Feature selection in functional classification of micro-array gene expression data

Mahesan Niranjan
University of Sheffield

The widely used tool used for making statistical inference on data generated by micro-array gene expression experiments is cluster analysis. Genes that cluster together exhibiting co-expression can be said to have functional similarities or regulated by the same mechanism. In some cases, however, the inference problem is formulated as a classification problem – often referred to in the micro-array literature as `class-prediction' problems. Recently many powerful machine learning algorithms, such as Support Vector Machines, have been applied to functional classification of genes from micro-array data. Such techniques operate in high dimensional spaces and impose complex non-linear transformations of the data. In this paper I will discuss feature reduction applied in such problem settings. Being able to select a subset of features, without loss of discrimination ability, reduces the dimensionality of the problem. With reduced feature sets, interpretation of the data is likely to be easier. These ideas will be demonstrated using a public domain dataset of yeast gene expressions.