A New Branch and Bound Feature Selection Algorithm

Frank, A.,[1] Geiger, D.[1] and Yakhini, Z.[1,2]
[1] Department of Computer Science, Technion
[2] Agilent Laboratories, Israel

Feature selection is an essential step to enhance correct classification in the presence of many irrelevant features and a small number of samples. For example gene expression data contains thousands of genes per sample, often with only a few dozens of samples. Most genes measured in a DNA microarray assay are irrelevant to the classification task. Identifying the few genes that affect classification is the task of feature selection.

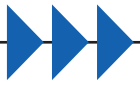
The efficient application of expression profiling as a diagnostic tool [1] is highly dependent on making decisions based on reasonably small numbers of genes. Simple assays will be more cost-effective and much more robust. Enabling decisions based on a small number of parameters will also allow for building redundancy and control measurements into the process. Thus feature selection in the context of expression data may be key to the development of Pharmacogenomics.

We present herein an approach that drastically reduces the number of different feature subsets that need to be evaluated for realistic data including gene expression data. Using bounds on the Bayesian classification error, derived from monotonic additive distance measures such as the *Bhattacharyya distance* [2], our algorithm prunes subsets of features that are no longer candidates for having the lowest error, given the subsets examined so far. When the computational savings are not sufficient, we augment our approach with a preprocessing greedy identification of a sufficiently small subset of the most promising features and use this subset as the input to the main algorithm.

The algorithm we present is a *Branch and Bound* algorithm. This approach can be described as an exploration of a state space tree for the problem at hand. At each point of searching the tree, a bound is computed of the best solution possible in the current subtree. Promising nodes in the tree are expanded, whereas nodes for which the lower bound is larger than the best solution found so far, are pruned.

We tested our algorithm on three gene expression datasets: leukemia data [3], breast cancer [4] and prostate cancer [5]. Our algorithm selected subsets of up to 5 genes from each dataset, usually pruning more than 90% of the subsets in the process. The selection was done from the 100 genes having the highest Bhattacharyya distance. Using the gene subsets selected by our algorithm, a naive Bayesian classifier [2] has shown high classification success rates.

In the leukemia dataset, none of the 5 genes selected by our algorithm, are in the list of 50 dis-



criminating genes used in [3], in which they had 5 misclassifications. We classified the entire test set correctly. In the breast cancer dataset, using a subset of 5 genes selected by our algorithm, we managed to classify correctly all the samples in the test set. Of these 5 genes, the first three appear in the list of 50 genes described in [4] (ranked 1, 12, and 27, respectively), while the last two do not. Perfect classification of the test set was also achieved by this work using all 3389 genes (which are reduced by PCA to 10 components). In the prostate cancer data, of the 5 most common genes from the subsets chosen in a LOOCV experiment for selecting subsets of 3 genes, the first two are amongst the top 100 genes mentioned in [5] (ranked 53 and 1, respectively). Using these 5 genes, our algorithm classified correctly 98% of the samples in a LOOCV test, higher than the any of the models using 4 to 256 genes mentioned in [5].

It is worthy to emphasize that our algorithm selects very small subsets, as small as 3-5 genes, compared to up to 50 genes or more needed in some of other approaches described in the literature. After selecting the gene subsets from microarray data consisting of thousands of genes, it might be possible to use these small subsets of genes for tissue classification employing small-scale low-cost gene expression measurement methods, such as RT-PCR. The development of assays based on profiling small sets of genes is crucial to cost-effectiveness of such methods in clinical practice.

References

[1] Bittner, M., *et al.* (2000), Nature, 406(13), p.536-540.

[2] Fukunaga, K., (1990), Introduction to Statistical Pattern Recognition, Academic Press.

[3] Golub T., *et al.* (1999), Science, 286, p. 531-537.

[4] Gruvberger, S., *et al.* (2001), Cancer Research, 16, p. 5979-5984.

[5] Singh, D., (2002), Cancer Cell, 1(2), p.203-209.