SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

Programa e Livro de Resumos

Edições SPE

Ficha Técnica:

Programa e Livro de Resumos

Isabel Pereira, Adelaide Freitas, Cláudia Neves, Eugénio Rocha,
Manuel Scotto, Maria Eduarda Silva, Nélia Silva

Editora: Sociedade Portuguesa de Estatística

Capa: Carina Sousa

Impressão: Instituto Nacional de Estatística

Tiragem: 200 exemplares

ISBN: 978-972-8890-31-5

Depósito Legal: 366448/13

Identificação simultânea de biomarcadores em estudos genéticos de associação: Desafios estatísticos e computacionais

(Comunicação)
Pedro Duarte Silva – *Faculdade de Economia e Gestão e CEGE*
Universidade Católica Portuguesa/Porto
psilva@porto.ucp.pt

Resumo

Em estudos genéticos de associação é comum pesquisar a informação contida em grandes bases genéticas de dados com o objectivo de identificar um pequeno conjunto de marcadores relacionados com alguma doença ou uma característica genética de interesse. A disponibilidade de bases de dados contendo dezenas ou centenas de milhar de pequenas variações genéticas conhecidas como "Single Nucleotide Polymorphisms"(SNPs) tem permitido encontrar algumas dezenas de SNPs associados, ou casualmente ligados, a condições clinicas de risco tais como anomalias cardíacas, diabetes, ou vários tipos de cancro [4].

Em termos estatísticos, o problema central dos estudos de associação genética pode ser formalizado como um problema de selecção de variáveis (SNPs) em modelos de previsão para uma condição clinica de interesse. Este problema tem no entanto característica especificas, nomeadamente devido ao elevado numero de variáveis potencialmente preditivas, que é tipicamente muito superior ao numero de observações disponíveis.

Devido às dificuldades computacionais envolvidas, muitas das metodologias inicialmente tentadas nestes problemas recorrem a técnicas univariadas, tais como o estabelecimento de rankings baseados em correlações marginais [2]. Essas abordagens revelam-se incapazes de identificar combinações de SNPs que só actuam quando agrupadas em conjunto, e mais recentemente várias técnicas multivariadas (ver, por exemplo, [1][3][5][6][7]) tem vindo a ser propostas e aplicadas com sucesso.

Nesta comunicação far-se-á uma revisão de propostas recentes para o problema da identificação simultânea de biomarcadores em estudos genéticos de associação, discutindo-se as vantagens e limitações das principais abordagens propostas, e identificando-se alguns dos desafios em aberto nesta área.

Referências

[1] Ding, Y., Wilkins, D. (2006) *Improving the Performance of SVM-RFE to Select Genes in Microarray Data*, BMC Bioinformatics. 7.Suppl 2: S12.

[2] Foulkes, A.S., (2009). Applied statistical genetics with R. New York: Springer.

[3] Guyon, I., Weston, J., Barnhill, S., Vapnick, V. (2002) *Gene selection for cancer classification using support vector machines*. Machine Learning. 46, 389-422.

[4] Hindorff, L., Sethupathy, P., Junkins, H.A., Ramos, E.M., Metha, J.P., Collins, F.S. Manolio, T.A. (2009) *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*. Proc. Natl. Acad. Sci. USA.

[5] Hoggart, C.J., Whittaker, J.C., De Iorio M., Balding D.J. (2008) *Analysis of all SNPs, in genome-wide and re-sequencing association studies*. PLoS Genetics. 4:e1000130.

[6] Zhang, C.H. (2010) *Nearly unbiased variable selection under minimax concave penalty, Annals of Statistics. 38, 894-942*

[7] Zuber, B., Duarte Silva, A.P., Strimmer, K. (2012) *A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies*, BMC Bioinformatics. 13:284.