COMUNICAÇÃO ORAL

Desempenho de uma abordagem de Combinação de Modelos em Análise Discriminante Discreta

Anabela Marques
Escola Superior de Tecnologia do Barreiro, IPS, *anabela.marques@estbarreiro.ips.pt*

Ana Sousa Ferreira
LEAD, Faculdade de Psicologia, Universidade de Lisboa, UNIDE e CEAUL, *asferreira@fp.ul.pt*

Margarida Cardoso
Dep. de Métodos Quantitativos do ISCTE- Instituto Universitário de Lisboa, UNIDE,
margarida.cardoso@iscte.pt

Anabela Marques
Escola Superior de Tecnologia do Barreiro, IPS, *anabela.marques@estbarreiro.ips.pt*

Palavras–chave: Análise Discriminante Discreta; Combinação de modelos; Modelo de Emparelhamento Hierárquico; Modelo Gráfico Decomponível; Modelo de Independência Condicional

Abstract: Deparamo-nos frequentemente nas mais diversas áreas da vida quotidiana com problemas de classificação onde vários modelos revelam um fraco desempenho, particularmente em presença de classes mal separadas e/ou amostras de pequena dimensão. A abordagem da combinação de modelos surgiu, então, naturalmente com o objectivo de encontrar novos métodos que se adaptassem melhor ao comportamento dos dados em estudo, usando a contribuição de vários modelos, minimizando assim o número de observações mal classificadas. Este trabalho insere-se no campo da Análise Discriminante Discreta, que tem vindo a despertar um interesse crescente, nomeadamente nas áreas das ciências sociais e da saúde. Neste trabalho, pretende-se avaliar o desempenho da combinação de modelos proposta por Marques *et al.* [3], sobre dados simulados com base no modelo de Bahadur.

1 Introdução

A Análise Discriminante Discreta (ADD) é utilizada em muitas situações da vida real sendo comum dispormos de classes *a priori* mal separadas e/ou de amostras de pequena dimensão, tornando difícil a estimação de um grande número de parâmetros habitualmente associados aos modelos de classificação no campo discreto, dificultando assim a tarefa de afetação nas classes definidas *a priori*. Ao longo das últimas décadas, a combinação de modelos começou a ser proposta por diversos investigadores com o objectivo de encontrar métodos de classificação que se adaptem melhor ao comportamento dos dados em estudo e que conduzam à minimização do número de parâmetro a estimar. Nos estudos desenvolvidos por Sousa Ferreira, (Sousa Ferreira [4]; Sousa Ferreira *et al.* [5]), verificou-se que a abordagem pela combinação de modelos conduzia a modelos mais eficientes e estáveis, tanto mais que frequentemente se observava que os erros de afetação obtidos por vários modelos não ocorriam sobre os mesmos objectos (Brito *et al.* [1]). Na sequência destes estudos, Marques *et al.* [3] propuseram uma combinação linear entre o Modelo de Independência Condicional (MIC) e o Modelo Gráfico Decomponível (MGD), recorrendo a um único coeficiente β ($0 \leq \beta \leq 1$), conduzindo a um modelo intermédio entre estes dois modelos de classificação.
Neste trabalho, iremos avaliar o desempenho da combinação de modelos proposta por Marques *et al.* [3], recorrendo para tal a uma bateria de dados simulados, procurando desta forma perceber o

seu campo privilegiado de aplicação.
Esta avaliação do desempenho vai basear-se na percentagem de observações correctamente classificadas estimada quer na amostra de treino (estimação por resubtituição), quer na amostra de teste ou ainda por validação cruzada.

2 Estrutura dos dados simulados

Para avaliar o desempenho do referido modelo, recorrendo a dados simulados com base no modelo de Bahadur (Goldstein e Dillon [2]), consideramos dois tipos de estrutura relacional entre as variáveis explicativas:
- IND - simulam-se observações a partir do modelo MIC, isto é considera-se que as variáveis explicativas são independentes dentro de cada classe;
- DIF - as observações simuladas evidenciam a existência de relações diferenciadas entre as variáveis explicativas nas várias classes em estudo.
Para simular os valores das variáveis binárias, o modelo de Bahadur considera as probabilidades condicionadas para cada classe $C_k, (k = 1, ..., K)$ como sendo:

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

onde X_{kp} é uma variável de Bernoulli com parâmetro $\theta_{kp} = E(X_{kp}), p = 1, ..., P$ tal que

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

Neste trabalho, iremos considerar os dois tipos de estruturas de relação entre as variáveis, P=6 variáveis binárias para o caso de duas, três ou quatro classes definidas *a priori*, e dimensões das amostras pequenas ou moderadas.

Referências

[1] 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(3) (2006), 201–225.

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

[3] Marques A., Sousa Ferreira A. and 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*, Ediões S.P.E, 393–403.

[4] 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. 181–186, Springer.

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