Exploring the Within and Between Class Correlation Distributions for Tumor Classification

李克昭
中央研究院統計科學研究所

Abstract

To many biomedical researchers, effective tumor classification methods such as the support vector machine often appear like a black box not only because the procedures are complex but also because the required specifications such as the choice of a kernel function suffer from a clear guidance either mathematically or biologically. As commonly observed, samples within the same tumor class tend to be more similar in gene expression than samples from different tumor classes. But can this well-received observation lead to a useful procedure of classification and prediction? To address this issue, we first conceived a statistical framework and derived general conditions that serve as the theoretical foundation which supports the aforementioned empirical observation. Then we constructed a classification procedure that fully utilized the information obtained by comparing the distributions of within-class correlations with between-class correlations via Kullback-Leibler divergence. We compared our approach with many machine-learning techniques by applying to 22 binary- and multi-class gene expression datasets involving human cancers. The results showed that our method performed as efficiently as support vector machine and Naïve Bayesian, and outperformed other learning methods (decision trees, linear discriminate analysis and k-nearest neighbor). In addition, we conducted a simulation study and showed that our method would be more effective if the arriving new samples are subject to the often-encountered baseline shift or increased noise level problems. Our method can be extended for general classification problems when only the similarity scores between samples are available. This talk is based on a joint work with Xuelian Wei.