Learning Predictors by Integrating Multiple Microarray Datasets

Keywords: microarray – gene expression – prediction – integration – multiple datasets –

Abstract: Our general challenge is to use a patients microarray to predict some important property or phenotype (eg "disease-free survival time, or "ER status) of that patient. Many projects try to learn this predictor from a single labeled dataset – typically one that they collected. This paper explores the challenge of combining this dataset with other datasets, in the hope of producing a more accurate predictor. In particular, we systematically explore (1) various ways to preprocess the raw data, (2) various ways to normalize the resulting data, and (3) whether to use the probeset information, or switch to using gene information. Our empirical results, over two different tasks – predicting a breast cancer patients disease-free survival, and her ER-status – show that one can improve performance by using the data in other datasets, provided you use the proper settings (preprocessing, normalizing, features); we also provide the setting that appear to work best.

Authors:

- Saman Vaisipour vaisipou@cs.ualberta.ca Canada Department of Computing Science
- Russell Greiner rgreiner@ualberta.ca Canada Department of Computing Science
- David Wishart dwishart@ualberta.ca Canada Department of Computing Science
- Meysam Bastani bastani@ualberta.ca Canada Department of Computing Science
- Chun-Nam Yu chunnam@ualberta.ca Canada Department of Computing Science