39. Data-adaptive test statistics for microarray data

Sach Mukherjee

Pattern Analysis & Machine Learning Group
University of Oxford.
www.robots.ox.ac.uk/~sach

An important task in microarray data analysis is the selection of genes which are differentially expressed between different tissue samples, such as healthy and diseased. However, microarray data contain an enormous number of dimensions (genes), and very few samples (arrays), a mismatch which poses fundamental statistical problems for the selection process, which have defied easy resolution. We present a novel approach to the selection of differentially expressed genes in which test statistics are learned from data, using a simple notion of reproducibility in selection results as the learning criterion. Reproducibility, as we define it, can be computed without any knowledge of the 'ground-truth', but takes advantage of certain properties of microarray data to provide an asymptotically valid guide to expected loss under the true data-generating distribution. We are therefore able to indirectly minimize expected loss, and obtain results substantially more robust than conventional methods. We present the results of experiments using both simulated and oligonucleotide array data, comparing our data-adaptive method to two widely-used gene selection procedures.