Classification of Tissue Samples Using Mixture Modeling of Microarray Gene Expression Data

Shili Lin and Roxana Alexandridis

Abstract

Accurate classification of tissue samples is an essential tool in disease diagnosis and treatment. The DNA microarray technology enables disease classification based only on gene expression analysis, without prior biological insights. We present a classification method based on modeling the distribution of the gene expression profile of a test sample as a mixture of distributions, each of which characterizes the levels of gene expression within a class. Class assignment for a test sample is based on the predictive probabilities of class memberships. We believe that this general modeling framework is a flexible scheme for multi-type classification. Since most of the thousands of genes whose expression levels are measured do not contribute to the separation between types of tissue samples, we also explore several measures for gene selection, including T, NPT, BW, NPBW, and a mixture modeling approach based on Markov chain Monte Carlo (MCMC) estimation of parameters. For a classifier based on a gene selection measure, such as the T classifier, the number of genes selected is achieved by cross-validation. The methods are applied to a leukemia dataset; our results are comparable with the best results achieved in a comparative study done by Professor Terry Speed and colleagues.

Keywords: microarray; gene expression; classifier; mixture; EM; MCMC

1 Introduction

DNA microarrays are biotech chips that enable researchers to measure the expression levels of thousands of genes simultaneously; see Schena [15] and The Chipping Forecast [5]. These measurements are obtained by quantifying the hybridization of the mRNA extracted from tissue samples to an array of spotted cDNA (cDNA arrays) or oligonucleotide probes (oligonucleotide arrays) immobilized on the surface of the chip. Details can be found in Schena *et al.* [16] for cDNA arrays and Lockhart *et al.* [9] for oligonucleotide arrays.

After proper image analysis, data processing and normalization (which entails nontrivial efforts, see for example, Dudoit *et al.* [4], Schadt *et al.* [14], Newton *et al.* [11], and Yang *et al.* [17]), a single number, referred to as the level of expression, is obtained for each gene on a microarray.