Evaluating discriminant analysis results

Ana Sousa Ferreira and Margarida Cardoso

Abstract In Discrete Discriminant Analysis (DDA) different models often exhibit different classification performances. Therefore, the idea of combining models has increasingly gained importance. In the present work we focus on the evaluation of alternative DDA models, including combined models. The proposed approach uses not only the classic indicators of classification precision, but also indices of agreement that regard the relationship between the actual classes and the ones predicted by discriminant analysis. The performance of the DDA methods is analyzed based on simulated binary data, using small and moderate sample sizes. The results obtained illustrate the potential of combining DDA models, offering different evaluation perspectives.

Key words: CART, combining models, hierarchical coupling model

1 Introduction

In Discrete Discriminant Analysis (DDA) different models often exhibit different classification performances for different individuals or observations. This seems to be a particularly relevant issue in the small or moderate sample setting and when the classes are not well separated. Therefore, the idea of combining models currently appears in an increasing number of DDA papers, in an attempt to obtain more robust and stable models.

In this paper we compare the performance of the Full Multinomial Model (FMM) ([9]) and the First-order Independence Model (FOIM) ([9]) with a model based on

Ana Sousa Ferreira
LEAD, FP, Universidade de Lisboa, CEAUL and UNIDE, Alameda da Universidade 1649-013 Lisboa
e-mail: asferreira@fp.ul.pt

Margarida Cardoso
Dep. Métodos Quantitativos and UNIDE, ISCTE-Instituto Universitário de Lisboa, Avenida das Forças Armadas, 1600-083 Lisboa
e-mail: mgsc@iscte.pt

the two referred models that produces an intermediate model between them. In order to deal with the multi-class case we use the Hierarchical Coupling Model (HIERM) (e.g., [3], [15]) that enables to reduce the problem into several bi-class problems embedded in a binary tree. The comparison is extended to the results of the CART - Classification and Regression Trees algorithm ([2]), a classical approach within the classification domain.
The performance of the alternative models considered is compared based on simulated data. To evaluate this performance we consider several measures of precision including traditional classification indices and indices of agreement between the actual classes and the ones predicted by the DDA methods. Results obtained refer to two-fold cross-validation.

2 Methodological approach

In the present study, a new methodology is proposed for the evaluation of DDA results. It enables the comparison of DDA classical models with the DDA combining models approach. The proposed methodology relies on indices of agreement between the actual and predicted (by DDA) classes and is illustrated using simulated data according to the Bahadur model.

2.1 Indices for evaluating classification results

When evaluating results from classification we focus on the $K \times K$ confusion matrix $\mathbf{M} = [n_{ij}]$ which is a contingency table of the actual classes (lines refer to partition $\prod_a^K$ with K classes) by the ones predicted by discriminant analysis (columns refer to partition $\prod_b^K$ with K classes). The row totals are $n_{i.}$, $(i = 1, \ldots, K)$.
Some commonly used indices depend only on the diagonal of the referred matrix, which adds up to the number of correctly classified observations (see Table 1). The percent agreement varies between 0 (null classification precision) and 100% (perfect classification precision). The Cohen's Kappa deducts agreement by chance and the Huberty index deducts the percentage of correctly classified by default (majority class rule).

Table 1 Indices of agreement based on the diagonal of the confusion matrix

Indices	Definition
Percent agreement	Perc-agree$(\prod_a^K, \prod_b^K) = (\sum_{k=1}^K n_{kk})/n$
Cohen's Kappa ([6])	Kappa$(\prod_a^K, \prod_b^K) = (\sum_{k=1}^K n_{kk} - \sum_{k=1}^K n_{k.} n_{.k}/n)/(n - \sum_{k=1}^K n_{k.} n_{.k}/n)$
Huberty ([12])	Huberty$(\prod_a^K, \prod_b^K) = ((\sum_{k=1}^K n_{kk})/n - \max_i n_{i.}/n)/(1 - \max_i n_{i.}/n)$, where $n_{i.}$, $(i = 1, \ldots, K)$ are the row totals

In this work we suggest further exploring the confusion matrix to evaluate the agreement between $\prod_a^K$ and $\prod_b^K$. We thus consider the indices on Table 1 and Table 2 (see [4], for further details). The Cramer's V statistic quantifies simple agreement and Variation of Information (VI) considers entropy and mutual information. They vary between 0 and 1 (0 indicating null agreement for Cramer's V and perfect agreement for the normalized VI). The Adjusted Rand quantifies paired agreement deducting agreement by chance. *A priori*, the advantage of using these indices is to complement the evaluation of agreement between partitions $\prod_a^K$ and $\prod_b^K$.

Table 2 Indices of agreement based on the complete confusion matrix

Indices	Definition
Cramer's V ([7])	$\mathrm{V}(\prod_a^K,\prod_b^K) = \sqrt{Chi - sq(\prod_a^K,\prod_b^K)/(nK-n)}$ where $Chi - sq(\prod_a^K,\prod_b^K) = \sum_{k=1}^K\sum_{q=1}^K (n_{kq} - \frac{n_{k.}n_{.q}}{n})^2 / \frac{n_{k.}n_{.q}}{n}$
Adjusted Rand ([11])	$\text{Adj-Rand}(\prod_a^K,\prod_b^K) = \frac{\sum_{k=1}^K\sum_{q=1}^K C_2^{n_{kq}} - \sum_{k=1}^K C_2^{n_{k.}} \sum_{q=1}^K C_2^{n_{.q}}/C_2^n}{\frac{1}{2}[\sum_{k=1}^K C_2^{n_{k.}} + \sum_{q=1}^K C_2^{n_{.q}}] - \sum_{k=1}^K C_2^{n_{k.}} \sum_{q=1}^K C_2^{n_{.q}}/C_2^n}$
Normalized Variation of Information([14])	$N\_VI(\prod_a^K,\prod_b^K) = [\mathrm{H}(\prod_a^K) + \mathrm{H}(\prod_b^K) - 2\mathrm{I}(\prod_a^K,\prod_b^K)]/\log n$ where H indicates the entropy $\mathrm{H}(\prod^K) = \sum_{k=1}^K \frac{n_k}{n}\log\frac{n_k}{n}$ and I indicates the mutual information $I(\prod_a^K,\prod_b^K) = \sum_{k=1}^K\sum_{q=1}^K \frac{n_{kq}}{n}\log\frac{n_{kq}}{n_{k.}n_{.q}/n}$

2.2 Simulated data

The performance of the DDA methods is analyzed based on simulated binary data. We use the Bahadur model, as proposed in Godstein and Dillon ([5], [9]), to simulate the predictive binary variables' values. This model representation defines class conditional probabilities for class $C_k, (k = 1,...,K)$ as

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

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

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

We consider two types of population structures with P=6 variables and for illustrative purposes, let us consider the case of K=2 classes and one of the most usual multi-class case, K=4 classes. Location parameters are described on Table 3.

For each structure, data sets generated have small sample sizes (60 observations for each class) and moderate sample sizes (200 observations for each class). The training and test samples represent 50% of the total of observations.

Table 3 Parameters for simulated Bernoulli variables

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

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

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

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

The prior probabilities are considered equal.

2.3 Discrete discriminant analysis

In discrete classification problems the most natural model is the Full Multinomial Model (FMM) where the conditional probabilities are estimated by the observed frequencies ([9]). This model involves $2^P - 1$ parameters to be estimated in each class. Hence, even for moderate P (e.g., ten binary variables leads to 1023 parameters to be estimated), generally, not all of the parameters are identifiable.

One way to deal with this problem consists in reducing the number of parameters to be estimated. The First-order Independence Model (FOIM) assumes that the P binary variables are independent in each class C_k, $k = 1,...,K$ ([9]). Then, the number of parameters to be estimated for each class is reduced from $2^P - 1$ to P.

Since we are mainly concerned with small or moderate sample sizes, we may encounter a problem of sparseness in which some of the multinomial cells may have no

data in the training sets. Therefore, we suggest to smooth the observed frequencies of model FMM as follows:

$$P(\mathbf{x}|\lambda) = \frac{1}{n}\sum_{i=1}^{n} \lambda^{P-\|\mathbf{x}-x_i\|}(1-\lambda)^{\|\mathbf{x}-x_i\|}, 0 < \lambda \leq 1 \tag{3}$$

where $\lambda = 1.00$, $\lambda = 0.99$, $\lambda = 0.95$ or $\lambda = 0.90$ according to the training sample size.
In this work, taking into account the size of our samples, we consider $\lambda = 1.00$ (no smoothing) or $\lambda = 0.95$ (moderate smoothing) for all samples.
Note that according to Hand ([10]), we opt for a computationally less demanding method since the choice of the smoothing method is not particularly important.

FMM and FOIM provide different classifications in many circumstances. Therefore, we expect a combining model (using a single coefficient β for the linear combination of FMM and FOIM) to yield better results.
There are several strategies to estimate the coefficient β (e.g., [3], [15]) that combines the two referred models. A natural way of deriving this coefficient is by minimizing the fitting error using a least squares criterion ([15],[16]). For the two classes case, we use an approach to estimate the coefficient β using a least squares regression (LSR) criterion:

$$\widehat{\beta}_{LSR} = \frac{\sum_{i=1}^{n}(l_2(x_i) - l_1(x_i))l_2(x_i) - \sum_{i=1}^{n} y_i(l_2(x_i) - l_1(x_i))}{\sum_{i=1}^{n}(l_2(x_i) - l_1(x_i))^2} \tag{4}$$

where y_i denotes a indicator of class membership for observation i and l_1, l_2 represent, respectively, the log ratio of the class conditional probabilities for model FMM and FOIM (denoted by LSR1) or the *a posteriori* probabilities of the first class for FOIM and FMM models (denoted by LSR2), estimated by cross-validation in a sample of size n.

In the multi-class case, we use the Hierarchical Coupling Model (HIERM), inspired by Friedman's approach ([8]), for reducing the multi-class problem into several bi-class problems embedded in a binary tree. HIERM needs two decisions at each level:

1. Selecting the hierarchical coupling among the $2^{K-1} - 1$ possible classes couple;
2. In each node of the tree, selecting the combining model that gives the best classification rule for the chosen couple.

At the beginning we have K classes that we want to reorganize into two classes. So, we propose to select the two new classes that are the most separable. The basic affinity coefficient ([1], [13]) can be used to select the hierarchical coupling at each level of the tree.
Denoting $F_1 = q_j^1$ and $F_2 = q_j^2$, $j = 1,...,P$ two discrete distributions defined in the same space, the affinity coefficient is defined by

$$\rho(F_1, F_2) = \sum_j \sqrt{q_j^1}\sqrt{q_j^2}, j = 1,...P \tag{5}$$

and is easily computed in our classification problem. The individual vector **x** is assigned to the class associated with the last node of the tree on which **x** falls.
The main aim of this approach is to obtain a better prediction performance and improve results stability.

3 Experimental results

After running discriminant analysis for the simulated data we obtain the results presented in Table 4 and Table 5.
When referring to the combining models we simply present the results yielded by the best strategy (LSR1 or LSR2). For the sake of simplicity, we only report the best FMM results (smoothed or not).

Table 4 Small samples results/ Cross-validation (two-fold results)

Data	Methods	Perc-Agree	Kappa	Huberty	Cramer's V	Adj-Rand	N-VI
	CART	52%	5%	-7%	0.048	-0.019	0.355
IND	FMM	31%	21%	-2%	0.226	0.040	0.340
K=2	FOIM	58%	16%	4%	0.198	0.013	0.328
	LSR2	60%	21%	11%	0.222	0.025	0.320
	CART	77%	54%	48%	0.559	0.291	0.242
DIF	FMM	65%	50%	30%	0.520	0.286	0.245
K=2	FOIM	58%	17%	0%	0.165	0.004	0.335
	LSR2	76%	52%	46%	0.400	0.097	0.278
	CART	28%	5%	-1%	0.156	-0.005	0.536
IND	FMM	0%	$*$	0%	$*$	$*$	$*$
K=4	FOIM	30%	6%	3%	0.173	0.005	0.534
	LSR2	50%	34%	30%	0.505	0.208	0.368
	CART	23%	-1%	-6%	$*$	-0.010	$*$
DIF	FMM	10%	-20%	-23%	0.347	0.083	0.472
K=4	FOIM	32%	12%	6%	0.241	0.036	0.510
	LSR1	48%	31%	29%	0.426	0.135	0.474

$*$ Not defined (null observed frequency in denominator)

Table 5 Moderate samples results/ Cross-validation (two-fold results)

Data	Methods	Perc-agree	Kappa	Huberty	Cramer's V	Adj-Rand	N-VI
	CART	54%	8%	6%	0.078	0.004	0.258
IND	FMM	55%	14%	10%	0.139	0.014	0.259
K=2	FOIM	59%	17%	15%	0.172	0.025	0.255
	LSR2	60%	19%	17%	0.195	0.031	0.253
	CART	69%	37%	36%	0.398	0.138	0.219
DIF	FMM	61%	32%	23%	0.346	0.120	0.208
K=2	FOIM	50%	-1%	-3%	0.039	-0.022	0.261
	LSR2	63%	30%	24%	0.333	0.100	0.224
	CART	33%	11%	9%	0.154	0.016	0.447
IND	FMM	0%	∗	0%	∗	∗	∗
K=4	FOIM	35%	13%	12%	0.225	0.043	0.429
	LSR2	44%	26%	25%	0.327	0.093	0.407
	CART	29%	6%	4%	0.105	0.002	0.425
DIF	FMM	11%	-18%	-20%	0.221	0.039	0.431
K=4	FOIM	35%	13%	12%	0.220	0.038	0.433
	LSR1	46%	28%	27%	0.393	0.130	0.362

∗ Not defined (null observed frequency in denominator)

In these results, the DDA methods seem to perform similarly for the small and moderate sized samples. Except for the case of DIF and K=2 (where the best results are attained by CART) the combined models evidence the best performances.

4 Discussion and perspectives

In general, the best DDA results are obtained using the combining models approach, with the LSR2 strategy where the *a posteriori* probabilities characterize the class conditional probabilities.

The various indicators used to evaluate DDA results offer different insights regarding the confusion matrix and the corresponding results do not necessarily agree (see correlations in Table 6). Note that we consider small and moderate size samples when computing correlations, since they exhibit similar (correlation) patterns.

Table 6 Pearson correlations (r)

Methods	Perc-agree	Kappa	Huberty	Cramer's V	Adj-Rand	N-VI
Perc-agree	1					
Kappa	0.807	1				
Huberty	0.790	0.952	1			
Cramer's V	0.339	0.709	0.699	1		
Adj-Rand	0.436	0.739	0.711	0.948	1	
N-VI	-0.807	-0.516	-0.464	-0.181	-0.307	1

The Percent Agreement index is strongly related with the Normalized Variation of Information index which has the advantage of quantifying not only the correctly classified cases, but also the relationship between the incorrectly classified ones. The Cramer's V statistic and the Adjusted Rand index are strongly related as well as the Kappa and the Huberty indices. These indicators offer a different perspective, quantifying simple agreement and paired agreement between the actual classes and the predicted ones.
In future research, the advantages of using indices of agreement for evaluating DDA results should be further explored. In addition, real data should be used to further illustrate the utility of the proposed approach.

References

1. Bacelar-Nicolau, H.: The affinity coefficient in cluster analysis. Meth.Oper. Res. **53**, 507–512 (1985)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. California, Wadsworth, Inc. (1984)
3. Brito, I., Celeux, G., Sousa Ferreira, A.: Combining methods in supervised classification: A comparative study on discrete and continuous problems. REVSTAT - Statistical Journal **4**(3), 201–225 (2006)
4. Cardoso, M.G.M.S.: Clustering and cross-validation. In: Paper presented at the IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction (2007)
5. Celeux, G., Mkhadri, A.: Discrete regularized discriminant analysis. Statistics and Computing **2**(3), 143–151, DOI:10.1007/BF01891,206 (1992)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**, 37–46 (1960)
7. Cramér, H.: Mathematical Methods of Statistics. Princeton, NJ:Princeton University Press (1946)
8. Friedman, J.: Another approach to polychotomous classification. Tech. rep., Stanford University (1996)
9. Goldstein, M., Dillon, W.: Discrete Discriminant Analysis. New York, Wiley (1978)
10. Hand, D.: Kernel Discriminant Analysis. Research Studies Press. Wiley, Chichester (1982)
11. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985)
12. Huberty, C.J., Olejnik, S.: Applied MANOVA and Discriminant Analysis. Wiley-Interscience. John Wiley and Sons (2006)
13. Matusita, K.: Decision rules based on distance for problems of fit, two samples and estimation. Ann. Inst. Stat. Math. **26**(4), 631–640 (1955)
14. Meila, M.: Comparing clusterings-an information based distance. Journal of Multivariate Analysis **98**(5), 873–895, DOI:10.1007/BF01891,206 (2007)
15. Sousa Ferreira, A.: Combinação de modelos em análise discriminante sobre variáveis qualitativas. Ph.D. thesis, Univ. Nova de Lisboa (2000)
16. Sousa Ferreira, A.: Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization, chap. A Comparative Study on Discrete Discriminant Analysis through a Hierarchical Coupling Approach. Hermann Locarek-Junge, Claus Weihs (eds.) Springer, Berlin-Heidelberg-New York (2009)