Understanding Microarray Data through Applying Competent Program Evolution

Moshe Looks

Department of Computer Science and Engineering
Washington University
Saint Louis, MO 63130, USA
moshe@metacog.org

Ben Goertzel, Lucio de Souza Coelho,
Mauricio Mudado, Cassio Pennachin
Biomind LLC
1405 Bernerd Place
Rockville, MD 20851, USA
ben@goertzel.org,
{lucio, mauricio, cassio}@vettalabs.com

Categories and Subject Descriptors

I.2.2 [**Artificial Intelligence**]: Automatic Programming – *Program synthesis*

General Terms

Algorithms, Design, Experimentation

Keywords

Empirical Study, Heuristics, Optimization, Representations

Many researchers have used supervised categorization algorithms such as GP and SVMs, to analyze gene expression microarray data. Overall, the results in this area using SVMs have been stronger than those for GP. However, GP is sometimes preferable to SVMs because of the relative transparency of the models it produces. Studying the GP models themselves can indicate exactly how the classification is being performed, which can lead to biological insights.

We ask here first whether the use of an alternate program evolution technique, MOSES (meta-optimizing semantic evolutionary search) [2], can improve GP's results in this domain (in terms of both accuracy and model simplicity), and second, if MOSES might succeed in providing "important gene" lists with substantial biological relevance. Here we report results for two datasets: (1) distinguishing between types of lymphoma based on gene expression data [4]; and (2) classifying between young and old human brains [3].

Three issues are relevant to any classification approach to microarray analysis: (1) dealing with a huge number of problem variables; (2) dealing with noisy continuous data; (3) avoiding overfitting to the data. We dealt with (1) by selecting the 50 most-differentiating features to use in all experiments, (2) by considering gene expression levels as *Boolean* features determined by median-thresholding (to eliminates concerns regarding noise and scaling), and (3) by using $TP + TN - s/2$ as our fitness function, where s is the number of nodes in the classifier, TP is the number of true positives, and TN is the number of true negatives (i.e., high parsimony pressure). See [2] for details and justification, along with algorithm parameter settings (which were fixed across a variety of experiments). Results are presented in the table for SVMs (from [5] for Lymphoma, from [1] for aging), "standard" GP operating on floating-point gene-expression levels (operator set $\{+, -, *, /, sin, cos, log, exp\}$), and GP and MOSES on Boolean features (operator set *AND*, *OR*, and *NOT*). The "MOSES+Voting" scheme allows the top classifiers in a run a single vote each with ties going to max-prior (based on 10 runs each). GP and MOSES were both allowed $100,000$ fitness evaluations per run - results are for 10 runs of 10-fold cross-validation (i.e. 100 for voting).

Table 1: Avg. test accuracy, 10-fold cross-validation (max-prior = freq. of most common classification).

Technique	Lymphoma	Aging Brain
Max-prior	75.3%	52.9%
SVM	97.5%	95.0%
Standard GP	84.3%	85.8%
Boolean GP	86.9%	91.1%
MOSES	93.5%	95.3%
MOSES+Voting	98.6%	100%

The best models found by MOSES were extremely simple, e.g., `ALDOA OR NOT(GPR18) OR PGAM1` for Lymphoma, and `HBB OR SPON1` for aging brain (98.7% and 100% overall accuracies, respectively). The 5 genes occurring most frequently in the top models for both problems were analyzed based on the gene ontology and all found to be plausible, with some leading to novel biological hypotheses.

1. REFERENCES

[1] B. Goertzel *et al.* Learning comprehensible classification rules from gene expression data using genetic programming and biological ontologies. In *CIBB*, 2006.

[2] M. Looks. *Competent Program Evolution*. PhD thesis, Washington University in St. Louis, 2006.

[3] T. Lu *et al.* Gene regulation and DNA damage in the aging human brain. *Nature*, 2004.

[4] M. A. Shipp *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 2002.

[5] A. Statnikov *et al.* A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 2004.

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
ACM 978-1-59593-697-4/07/0007.