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 $K \geq 2$ *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

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

then assign $\underline{\mathrm{x}}$ to $\mathrm{G_k}$. π_l represents the *a priori* probability of group l ($\mathrm{G_l}$), and $\mathrm{P}(\underline{\mathrm{x}}|\mathrm{G_l})$ 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 $\mathrm{P}(\underline{\mathrm{x}}|\mathrm{G_l})$ 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 $\mathrm{G_k}$, the corresponding conditional probability being estimated by:

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

where $\mathrm{n_k}$ represents the $\mathrm{G_k}$'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 $\mathrm{I(X_i, X_j)}$ between two variables

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

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./ Gk	$\hat{P}$ (x1,x2,x3,x4) for group Gk		
		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 $F_2=\{q_l\}$, $l=1,\ldots,L$ 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: G_2+G_1 vs G_3+G_4 β*FOIM+ (1-β)*DTM	β = 0 (DTM)	45,6%	20,00%
	β = 0,25	59,2%	40,00%
	β = 0,50	**60,8%**	**42,35%**
	β = 0,75	**60,8%**	**42,35%**
	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**	
	n_k	**%**	**n_k**	**%**	**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+	β = 0 (DTM)	45,00%	-3,77%
(1-β)*DTM	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:	β = 0 (DTM)	47,80%	1,51%
G_2 vs G_1+G_3+G_4	β = 0,20	48,10%	2,08%
	β = 0,40	49,30%	4,34%

β*FOIM+	β = 0,50	49,30%	4,34%
(1-β)*DTM	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).