **69.9%** | **69.9%** |
| % CC in $C_1$ | 70.8% | 70.8% | **72.7%** | 67.5% | 64.9% | 64.9% |
| % CC in $C_2$ | 68.6% | 68.6% | 66.3% | 68.6% | **72.9%** | **72.9%** |

[a] Comunication presented at: XV Jornadas de Classificação e Análise de Dados (JOCLAD 2008), in ESCE/IPS, 27-29 March 2008.

### CASE 2[b]

***Data description:***

Data collected from Prazeres (1996) concerning the evaluation of alexithymia ("alexithymia"means no words to express emotions). Experiments with the proposed approach regard a data set which consists of 34 dermatology's patients evaluated by the psychological test TAS-20 (Twenty Item Toronto Alexithymia Scale)- conceived to evaluate the presence of alexithymia.
For each subject, the values of six binary variables of another psychological test Rorschach test - are available.

- $N = 34$ subjects;
- $K = 3$;
- $C_1$ - Nonalexithymics Class with 7 subjects;
- $C_2$ - Alexithymics Class with 14 subjects;
- $C_3$ - Intermediate Class with 13 subjects;
- $P = 6$.

In this study, several combinations FOIM-DTM were trained, using some values for the $\beta$ coefficient.

***Results:***

Table 2: Percentage of correctly classified cases

| | | HIERM:$(1-\beta)*\hat{P}_{FOIM}+(\beta)*\hat{P}_{DTM}$ | | |
|---|---|---|---|---|
| FOIM | DTM | $\beta = 0$ | $\beta = 1$ | $\beta = 0.25; 0.50; 0.75$ |
| $64.7\%$ | $64.7\%$ | $82.5\%$ | $76.5\%$ | **85.3%** |

[b] Poster presented at: $18^{th}$ International conference on Computational Statistics (COMPSTAT 2008), in FEUP, 24-29 August 2008.

## CASE 3[c]

***Data description:***

Table 3:

| | Sample | N | Classes | Number variables |
|---|---|---|---|---|
| Real Data* | Monk | 432 | $C_1 = 216$ e $C_2 = 216$ | 6 variables, with 2 binary |
| | Balance | 623 | $C_1 = 48$; $C_2 = 288$ e $C_3 = 287$ | 4 non-binary variables |
| Simulated Data | 2 Classes | 200 | $C_1 = 130$ e $C_2 = 70$ | 4 variables, with 2 binary |
| | 4 Classes | 250 | $C_1 = 80$ e $C_2 = 70$<br>$C_3 = 30$ e $C_4 = 70$ | 3 binary variables |

- A. Asuncion and D.J. Newman, UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA:, Technical Report University of California, School of Information and Computer Science., 2007.

In this study, several combinations FOIM-DTM were trained, using some values for the $\beta$ coefficient.

[c] Comunication presented at: $11^{th}$ Conference of the International Federation of classification Societes (IFCS 2009), in Techinische Universitat Dresden - Germany, 13-18 March 2009.

***Results:***

Table 4: Percentage of correctly classified cases: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Data Base | Majority rule | Classification Method | | Resubst. | Training sample (50%) | Test sample (50%) | 2-Fold |
|---|---|---|---|---|---|---|---|
| Monks | 50.0% | CART | Comp. | 75.0% | 80.6% | 65.3% | 82.2% |
| | | | Prun. | 75.0% | 80.1% | 71.3% | - |
| | | FOIM-DTM | $\beta = 0$ | 83.3% | **100.0%** | **100.0%** | **100.0%** |
| | | | $\beta = 0.25$ | 83.3% | **100.0%** | **100.0%** | **100.0%** |
| | | | $\beta = 0.50$ | **100.0%** | **100.0%** | **100.0%** | **100.0%** |
| | | | $\beta = 0.75$ | **100.0%** | 89.4% | 81.5% | 83.6% |
| | | | $\beta = 1$ | 75.0% | 76.9% | 72.7% | 74.1% |
| Balance | 46.2% | CART | Comp. | 84.9% | 85.6% | 74.9% | 74.6% |
| | | | Prun. | 83.8% | 84.0% | 74.6% | - |
| | | FOIM-DTM | $\beta = 0$ | 82.0% | 84.0% | 76.9% | 74.8% |
| | | | $\beta = 0.25$ | 84.0% | 86.2% | 79.1% | 77.3% |
| | | | $\beta = 0.50$ | 87.3% | 88.8% | 83.0% | 80.7% |
| | | | $\beta = 0.75$ | 90.4% | **91.3%** | 86.5% | 86.5% |
| | | | $\beta = 1$ | **92.3%** | **91.3%** | **89.7%** | **89.8%** |

Table 5: Percentage of correctly classified cases: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Data Base | Majority rule | Classification Method | | Resubst. | Training sample (50%) | Test sample (50%) | 2-Fold |
|---|---|---|---|---|---|---|---|
| 2 Simulated | 65.0% | CART | Comp. | **70.0%** | **71.0%** | 63.0% | 60.0% |
| | | | Prun. | 69.5% | 70.0% | 64.0% | - |
| | | FOIM-DTM | $\beta = 0$ | 67.0% | 67.0% | 57.0% | 54.0% |
| | | | $\beta = 0.25$ | 66.0% | 68.0% | 58.0% | 54.0% |
| | | | $\beta = 0.50$ | 66.0% | 70.0% | 60.0% | 57.0% |
| | | | $\beta = 0.75$ | 66.5% | 68.0% | **67.0%** | **64.0%** |
| | | | $\beta = 1$ | 66.0% | 70.0% | 64.0% | 62.0% |
| 4 Simulated | 32.0% | CART | Comp. | 56.4% | 61.6% | 45.6% | 54.8% |
| | | | Prun. | 56.0% | 59.2% | 52.8% | - |
| | | FOIM-DTM | $\beta = 0$ | 53.6% | 71.2% | 47.2% | 49.2% |
| | | | $\beta = 0.25$ | 53.6% | 72.0% | 52.8% | 52.8% |
| | | | $\beta = 0.50$ | 53.6% | **74.4%** | 51.2% | 52.0% |
| | | | $\beta = 0.75$ | 53.6% | 72.8% | 51.2% | 52.0% |
| | | | $\beta = 1$ | 56.0% | 72.8% | 51.2% | 55.2% |
| | | HIERM | $\beta = 0$ | 61.6% | 71.2% | 59.2% | 60.8% |
| | | | $\beta = 0.25$ | 61.6% | 72.0% | **60.8%** | 61.6% |
| | | FOIM-DTM | $\beta = 0.50$ | 61.6% | **74.4%** | **60.8%** | 62.4% |
| | | | $\beta = 0.75$ | 64.0% | 72.8% | 59.2% | 61.6% |
| | | | $\beta = 1$ | **64.8%** | 72.8% | 57.6% | **64.8%** |

## CASE 4[d]

### *Data description:*

The GSS1 - Gudjonsson Suggestibility Scale (Gudjonsson, 1997) was developed to assess the tendency that some people have to distort facts when interviewed. The experiment consists of:

- It's orally presented a story about a robbery, followed by one task of immediate recall and one delayed recall task (with a range of about 50 minutes;
- At the end of the deferred memory tasks, each subject answers to 20 questions, 15 of which are constructed so as to induce the subject in error. At the end of the 20 questions, subject is told that he had made some mistakes (even if he didn't committed any one) and therefore, he answers again to the 20 questions, trying to be this time be more precise;
- Answers to the 20 questions are listed as amendment or transfer depending if the answer changes from 1st to 2nd time or if he is influenced by the issues created for misleading.

In this work is considered a general hypothesis that exist individual differences on vulnerability to suggestion. In particular, were analyzed the demographic characteristics and its association with vulnerability to suggestion. Classification task considered, in particular, the classes offered by demographic variables like gender, age group and educational level of individuals and suggestibility measured using binary variables (Pires, 2010). In this study, several combinations FOIM-DTM were trained, using some values for the $\beta$ coefficient.

[d] Comunication presented at: XVI Jornadas de Classificação e Análise de Dados (JOCLAD 2009), in Universidade do Algarve, 2-4 April 2009.

Table 6: Data Characteristics

| Class | N |
|---|---|
| Gender | $C_1 = 30$<br>$C_2 = 68$ |
| Age group | $C_1 = 72$<br>$C_2 = 7$<br>$C_3 = 19$ |
| Schooling level | $C_1 = 17$<br>$C_2 = 46$<br>$C_3 = 22$<br>$C_4 = 13$ |

***Results:***

Table 7: Percentage of correctly classified cases for gender classes

| Majority rule | Classification Method | | Resubst. | Training sample (50%) | Test sample (50%) |
|---|---|---|---|---|---|
| 69.4% | CART | Comp. | 79.6% | 87.8% | 65.3% |
| | | Prun. | 77.6% | 85.7% | 67.7% |
| | FOIM-DTM | $\beta = 0$ (DTM) | 71.4% | 73.5% | 63.3% |
| | | $\beta = 0.25$ | 71.4% | 73.5% | 65.3% |
| | | $\beta = 0.50$ | 73.5% | 79.6% | 65.3% |
| | | $\beta = 0.75$ | 75.5% | 77.6% | **73.5%** |
| | | $\beta = 1$ (FOIM) | 75.5% | 81.6% | 69.4% |
| | FOIM-FMM* | $\beta = 0$ (FOIM) | **83.7%** | **89.8%** | 46.9% |
| | | $\beta_1 = 0.301$ | | | **71.4%** |
| | | $\beta_2 = 0.126$ | | | **72.7%** |
| | | $\beta_3 = 0.147$ | | | **89.8%** |
| | | $\beta = 1$ (FMM) | 75.5% | 81.6% | 69.4% |

* (Sousa Ferreira, 2000)

Table 8: Percentage of correctly classified cases for age group classes

| Majority rule | Classification Method | | Resubst. | Training sample (50%) | Test sample (50%) |
|---|---|---|---|---|---|
| 73.5% | CART | Comp. | 85.7% | 85.7% | 73.5% |
| | | Prun. | 85.7% | 85.7% | 69.4% |
| | FOIM-DTM | $\beta = 0$ (DTM) | 77.5% | 69.4% | 67.3% |
| | | $\beta = 0.25$ | 79.6% | 69.4% | 67.3% |
| | | $\beta = 0.50$ | 78.6% | 75.5% | 67.3% |
| | | $\beta = 0.75$ | 77.5% | 85.7% | 75.5% |
| | | $\beta = 1$ (FOIM) | 75.5% | 81.6% | 77.5% |
| | HIERM FOIM-DTM | $\beta = 0$ (DTM) | 77.6% | 79.6% | **77.6%** |
| | | $\beta = 0.25$ | 79.6% | 79.6% | 75.5% |
| | | $\beta = 0.50$ | 78.6% | 79.6% | 71.4% |
| | | $\beta = 0.75$ | 77.6% | 83.7% | **77.6%** |
| | | $\beta = 1$ (FOIM) | 76.5% | 85.7% | **77.6%** |
| | FOIM-FMM* | $\beta = 0$ | **92.9%** | **93.9%** | 65.3% |
| | | $\beta_1 = 0.301$ | | | **77.6%** |
| | | $\beta_2 = 0.126$ | | | **77.6%** |
| | | $\beta_3 = 0.147$ | | | 75.5% |
| | | $\beta = 1$ | 76.5% | 83.7% | 75.5% |

* (Sousa Ferreira, 2000)

Table 9: Percentage of correctly classified cases for Schooling level classes

| Majority rule | Classification Method | | Resubst. | Training sample (50%) | Test sample (50%) |
|---|---|---|---|---|---|
| 46.9% | CART | Comp. | 68.4% | 75.5% | 40.8% |
| | | Prun. | 68.4% | 75.5% | 40.8% |
| | FOIM-DTM | $\beta = 0$ (DTM) | 50.0% | 63.3% | 30.6% |
| | | $\beta = 0.25$ | 62.2% | 69.4% | 34.7% |
| | | $\beta = 0.50$ | 58.2% | 73.5% | 30.6% |
| | | $\beta = 0.75$ | 59.2% | 67.3% | 44.9% |
| | | $\beta = 1$(FOIM) | 56.1% | 65.3% | 44.9% |
| | HIERM<br>FOIM-DTM | $\beta = 0$ (DTM) | 63.3% | 71.4% | 49.0% |
| | | $\beta = 0.25$ | 72.4% | 75.5% | 49.0% |
| | | $\beta = 0.50$ | 69.4% | 73.5% | 51.0% |
| | | $\beta = 0.75$ | 65.3% | 73.5% | **55.1%** |
| | | $\beta = 1$ (FOIM) | 62.2% | 73.5% | **55.1%** |
| | FOIM-FMM* | $\beta = 0$ | 53.1% | **83.7%** | 34.7% |
| | | $\beta_1 = 0.4999/0.9921/0.8339$ | | | 49.0% |
| | | $\beta = 0.6139/0.6492/0.6968$ | | | 49.0% |
| | | $\beta = 0.2492/0.1204/0.5454$ | | | 53.1% |
| | | $\beta = 1$ | **76.5%** | 63.3% | 40.8% |

* (Sousa Ferreira, 2000)

**CASE 5**[e]

***Data description:***
This work refers to the study of data on clients of Belém Cultural Centre (CCB) and their evaluation of the quality of products/services in CCB (Inquiry realized in 2007, by. Ana Duarte (2009) to whom we thanks the availability of the data set.

Table 10: Data Characteristics

| Class | Schooling | n | % |
|---|---|---|---|
| 1 | Secondary grade | 177 | 17.9 |
| 2 | University frequency | 136 | 13.8 |
| 3 | Graduation | 462 | 46.8 |
| 4 | Master or Phd | 213 | 21.6 |
| | Total | 988 | |

In this study, several combinations FOIM-DTM were trained, using some values for the $\beta$ coefficient.

***Results:***

Table 11: Percentage of correctly classified cases

| Classification Method | | Test sample (35%) |
|---|---|---|
| CART | | 46.1% |
| FOIM-DTM | $\beta = 0$ (DTM) | 45.0% |
| | $\beta = 0.10$ | 44.4% |
| | $\beta = 0.20$ | 45.8% |
| | $\beta = 0.30$ | 46.4% |
| | $\beta = 0.40$ | 46.4% |
| | $\beta = 0.50$ | 47.6% |
| | $\beta = 0.60$ | 47.3% |
| | $\beta = 0.70$ | **47.8%** |
| | $\beta = 0.80$ | **47.8%** |
| | $\beta = 0.90$ | 47.0% |
| | $\beta = 1$ (FOIM) | 47.0% |

[e] Comunication presented at: XVII Jornadas de Classificação e Análise de Dados (JOCLAD 2010), in ISCTE, 25-27 March 2010.

Table 12: Percentage of correctly classified cases by HIERM model

| Classification<br>Method | | Test sample<br>(35%) |
|---|---|---|
| CART | | 46.1% |
| HIERM<br>FOIM-DTM | $\beta = 0$ (DTM) | 47.8% |
| | $\beta = 0.10$ | 47.7% |
| | $\beta = 0.20$ | 48.1% |
| | $\beta = 0.30$ | 48.1% |
| | $\beta = 0.40$ | 49.3% |
| | $\beta = 0.50$ | 49.3% |
| | $\beta = 0.60$ | 49.3% |
| | $\beta = 0.70$ | 48.7% |
| | $\beta = 0.80$ | 48.4% |
| | $\beta = 0.90$ | **49.9%** |
| | $\beta = 1$ (FOIM) | **49.9%** |

**CASE 6**[f]

These work was focused in feature selection, comparing some criterias as: Chi-Square statistic ($Q^2$), Mutual Information ($I$), Bonferroni Correction (BON) and the False Discovery Rate (FDR).

***Data description:***
**GSS1** - The Gudjonsson Suggestibility Scale (GSS1) (Gudjonsson, 1997) was developed to evaluate the trend in forensics, that some people have to distorting facts when interviewed (Pires, 2010).

**MVS** - The psychological test My Vocational Situation (MVS) (Lima, 1998) is organized into two scales: Occupational Information and Barriers (Difficulties). In this case, the aim is to study the relationships between features personality and career concerns.

Table 13: Data Characteristics

| Sample | Class | | Nº of variables |
|---|---|---|---|
| GSS1 (n=98) | Gender | $C_1 = 30$ (M)<br>$C_2 = 68$ (F) | 10 |
| MVS (n=1203) | Course | $C_1 = 480$ (Biol+Psic.)<br>$C_2 = 297$ (Letras)<br>$C_3 = 426$ (Eng.) | 8 |
| | Gender | $C_1 = 560$ (M.)<br>$C_2 = 643$ (F) | 8 |

In this study, several combinations FOIM-DTM were trained, using some values for the $\beta$ coefficient.

[f] Poster presented at: XVIII Congresso Anual da Sociedade Portuguesa de Estatística (SPE 2010), in S. Pedro do Sul, 29 September - 2 October 2010.

***Results:***

Table 14: Percentage of correctly classified cases: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Data Base | Majority Rule | $\beta$ | Training Sample | | | | 2-fold | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 10 Var. | QQ<br>IM<br>5 Var. | QQ<br>IM<br>3 Var. | FDR<br>BON | 10 Var. | QQ<br>IM<br>5 Var. | QQ<br>IM<br>3 Var. | FDR<br>BON |
| GSS1<br>(Gender) | 69.4% | $\beta = 0$ | **77.6%** | 53.1% | 57.1% | - | 54.1% | 57.1% | 49.0% | - |
| | | $\beta = 0.10$ | **77.6%** | 53.1% | 57.1% | - | 62.3% | **58.2%** | 51.0% | - |
| | | $\beta = 0.20$ | **77.6%** | 53.1% | 57.1% | - | 62.3% | **58.2%** | 51.0% | - |
| | | $\beta = 0.30$ | **77.6%** | 40.8% | 57.1% | - | 60.2% | 49.9% | 51.0% | - |
| | | $\beta = 0.40$ | 73.5% | 53.1% | 57.1% | - | 58.2% | 46.9% | 54.1% | - |
| | | $\beta = 0.50$ | 73.5% | 59.2% | 57.1% | - | 60.2% | 50.0% | 54.1% | - |
| | | $\beta = 0.60$ | 71.4% | 59.2% | 57.1% | - | **63.3%** | 50.0% | 54.1% | - |
| | | $\beta = 0.70$ | 67.4% | **67.4%** | 57.1% | - | **63.3%** | 53.1% | 54.1% | - |
| | | $\beta = 0.80$ | 67.4% | **67.4%** | 57.1% | - | **63.3%** | 53.1% | 54.1% | - |
| | | $\beta = 0.90$ | 67.4% | **67.4%** | **61.2%** | - | **63.3%** | 53.1% | **62.2%** | - |
| | | $\beta = 1$ | 63.3% | **67.4%** | **61.2%** | - | **63.3%** | 53.6% | **62.2%** | - |

Table 15: Percentage of correctly classified cases: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Data Base | Majority Rule | $\beta$ | Test Sample(30%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | QQ | IM | QQ | IM | FDR e BON |
| | | | 8 Var. | 5 Var. | 5 Var. | 3 Var. | 3 Var. | 1 Var. |
| MVS (Gender) | 53.5% | $\beta = 0$ | 52.6% | 49.0% | 49.3% | 44.6% | 49.0% | - |
| | | $\beta = 0.10$ | 54.1% | 49.0% | 50.7% | 44.6% | **50.1%** | - |
| | | $\beta = 0.20$ | 53.7% | 49.0% | 50.7% | 44.6% | **50.1%** | - |
| | | $\beta = 0.30$ | 53.7% | 49.9% | 50.7% | 44.6% | **50.1%** | - |
| | | $\beta = 0.40$ | 54.0% | 51.8% | 50.7% | 44.6% | **50.1%** | - |
| | | $\beta = 0.50$ | 53.2% | 52.1% | 50.7% | 44.6% | **50.1%** | - |
| | | $\beta = 0.60$ | 53.5% | 52.1% | 50.7% | 44.6% | **50.1%** | - |
| | | $\beta = 0.70$ | 54.9% | 52.1% | 54.3% | 44.6% | **50.1%** | - |
| | | $\beta = 0.80$ | 53.7% | 52.1% | 54.9% | **52.1%** | **50.1%** | - |
| | | $\beta = 0.90$ | 53.7% | **53.5%** | 54.9% | **52.1%** | **50.1%** | - |
| | | $\beta = 1$ | **55.1%** | **53.5%** | **56.0%** | **52.1%** | **50.1%** | - |

Table 16: Percentage of correctly classified cases: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Data Base | Majority | $\beta$ | Training Sample | | | | 2-fold | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | QQ | QQ | FDR | | QQ | QQ | FDR |
| | | | 8 Var. | IM | IM | BON | 8 Var. | IM | IM | BON |
| | Rule | | | 5 Var. | 3 Var. | 6 Var. | | 5 Var. | 3 Var. | 6 Var. |
| MVS<br>(Course) | 39.9% | $\beta = 0$ | 45.8% | 46.6% | **45.9%** | 44.2% | 54.1% | 57.1% | **43.9%** | 42.5% |
| | | $\beta = 0.10$ | 45.7% | 46.6% | **45.9%** | 46.1% | 48.6% | 44.8% | **43.9%** | 43.9% |
| | | $\beta = 0.20$ | 47.1% | 46.7% | **45.9%** | 46.0% | 48.9% | 44.8% | **43.9%** | 44.2% |
| | | $\beta = 0.30$ | **47.2%** | 46.7% | **45.9%** | 46.0% | 49.2% | 44.8% | **43.9%** | 44.2% |
| | | $\beta = 0.40$ | 46.3% | 46.7% | 45.9% | **45.9%** | 49.2% | 45.0% | **43.9%** | 43.9% |
| | | $\beta = 0.50$ | 46.4% | **46.9%** | **45.9%** | 46.5% | 49.2% | 45.8% | **43.9%** | 45.9% |
| | | $\beta = 0.60$ | 45.8% | **46.9%** | **45.9%** | 46.7% | 50.6% | 44.8% | **43.9%** | 46.1% |
| | | $\beta = 0.70$ | 45.8% | **46.9%** | **45.9%** | 46.9% | **51.1%** | 44.8% | **43.9%** | **46.7%** |
| | | $\beta = 0.80$ | 45.5% | **46.9%** | **45.9%** | **47.0%** | **51.1%** | 44.8% | **43.9%** | 46.1% |
| | | $\beta = 0.90$ | 44.8% | **46.9%** | **45.9%** | 46.6% | 50.0% | 44.8% | **43.9%** | 46.4% |
| | | $\beta = 1$ | 44.6% | **46.9%** | **45.9%** | 45.9% | 50.0% | 44.8% | **43.9%** | 45.9% |

Table 17: Percentage of correctly classified cases by HIERM model and $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Data Base | Majority Rule | $\beta$ | Training Sample | | | | 2-fold | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 8 Var. | QQ IM 5 Var. | QQ IM 3 Var. | FDR BON 6 Var. | 8 Var. | QQ IM 5 Var. | QQ IM 3 Var. | FDR BON 6 Var. |
| MVS (Course) | 39.9% | $\beta = 0$ | 56.2% | 56.96% | 56% | 60.4% | 55.52% | 55.8% | 50.8% | 60.2% |
| | | $\beta = 0.10$ | 56.4% | 57.07% | 56.24% | 60.76% | 55.8% | 55.8% | 51.1% | 60.2% |
| | | $\beta = 0.20$ | 56.8% | 57.07% | 56.24% | 60.88% | 56.08% | 55.8% | 51.1% | 60.5% |
| | | $\beta = 0.30$ | 58.0% | 57.07% | 56.24% | 61.12% | 58.01% | 55.8% | 51.1% | 60.8% |
| | | $\beta = 0.40$ | 58.0% | 57.07% | 56.24% | 61.12% | 58.01% | 55.8% | 51.1% | 61.1% |
| | | $\beta = 0.50$ | 58.2% | 57.07% | 56.24% | **61.47%** | 58.01% | 55.8% | 51.1% | 61.1% |
| | | $\beta = 0.60$ | 58.6% | 58.74% | 56.24% | **61.47%** | 60.77% | 57.73% | 51.1% | **61.3%** |
| | | $\beta = 0.70$ | 58.7% | **61.4%** | 56.24% | 61.12% | 60.77% | **61.6%** | 51.1% | **61.3%** |
| | | $\beta = 0.80$ | **60.6%** | 61% | **58.5%** | 60.8% | **63.5%** | 60.77% | 55% | 60.8% |
| | | $\beta = 0.90$ | 60.2% | 60.52% | 58.38% | 60.4% | 62.98% | 60.5% | **55.3%** | 61.1% |
| | | $\beta = 1$ | 60.3% | 60.52% | 58.38% | 60.4% | **63.5%** | 60.22% | **55.3%** | 61.1% |

## CASE 7[g]

***Data description:***
**NINA** - This data set which consists of 34 dermatology's patients evaluated by a psychological test set. The whole sample is divided into two classes - Unhealthy ($C_1$) and Healthy ($C_2$). In this data set we considered eleven binary variables.(Prazeres, 1996)

**GSS1** - This data set refers to measurements of susceptibility to changes the testimony of 98 individuals. The target classes are related to gender - Men ($C_1$) and Women ($C_2$). In this data set we considered eight binary variables.(Pires, 2010)

Table 18: Data Characteristics

| Sample | Class | | Nº of variables |
|---|---|---|---|
| NINA (n=34) | Healthy | $C_1 = 14$<br>$C_2 = 20$ | 11 |
| GSS1 (n=98) | Gender | $C_1 = 30$<br>$C_2 = 68$ | 8 |

In this study, several combinations FOIM-DTM were trained, using some values for the $\beta$ coefficient.

[g] Poster presented at: $14^{th}$ Applied Stochastic Models and Data Analysis Conference (ASMDA 2011) in Roma, 7-10 June 2011.

***Results:***

Table 19: Percentage of correctly classified cases for Nina Data by 2-fold: $(\beta) * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$

| Methods of Selection | nº selected | $\beta = 0$ | $\beta = 0.25$ | $\beta = 0.50$ | $\beta = 0.75$ | $\beta = 1$ | Runtime |
|---|---|---|---|---|---|---|---|
| All variables | 11 | 67.7% | 79.4% | 79.4% | 79.4% | 76.5% | 1.72 days |
| Core set | 4 | 61.8% | 70.6% | 76.5% | 76.5% | 76.5% | 1.2 minutes |
| FDR ($\alpha = 0.2$) | 2 | (DTM approach is not possible since Mutual Information is null) | | | | | |
| FDR ($\alpha = 0.25$) | 3 | 76.5% | 76.5% | 73.5% | 73.5% | 73.5% | 39.3 second |
| BON($\alpha = 0.30$) | 7 | 70.6% | 73.5% | **79.4%** | **79.4%** | **79.4%** | 17.8 minutes |
| Cramer Statistic | 6 | 70.6% | 76.5% | **79.4%** | **79.4%** | **79.4%** | 6.7 minutes |
| Mutual Inf. | 6 | (similar to Cramer Statistic based selection results) | | | | | |

Table 20: Percentage of correctly classified cases for GSS1 Data by 2-fold:: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Methods of Selection | n$^o$ selected | $\beta = 0$ | $\beta = 0.25$ | $\beta = 0.50$ | $\beta = 0.75$ | $\beta = 1$ | Runtime |
|---|---|---|---|---|---|---|---|
| Core set | 8 | 58.2% | 63.3% | 62.2% | 59.2% | 58.2% | 57.4 minutes |
| FDR ($\alpha = 0.6$) | 2 | 55.1% | 55.1% | 55.1% | 55.1% | 55.1% | 22.8 minutes |
| BON | - | - | - | - | - | - | |
| Cramer Statistic | 3 | 59.2% | 59.2% | 61.2% | 61.2% | 62.2% | 44.3 second |
| Mutual Inf. | 2 | **62.3%** | **62.3%** | **62.3%** | **62.3%** | **62.3%** | 30.4 second |

## CASE 8[h]

***Data description:***

To evaluate the performance of the proposed model, were considered two defined classes *a priori* and resort to simulation of binary data based on the model Bahadur proposed by Goldstein and Dillon (1978) and Celeux and Mkhadri (1992). Based on this model 10 simulations were performed considering two types of structures with P = 6 binary variables. The parameters $\theta_k$ considered in the simulation of the Bernoulli variables were:

$\theta_1 = (0.6, 0.4, 0.6, 0.5, 0.5, 0.6)$ and $\theta_2 = (0.5, 0.3, 0.5, 0.4, 0.4, 0.5)$.

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 $p \neq g,\ p,g = 1, ..., 6;$.For each of the structures are considered very small samples (30 observations in each class) and samples of small size (60 observations in each class). The *a priori* probabilities were considered equal.

In this study, several combinations FOIM-DTM were trained, using values for the $\beta$ coefficient.

[h] Poster presented at: XIX Jornadas de Classificação e Análise de Dados (JOCLAD 2012) in Instituto Politécnico de Tomar, 28-31 March 2012.
and
Poster presented at: International Conference on Trends and Perspectives in Linear Statistical Inference (LINSTAT 2010), in Tomar, 27-31 July 2010.

***Results:***

Table 21: Percentage of correctly classified cases for IND structure: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| $\beta$ | Very small sized ($n_1 = n_2 = 30$) | | | | Small sized ($n_1 = n_2 = 60$) | | | |
|---|---|---|---|---|---|---|---|---|
| | Training sample | | 2-fold | | Training sample | | 2-fold | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| $\beta = 0$ | **0.797** | 0.071 | 0.472 | 0.059 | 0.672 | 0.059 | 0.463 | 0.032 |
| $\beta = 0.10$ | 0.793 | 0.066 | 0.485 | 0.043 | 0.673 | 0.059 | 0.464 | 0.034 |
| $\beta = 0.20$ | 0.790 | 0.063 | 0.485 | 0.044 | 0.678 | 0.060 | 0.463 | 0.035 |
| $\beta = 0.30$ | 0.787 | 0.065 | 0.490 | 0.040 | **0.683** | 0.056 | 0.459 | 0.026 |
| $\beta = 0.40$ | 0.777 | 0.067 | 0.495 | 0.033 | 0.677 | 0.059 | 0.464 | 0.033 |
| $\beta = 0.50$ | 0.787 | 0.053 | 0.498 | 0.035 | 0.675 | 0.059 | 0.466 | 0.034 |
| $\beta = 0.60$ | 0.783 | 0.059 | **0.504** | 0.038 | 0.678 | 0.063 | 0.461 | 0.032 |
| $\beta = 0.70$ | 0.787 | 0.067 | 0.500 | 0.044 | 0.670 | 0.067 | 0.464 | 0.047 |
| $\beta = 0.80$ | 0.760 | 0.068 | 0.488 | 0.063 | 0.648 | 0.059 | 0.470 | 0.050 |
| $\beta = 0.90$ | 0.747 | 0.045 | 0.490 | 0.079 | 0.643 | 0.054 | 0.468 | 0.046 |
| $\beta = 1$ | 0.703 | 0.033 | 0.495 | 0.082 | 0.618 | 0.051 | **0.472** | 0.052 |

Table 22: Percentage of correctly classified cases for DIF structure: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| $\beta$ | Very small sized ($n_1 = n_2 = 30$) | | | | Small sized ($n_1 = n_2 = 60$) | | | |
|---|---|---|---|---|---|---|---|---|
| | Training sample | | 2-fold | | Training sample | | 2-fold | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| $\beta = 0$ | **0.692** | 0.080 | 0.484 | 0.084 | **0.633** | 0.044 | 0.514 | 0.059 |
| $\beta = 0.10$ | 0.692 | 0.080 | 0.487 | 0.086 | **0.633** | 0.044 | 0.510 | 0.069 |
| $\beta = 0.20$ | 0.692 | 0.080 | 0.480 | 0.085 | 0.630 | 0.048 | 0.515 | 0.072 |
| $\beta = 0.30$ | 0.689 | 0.080 | 0.483 | 0.084 | 0.625 | 0.049 | 0.516 | 0.068 |
| $\beta = 0.40$ | 0.689 | 0.080 | 0.483 | 0.084 | 0.622 | 0.043 | **0.517** | 0.066 |
| $\beta = 0.50$ | 0.689 | 0.080 | 0.483 | 0.084 | 0.618 | 0.040 | **0.517** | 0.066 |
| $\beta = 0.60$ | 0.666 | 0.058 | 0.477 | 0.079 | 0.615 | 0.036 | 0.515 | 0.069 |
| $\beta = 0.70$ | 0.635 | 0.040 | **0.503** | 0.087 | 0.610 | 0.033 | 0.512 | 0.070 |
| $\beta = 0.80$ | 0.610 | 0.045 | 0.500 | 0.089 | 0.595 | 0.036 | 0.508 | 0.065 |
| $\beta = 0.90$ | 0.597 | 0.043 | 0.497 | 0.091 | 0.588 | 0.038 | 0.512 | 0.063 |
| $\beta = 1$ | 0.583 | 0.045 | 0.502 | 0.094 | 0.580 | 0.036 | 0.509 | 0.061 |

## CASE 9[i]

***Data description:***
In this numerical experiments for simulated data using the R function rmultinom (n, size, prob) that generates vectors according to multinomial distribution, where:

- n - number of random vectors to draw;
- size - integer, say N, specifying the total number of objects that are put into K boxes in the typical multinomial experiment;
- prob - numeric non-negative vector of length K, specifying the probability for the K classes; is internally normalized to sum 1.

In this work, we consider the bi-class case and four types of population structures, using very small ($n = 60$), small ($n = 120$) and moderate ($n = 400$) samples sizes, with $P = 4$ binary variables:

*A Structure:*
$C_1$ - $prob = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$
$C_2$ - $prob = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$

*B Structure:*
$C_1$ - $prob = (0.4, 0.6, 0.6, 0.4, 0.4, 0.6, 0.6, 0.4)$
$C_2$ - $prob = (0.70.3, 0.3, 0.7, 0.7, 0.3, 0.3, 0.7)$

*C Structure:*
$C_1$ - $prob = (0.4, 0.6, 0.6, 0.4, 0.4, 0.6, 0.6, 0.4)$
$C_2$ - $prob = (0.9, 0.1, 0.1, 0.9, 0.9, 0.1, 0.1, 0.9)$

*D Structure:*
$C_1$ - $prob = (0.1, 0.9, 0.7, 0.3, 0.2, 0.80.6, 0.4)$
$C_2$ - $prob = (0.9, 0.1, 0.3, 0.7, 0.8, 0.2, 0.1, 0.9)$

Twenty random samples are generated for each structure. *Prior* probabilities are considered equal.

[i] Poster presented at: $7^{th}$ Workshop on Statistics, Mathematics and Computation, in Instituto Politécnico de Tomar, 28-29 May 2013.

***Results:***

Table 23: Percentage of correctly classified cases for A structure: $(\beta) * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$

| $\beta$ | very small sized | | small sized | | moderate samples sized | |
|---|---|---|---|---|---|---|
| | $n_1 = n_2 = 30$ | | $n_1 = n_2 = 60$ | | $n_1 = n_2 = 200$ | |
| | mean | s.d. | mean | s.d. | mean | s.d. |
| $\beta = 0$ | 48.70% | 0.088 | **52.10%** | 0.034 | 50.10% | 0.030 |
| $\beta = 0.10$ | 49.00% | 0.094 | 52.00% | 0.035 | 50.20% | 0.030 |
| $\beta = 0.20$ | 49.60% | 0.095 | 51.70% | 0.038 | 50.00% | 0.029 |
| $\beta = 0.30$ | 49.90% | 0.083 | 51.40% | 0.039 | 50.20% | 0.032 |
| $\beta = 0.40$ | 49.70% | 0.086 | 51.10% | 0.041 | 50.20% | 0.030 |
| $\beta = 0.50$ | 50.40% | 0.092 | 51.50% | 0.043 | 50.40% | 0.030 |
| $\beta = 0.60$ | 51.70% | 0.090 | 51.40% | 0.045 | **50.50%** | 0.030 |
| $\beta = 0.70$ | 52.40% | 0.083 | 51.60% | 0.051 | 50.30% | 0.032 |
| $\beta = 0.80$ | 52.30% | 0.080 | 50.70% | 0.055 | 50.30% | 0.030 |
| $\beta = 0.90$ | 52.80% | 0.077 | 50.10% | 0.052 | 50.30% | 0.029 |
| $\beta = 1$ | **53.60%** | 0.078 | 49.80% | 0.052 | 50.00% | 0.028 |
| $aff$ | 0.791 | 0.047 | 0.921 | 0.030 | 0.981 | 0.007 |

Table 24: Percentage of correctly classified cases for B structure: $(\beta) * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$

| $\beta$ | very small sized | | small sized | | moderate samples sized | |
|---|---|---|---|---|---|---|
| | $n_1 = n_2 = 30$ | | $n_1 = n_2 = 60$ | | $n_1 = n_2 = 200$ | |
| | mean | s.d. | mean | s.d. | mean | s.d. |
| $\beta = 0$ | 65.60% | 0.069 | 68.10% | 0.059 | 70.70% | 0.025 |
| $\beta = 0.10$ | 65.70% | 0.077 | 69.10% | 0.055 | 71.40% | 0.028 |
| $\beta = 0.20$ | 66.20% | 0.078 | 69.30% | 0.054 | 72.00% | 0.025 |
| $\beta = 0.30$ | 66.10% | 0.076 | 69.50% | 0.052 | 72.30% | 0.025 |
| $\beta = 0.40$ | 66.80% | 0.080 | 70.30% | 0.053 | 72.80% | 0.026 |
| $\beta = 0.50$ | 66.40% | 0.081 | 70.40% | 0.054 | 73.20% | 0.026 |
| $\beta = 0.60$ | 67.10% | 0.088 | 70.30% | 0.051 | 73.30% | 0.026 |
| $\beta = 0.70$ | 68.80% | 0.086 | 70.70% | 0.050 | 73.20% | 0.026 |
| $\beta = 0.80$ | 69.50% | 0.083 | 70.80% | 0.049 | 73.30% | 0.027 |
| $\beta = 0.90$ | 69.30% | 0.090 | **71.10%** | 0.051 | **73.60%** | 0.025 |
| $\beta = 1$ | **69.80%** | 0.087 | 71.00% | 0.048 | **73.60%** | 0.024 |
| $aff$ | 0.571 | 0.136 | 0.716 | 0.060 | 0.786 | 0.031 |

Table 25: Percentage of correctly classified cases for C structure: $(\beta) * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$

| $\beta$ | very small sized | | small sized | | moderate samples sized | |
|---|---|---|---|---|---|---|
| | $n_1 = n_2 = 30$ | | $n_1 = n_2 = 60$ | | $n_1 = n_2 = 200$ | |
| | mean | s.d. | mean | s.d. | mean | s.d. |
| $\beta = 0$ | 80.00% | 0.068 | 84.00% | 0.043 | 84.90% | 0.028 |
| $\beta = 0.10$ | 80.60% | 0.067 | 84.30% | 0.045 | 85.20% | 0.028 |
| $\beta = 0.20$ | 81.20% | 0.068 | 84.50% | 0.045 | 85.80% | 0.029 |
| $\beta = 0.30$ | 82.30% | 0.060 | 85.20% | 0.042 | 86.20% | 0.031 |
| $\beta = 0.40$ | 83.60% | 0.058 | 86.10% | 0.039 | 87.10% | 0.027 |
| $\beta = 0.50$ | 83.40% | 0.056 | 86.10% | 0.034 | 87.40% | 0.025 |
| $\beta = 0.60$ | 85.00% | 0.056 | 86.20% | 0.037 | 87.90% | 0.023 |
| $\beta = 0.70$ | 85.30% | 0.053 | 86.70% | 0.034 | 88.30% | 0.021 |
| $\beta = 0.80$ | **85.60%** | 0.043 | 86.80% | 0.032 | 88.40% | 0.021 |
| $\beta = 0.90$ | 85.20% | 0.038 | **86.90%** | 0.032 | **88.50%** | 0.021 |
| $\beta = 1$ | 85.20% | 0.039 | **86.90%** | 0.030 | 88.40% | 0.020 |
| $aff$ | 0.321 | 0.112 | 0.383 | 0.082 | 0.451 | 0.044 |

Table 26: Percentage of correctly classified cases for D structure: $(\beta) * \hat{P}_{FOIM} + (1-\beta) * \hat{P}_{DTM}$

| $\beta$ | very small sized |  | small sized |  | moderate samples sized |  |
|---|---|---|---|---|---|---|
|  | $n_1 = n_2 = 30$ |  | $n_1 = n_2 = 60$ |  | $n_1 = n_2 = 200$ |  |
|  | mean | s.d. | mean | s.d. | mean | s.d. |
| $\beta = 0$ | 85.20% | 0.043 | 88.10% | 0.050 | 90.50% | 0.024 |
| $\beta = 0.10$ | 87.10% | 0.051 | 88.60% | 0.046 | 90.80% | 0.024 |
| $\beta = 0.20$ | 88.10% | 0.054 | 89.40% | 0.041 | 91.20% | 0.024 |
| $\beta = 0.30$ | 89.00% | 0.049 | 89.40% | 0.042 | 91.40% | 0.023 |
| $\beta = 0.40$ | 90.40% | 0.046 | 90.60% | 0.034 | 91.80% | 0.015 |
| $\beta = 0.50$ | 90.40% | 0.042 | 91.10% | 0.033 | 92.00% | 0.016 |
| $\beta = 0.60$ | 90.50% | 0.033 | 91.40% | 0.033 | 92.30% | 0.015 |
| $\beta = 0.70$ | **90.70%** | 0.034 | 91.80% | 0.031 | 92.40% | 0.015 |
| $\beta = 0.80$ | 90.50% | 0.041 | 92.00% | 0.030 | 92.50% | 0.016 |
| $\beta = 0.90$ | **90.70%** | 0.042 | 92.30% | 0.030 | 92.50% | 0.014 |
| $\beta = 1$ | 90.50% | 0.038 | **92.40%** | 0.025 | **92.60%** | 0.014 |
| $aff$ | 0.114 | 0.076 | 0.208 | 0.087 | 0.297 | 0.046 |

Table 27: Best classifier model and frequency of unobserved states by level of separability and sample size

| Structure | Level of separability | Sample size | | |
|---|---|---|---|---|
| | | Small | Moderate | |
| | | $n = 60$ | $n = 120$ | $n = 400$ |
| A | poorly | many states unobserv.<br>FOIM | some states unob.<br>DTM | all states obser.<br>Comb.($\beta = 0.6$) |
| B | moderately | many states unobserv.<br>FOIM | some states unob.<br>Comb.($\beta = 0.9$) | all states obser.<br>FOIM |
| C | moderately | many states unobserv.<br>Comb.($\beta = 0.8$) | many states unobserv.<br>FOIM | some states unob.<br>Comb.($\beta = 0.9$) |
| D | well | many states unobserv.<br>Comb.($\beta = 0.7$<br>and $\beta = 0.9$ ) | many states unobserv.<br>FOIM | some states unob.<br>FOIM |

**CASE 10**[j]

***Data description:***

In order to evaluate the impact of separability on the performance of ADD we consider simulated data - multinomial distributed (R function rmultinom (n, size, prob) ) - with poorly and well separated classes (affinity coefficient >0,7 for poorly separated and < 0,4 for well separated classes). We control for three additional factors on the experiments considering: the number of classes (C=2 and C=3), the sample dimension (e.g n=60 and n=120 for C=2 ) and balance - unbalance (1:2, for C=2, for example). Thirty random samples are generated for each structure with equal *prior* probabilities. Finally, we report the Pearson correlation coefficients between separability and performance measures (r averaged for 30 samples in each scenario).

Table 28: Best classifier model and frequency of unobserved states by level of separability and sample size

| Structure | | | 4 binary predictors vector: $P(X_i = 1)$ e $P(X_i = 0)$, i=1,...,4 |
|---|---|---|---|
| 2 Classes | poor separated | $C_1$ | (0.5; 0.5; 0.5; 0.5; 0.5; 0.5; 0.5; 0.5) |
| | | $C_2$ | (0.5; 0.5; 0.5; 0.5; 0.5; 0.5; 0.5; 0.5) |
| | well separated | $C_1$ | (0.1; 0.9; 0.7; 0.3; 0.2; 0.8; 0.6; 0.4) |
| | | $C_2$ | (0.9; 0.1; 0.3; 0.7; 0.8; 0.2; 0.1; 0.9) |
| 3 Classes | poor separated | $C_1$ | (0.45; 0.55; 0.55; 0.45; 0.45; 0.55; 0.55; 0.45) |
| | | $C_2$ | (0.6; 0.4; 0.4; 0.6; 0.6; 0.4; 0.4; 0.6) |
| | | $C_3$ | (0.4; 0.6; 0.6; 0.4; 0.4; 0.6; 0.6; 0.4) |
| | well separated | $C_1$ | (0.1; 0.9; 0.7; 0.3; 0.2; 0.8; 0.6; 0.4) |
| | | $C_2$ | (0.9; 0.1; 0.3; 0.7; 0.8; 0.2; 0.1; 0.9) |
| | | $C_2$ | (0.5; 0.5; 0.1; 0.9; 0.5; 0.5; 0.8; 0.2) |

[j] Poster presented at: The Twelfth International Symposium on Intelligent Data Analysis (IDA 2013), in Royal Statistical Society - London, 17-19 October 2013.

***Results:***

Table 29: Percentage of correctly classified: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Structure | size (balanced) | Best $P_C$ | s.d | Huberty Index | r(HI,aff) |
|---|---|---|---|---|---|
| poor | 60(30,30) | 53.7% ($\beta = 0.7$) | 0.078 | 13.10% | -0.20 |
| | 60(20,40) | 54.2% ($\beta = 0.6$) | 0.079 | -28.50% | -0.32 |
| | 120(60,60) | 51.6% (DTM) | 0.034 | 7.60% | -0.23 |
| | 120(30,90) | 52.4% ($\beta = 0.9$) | 0.052 | -80.30% | -0.53 |
| well | 60(30,30) | 90.7% ($\beta = 0.9$) | 0.042 | 85.4% | -0.76 |
| | 60(20,40) | 54.2% ($\beta = 0.9$) | 0.042 | 72.40% | -0.73 |
| | 120(60,60) | 92.1% (FOIM) | 0.026 | 85.40% | -0.75 |
| | 120(30,90) | 91.0% ($\beta = 0.9$) | 0.035 | 68.40% | -0.72 |

Table 30: Percentage of correctly classified: $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Structure | size (balanced) | Best $P_C$ | s.d | Huberty Index | r(HI,aff) |
|---|---|---|---|---|---|
| poor | 90(30,30,30) | 39.0% (FOIM) | 0.073 | -16.50% | -0.52 |
| | 90(20,20,50) | 42.0% ($\beta = 0.8$) | 0.054 | -24.60% | -0.28 |
| | 180(60,60,60) | 41.9% ($\beta = 0.7$) | 0.048 | -12.40% | -0.64 |
| | 180(30,50,90) | 44.8% ($\beta = 0.8$) | 0.055 | -6.70% | -0.28 |
| well | 90(30,30,30) | 74.7% ($\beta = 0.7$) | 0.050 | 52.6% | -0.57 |
| | 90(20,20,50) | 75.3% ($\beta = 0.9$) | 0.059 | 47.00% | -0.35 |
| | 180(60,60,60) | 76.8% (FOIM) | 0.032 | 54.90% | -0.74 |
| | 180(30,50,90) | 75.5% (FOIM) | 0.038 | 52.70% | -0.50 |

Table 31: Percentage of correctly classified: (HIERM) - $(\beta) * \hat{P}_{FOIM} + (1 - \beta) * \hat{P}_{DTM}$

| Structure | size (balanced) | Best $P_C$ | s.d | Huberty Index | r(HI,aff) |
|---|---|---|---|---|---|
| poor | 90(30,30,30) | 52.6% ($\beta = 0.3$) | 0.059 | 11.30% | -0.38 |
| | 90(20,20,50) | 56.4% ($\beta = 0.8$) | 0.055 | 8.40% | -0.25 |
| | 180(60,60,60) | 54.8% ($\beta = 0.4$) | 0.061 | 13.20% | -0.27 |
| | 180(30,50,90) | 58.7% ($\beta = 0.7$) | 0.040 | 21.40% | -0.02 |
| well | 90(30,30,30) | 80.9% ($\beta = 0.7$) | 0.050 | 65.90% | -0.37 |
| | 90(20,20,50) | 81.8% (FOIM) | 0.064 | 61.90% | -0.25 |
| | 180(60,60,60) | 83.3% (FOIM) | 0.027 | 67.80% | -0.66 |
| | 180(30,50,90) | 83.6% (FOIM) | 0.032 | 69.10% | -0.10 |

Prazeres (1996) Gudjonsson (1997) Duarte (2009) Lima (1998)

# References

Abbott, D. W. (1999). Combining models to improve classifier accuracy and robustness. In *Proceedings of Second International Conference on Information Fusion, Fusion'99*, volume 1, pages 289–295. 101

Amershi, S. and Conati, C. (2009). Combining unsupervised and supervised classification to build user models for exploratory. *JEDM-Journal of Educational Data Mining*, 1(1):18–71. 14, 17, 101

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588. 7

Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis. *Methods of operations research*, 53:507–512. 19

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous variables. *In Studies in analysis and prediction, Solomon H. (éd.), Palo Alto*, pages 185–188. 11

Benjamini, Y. and 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)*, pages 289–300. 26, 27

Bostanci, B. and Bostanci, E. (2013). An evaluation of classification algorithms using mc nemar's test. In *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, pages 15–26. Springer. 23

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140. 14, 15, 16

Breiman, L. (1998). Half & half bagging and hard boundary points. Technical report, Technical Report, Statistics Department, University of California. 14, 16

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. 6, 7, 8

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press. 6, 7, 8

Brito, I. (2002). *Combinaison de modéles en analyse discriminante dans un contexte gaussien*. PhD thesis, Grenoble 1. iii, v, vii, 14, 16, 17, 101

Brito, I., Celeux, G., and Sousa Ferreira, A. (2006). Combining methods in supervised classification: A comparative study on discrete and continuous problems. *REVSTAT–Statistical Journal*, 4(3):201–225. iii, v, vii, 16, 17, 19, 101

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254. 23

Celeux, G. (1990). *Analyse discriminante sur variables continues*. INRIA. 20

Celeux, G. and Nakache, J.-P. (1994). *Analyse discriminante sur variables qualitatives*. Politechnica. iii, v, vii, 10, 11, 12, 19, 101

Cesa-Bianchi, N., Claudio, G., and Luca, Z. (2006). Hierarchical classification: combining bayes with svm. *Proceedings of the 23rd international conference on Machine learning. ACM*. 14, 17, 101

Cook, R. D. and Yin, X. (2001). Theory & methods: Special invited paper: Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199. 24

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297. 6, 8

Das Gupta, S. (1973). Theories and methods in classification: a review. *Discriminant analysis and applications*, pages 77–137. 1

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(3):131–156. 24, 25

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30. 23

Dietterich, T. G. (1997). Machine-learning research. *AI magazine*, 18(4):97. 14

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923. 23

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157. 16

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130. 17

Duarte, A. (2009). *A satisfação do consumidor nas instituições culturais. O caso do Centro Cultural de Belém*. PhD thesis, Tese de Mestrado em Marketing do ISCTE. 136

Elder, J. and Pregibon, D. (1995). A statistical perspective on knowledge discovery in databases. advances in knowledge discovery and data mining. 14

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188. 2

Foody, G. M. (2004). Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric engineering and remote sensing*, 70(5):627–634. 23

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156. 14, 15, 16

Friedman, J., Hastie, T., and Tibshirani, R. (1998). Additive logistic regression: A statistical view of boosting. *Technical Report*. 14, 16

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232. 14, 16

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954. 14

Goldstein, M. and Dillon, W. (1978). *Discrete Discriminant Analysis*. New York: Wiley. iii, v, vii, 9, 10

Gomez, D. and Montero, J. (2011). Determining the accuracy in image supervised classification problems. In *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology*, pages 342–349. Atlantis Press. 23

Gudjonsson, G. H. (1997). *The Gudjonsson suggestibility scales manual.* Psychology Press, Publishers. 136

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182. 24

Habema, J., H. J. and Van Den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In *Proceedings of Computational Statistics, Compstat 1974*, pages 101–110. Physica-Verlag. 17

Hill, T. L. (1967). Proceedings of the national academy of sciences. pages 111–114. USA 58. 6, 8

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, Canada*, pages 278–282. 7

Ho, T. K. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 20(8):832–844. 7

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70. 27

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37. 14

Janusz, A. (2010). Combining multiple classification or regression models using genetic algorithms. In *Rough Sets and Current Trends in Computing*, pages 130–137. Springer. 14, 17, 101

Kotsiantis, S. B. (2011). A random subspace method that uses different instead of similar models for regression and classification problems. *International Journal of Information and Decision Sciences*, 3(2):173–188. 14, 17, 101

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190. 14

Lemeshow, S. and Hosmer, D. (2000). *Applied Logistic Regression (Wiley Series in Probability and Statistics*. Wiley-Interscience; 2 Sub edition. 5, 8

Lima, M. (1998). *Orientação e Desenvolvimento da Carreira em Estudantes Universitários. Estudo das Atitudes de Planeamento e Exploração, Identidade Vocacional, Saliência dos Papéis e Fatores de Carreira*. PhD thesis, Dissertação de doutorado não publicada, Universidade de Lisboa, Lisboa, Portugal. 136

Marques, A., Sousa Ferreira, A., and Cardoso, M. G. (2008). Uma proposta de combinação de modelos em análise discriminante discreta. *Estatística - Arte de Explicar o Acaso, in Oliveira, I. et al. Editores, Ciência Estatística*, pages 393–403. 18, 19, 28

Marques, A., Sousa Ferreira, A., and Cardoso, M. G. (2010). Classification and combining models. *In Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010)*, pages 495–504. 18, 19, 28

Marques, A., Sousa Ferreira, A., and Cardoso, M. G. (2013). Selection of variables in discrete discriminant analysis. *Biometrical Letters*, 50(1):1–14. 20, 28, 102

Marques, A., Sousa Ferreira, A., and Cardoso, M. G. (2014a). Combining models in discrete discriminant analysis. *Int. J. of Data Analysis Techniques and Strategies. (to appear)* 18, 19, 23, 28, 103

Marques, A., Sousa Ferreira, A., and Cardoso, M. G. (2014b). Performance of combined models on binary discrete classification. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. (under review)* 18, 19, 28, 103, 104

Matusita, K. (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. *The Annals of Mathematical Statistics*, pages 631–640. 19

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133. 2

McKay, R. and Campbell, N. (1982). Variable selection techniques in discriminant analysis: Ii. allocation. *British Journal of Mathematical and Statistical Psychology*, 35(1):30–41. 25

McLachlan, G. (1992). Discriminant analysis and statistical pattern recognition. *Wiley New York.* 24, 25

Milgram, J., Mohamed, C., and Robert, S. (2004). Speeding up the decision making of support vector classifiers. *Frontiers in Handwriting Recognition*, IWFHR-9 2004. 14, 17

Murphy, T. B., Dean, N., and Raftery, A. E. (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The annals of applied statistics*, 4(1):396. 24, 25

Opitz, D. W. and Shavlik, J. W. (1996). Generating accurate and diverse members of a neural-network ensemble. *Advances in neural information processing systems*, pages 535–541. 16

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann. iii, v, vii, 11, 12

Prazeres, N. (1996). *Ensaio de um Estudo sobre Alexitimia com o Rorschach ea Escala de Alexitimia de Toronto (TAS-20).* 136

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann. 6, 17

Re, M. and Valentini, G. (2011). Ensemble methods: a review. *Chapman and Hall.* 14

Rebouças, S. M. D. P. (2011). Metodologias de classificação supervisionada para análise de dados de microarrays. 24, 25

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). *Learning representations by back-propagating errors*. MIT Press, Cambridge, MA, USA. 2

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *Artificial Neural Networks–ICANN 2009*, pages 175–184. Springer. 23

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2):197–227. 16

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press. 5, 8

Silva, P. D. (1999). Análise discriminante com seleçao de variáveis. pages 5–41. *Revista de Estatística.* 24

Silva, P. D. (2010). Classificação supervisionada para dados de elevada dimensão. *ISCTE - Instituto Universitário de Lisboa.* 27

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press. 6

Sousa Ferreira, A. (1987). *Análise Factorial Discriminante.* PhD thesis, Dissertação de Mestrado apresentada á Faculdade de Ciências da Universidade de Lisboa. 25

Sousa Ferreira, A. (2000). *Combinação de modelos em análise discriminante sobre variáveis qualitativas*. PhD thesis, Ph. D. thesis, University of Nova de Lisboa. iii, v, vii, 16, 17, 18, 19, 101, 102

Sousa Ferreira, A. (2004). Combining models approach in discrete discriminant analysis through a committee of methods. *Classification, Clustering, and Data Mining Applications*, pages 151–156. 101

Sousa Ferreira, A. (2010). A comparative study on discrete discriminant analysis through a hierarchical coupling approach. In *Classification as a Tool for Research*, pages 137–145. Springer. 101

Sousa Ferreira, A. and Cardoso, M. (2013). Evaluating discriminant analysis results. In *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*, pages 155–162. Springer. 23

Sousa Ferreira, A., Celeux, G., and Bacelar-Nicolau, H. (1999). Combining models in discrete discriminant analysis by a hierarchical coupling approach. *Applied Stochastic Models and Data Analysis, ASMDA*, 99:159–164. 16

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*, pages 181–186. Springer. 14, 16

Steinberg, D. (1997). Cart users manual. *Salford Systems*. 14

Wald, A. (1955). On a statistical problem arising in the classification of an individual into one of two groups1. *Selected Papers in Statistics and Probability*, page 391. 1, 2, 4

Welch, B. L. (1939). Note on discriminant functions. *Biometrika*, pages 218–220. 1, 2, 4

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560. 2

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259. 14, 16