Tumor Tissue Classification Using Support Vector Machines and k-Nearest Neighbor Methods

Chris H.Q. Ding
NERSC Division, Lawrence Berkeley National Laboratory
University of California, Berkeley, CA 94720, USA
chqding@lbl.gov FAX: 510-486-5812, Tel: 510-486-6901

The micro-array DNA expression data of leukemia tumor tissues of Golub et al is analized using support vector machines (SVM) and k nearest neighbor (kNN) classification methods. Based on 38 tumor samples of known cancer type, either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML), we build classification systems, and use them to classify 34 additional tumor samples.

We used F-statistic to select 50 genes out of about 7000 genes as variables for the 76 tumor samples. Leave-one-out cross validations are performed on the 38 training tumor samples; both SVM and kNN are accurate: error rates are 1/38 for both. On classifying the 34 unknown tumor samples, error rates for SVM and kNN are 1/34 and 6/34 respectively. SVM performs most accurate as in several other studies.

We will address the issues with gene (variable) selection, data pre-processing, consistency between training samples and test samples (they are obtained from different labs and under different conditions), prediction strength score, and visualization of results.