Classification of Tissue Samples on the Basis of Microarray Gene-Expression Data

Geoff McLachlan

Department of Mathematics and the Institute of Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

email: gjm@maths.uq.edu.au

Abstract

In the context of cancer diagnosis and treatment, we consider the problem of classifying a relatively small number of tumour tissue samples containing the expression data on very many (possibly thousands) of genes from microarray experiments. For the supervised problem where there are tumour samples of known classification, we discuss the need to correct for the selection bias in assessing the error rate of a prediction rule formed from a small subset of selected genes (Ambroise and McLachlan, 2002). We also consider the unsupervised problem where the aim is to cluster the tumour samples on the basis of the gene expressions. The associated problem of assessing the number of clusters is addressed. Attention is concentrated on the mixture model-based approach called EMMIX-GENE as proposed by McLachlan et al. (2002). Its performance is demonstrated on various microarray data sets available in the bioinformatics literature.

References

Ambroise, C. and McLachlan, G.J. (2002). Selection bias in gene extraction on basis of microarray gene expression data. *Proceedings of the National Academy of Sciences USA* **99**, 6562–6566.

McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.