Building Large-Scale Gene Expression Classification Databases Using Active Learning

Patrick R. Schmid*[,1], Nathan P. Palmer*[,1], Isaac Kohane[2], Bonnie Berger[1,◆]
* Equally contributing first authors
◆ Corresponding author: bab@mit.edu
1 Massachusetts Institute of Technology, Cambridge Massachusetts, USA
2 Harvard Medical School, Boston Massachusetts, USA

The notion of using large, public microarray repositories to create a classification system that can be used to prognosticate patients based on biological samples, or determine the efficacy of a drug, has been a goal since the advent of the first microarray experiment. One of the largest deterrents to using this publicly available data, however, is not only the sheer volume, but also the lack of consistent annotation among them. We overcome the annotation consistency issue by mapping all of the free text associated with each gene expression sample to the Unified Medical Language System (UMLS) ontology. However, manually inspecting all UMLS annotations for each sample is not only time consuming, but also error prone. We show that using an active learning framework that employs both the text (in the form of UMLS concepts) and the gene expression values, we can efficiently build a database relying on only a subset of all available data. Unlike previous methods that focus on a single phenotype, we use a concept enrichment statistic to allow us to efficiently update a heterogeneous database consisting of various tissue and disease types.