Comunicação oral

Selecção de variáveis em Análise Discriminante Discreta

Anabela Marques
Escola Superior de Tecnologia do Barreiro, IPS e CEAUL , anabela.marques@estbarreiro.ips.pt
Ana Sousa Ferreira
Faculdade de Psicologia, Universidade de Lisboa e UNIDE, asferreira@fp.ul.pt
Margarida Cardoso
Dep. Métodos Quantitativos, Escola de Gestão do ISCTE e UNIDE, margarida.cardoso@iscte.pt

Palavras–chave: Análise discriminante discreta, combinação de modelos, selecção de variáveis.

Abstract: Na Análise Discriminante Discreta deparamo-nos frequentemente com o problema da dimensionalidade das variáveis em estudo, dispondo muitas vezes de conjuntos de dados onde o número de variáveis é elevado face ao número de objectos em análise.
Neste trabalho, iremos avaliar o desempenho de métodos de classificação discreta, seguindo uma abordagem de combinação de modelos, recorrendo à aplicação de várias técnicas de selecção de variáveis.

1 Introdução

A Análise Discriminante Discreta é uma técnica de Análise de Dados Multivariada que perante um conjunto de objectos, descritos por variáveis explicativas, provenientes de grupos definidos *a priori*, tem como objectivos conhecer as váriaveis de entre as observadas, que melhor diferenciam estes grupos e mediante a aplicacão de uma regra de decisão, classificar novos objectos a um dos grupos definidos *a priori*.
Neste trabalho, iremos utilizar uma abordagem pela combinação de modelos, uma vez que se tem verificado que a sua aplicação conduz a métodos mais estáveis e robustos, recorrendo ainda à aplicação de várias técnicas de selecção de variáveis abordando a questão da dimensionalidade em classificação.
Apresentaremos um estudo comparativo de alguns métodos de selecção de variáveis, demonstrando a sua importância para a obtencão de bons resultados de classificação.
A combinação de modelos em análise será a proposta por Marques *et al.* ([5]), na continuação de estudos anteriores, conduzindo a um modelo intermédio entre o Modelo de Independência Condicional (MIC) e o Modelo Gráfico Decomponível (MGD)(Celeux e Nakache, [2]). Esta abordagem de combinação de modelos, no caso de mais de dois grupos *a priori*, utiliza o Modelo de Emparelhamento Hierárquico (MHIER), o qual decompõe o problema inicial em vários problemas de dois grupos *a priori*, utilizando uma estrutura de árvore binária.

2 Selecção de variáveis

Na aplicação da Análise Discriminante Discreta a um conjunto de dados, deparamo-nos muitas vezes com problemas de dimensionalidade, nomeadamente em dados provenientes das áreas das ciências sociais, humanas e da saúde, onde aparecem para classificar, objectos com um número de variáveis explicativas elevado face ao número de objectos em estudo. Têm surgido, na literatura, vários métodos para determinar as variáveis que discriminam mais eficazmente entre os grupos, desejavelmente em número consideravelmente inferior ao número inicial de variáveis, permitindo desta forma, reduzir o número de parâmetros a estimar, o tempo de execução dos métodos utilizados e facilitando a interpretação dos resultados.

Com esta finalidade, iremos recorrer a diversas técnicas de selecção de variáveis. Serão usadas algumas medidas descritivas - estatística de teste do Qui-Quadrado (QQ)e Informação Mútua (IM)- e, os valores de prova associados aos valores das medidas QQ irão viabilizar uma abordagem inferencial na qual serão consideradas as metodologias de correcção de Bonferroni e de control do FDR (False Discovery Rate) (Duarte, [3], Benjamini, [1], Shaffer, [8]). A análise será realizada usando os dados recolhidos no âmbito dos estudos de adaptação para a população portuguesa da Escala de Sugestionabilidade de Gudjonsson - GSS1 (Pires, [7], [4], [6]) . Esta escala permite avaliar a tendência que algumas pessoas têm para ceder perante questões falaciosas quando entrevistadas. O estudo de análise discriminante irá avaliar a associação entre a sugestionabilidade medida através das respostas binárias ás "cedências" e algumas características demográficas dos inquiridos. Na avaliação dos resultados serão apresentados os erros associados à amostra de treino, amostra teste e obtidos mediante validação cruzada.

Referências

[1] Benjamini, Y. e 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)*, Vol.57, 1, pp. 289-300.

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

[3] Duarte Silva, A. P. (2010). Classificação supervisionada para dados de elevada dimensão. *Livro de Resumos das XVII Jornadas de Classificação e Análise de Dados* (JOCLAD 2010), ISCTE eds., pp.17.

[4] Gudjonsson, G. H. (1997). The Gudjonsson Suggestibility Scales Manual. Hove: Psychology Press.

[5] Marques, A. Sousa Ferreira, A. e Cardoso, M. (2008). Uma proposta de combinação de modelos em Análise Discriminante Discreta . *Estatística - Arte de Explicar o Acaso*, em Oliveira, I. *et al.* (Eds.), Ciência Estatística, Edições SPE, pp. 393-403.

[6] Marques, A. Sousa Ferreira, A. e Cardoso, M. (2009). Resultados de uma Escala de Sugestionabilidade: Classificação em grupos demográficos. *Livro de Resumos das XVI Jornadas de Classificação e Análise de Dados* (JOCLAD 2009), Univ. Algarve, pp.69.

[7] Pires, R., Sousa Ferreira, A., Silva, D. (2010). Adaptação de uma Escala de Sugestionabilidade. *Livro de Resumos das XVII Jornadas de Classificação e Análise de Dados* (JOCLAD 2010), ISCTE eds., pp. 281.

[8] Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Rev. Psychol.* 46, pp.561-584.