Bootstrapped resampling of breast cancer microarray data

Lisa McMillan
Glasgow University

Abstract

Classification using microarray data often proceeds by building the classifier on the set of most significant genes, as measured by absolute correlation to the survival vector (e.g., Van 't Veer 2002). Recent work has shown that several gene sets exist within the microarray data that can classify data as accurately as the most highly correlated genes, using bootstrap sampling in alternative dataset composition (Ein-Dor et al., 2005). Accuracy is the predominant metric with which to measure classification performance.

Here, we have used bootstrap sampling in alternative dataset compositions to (a) assess the stability of gene correlations across bootstraps and dataset compositions (b) to characterise the sample population with respect to sample quality and (c) assess whether other performance metrics may be useful.

Our results suggest that the best classifiers are comprised of genes that are stable with respect to correlation; that the data may contain several samples that are noisy, and that sensitivity and NPV are the most consistent indicators of performance within a dataset composition. We discuss these findings in the context of improved variable selection and prediction methods.