208c An Integer Optimization Framework for Selecting Informative Genes

James Wu and Ioannis (Yannis) P. Androulakis

Microarray experiments are emerging as one of the main driving forces in biology. By allowing the simultaneous monitoring of the expression of an organism's entire genome, array experiments provide tremendous insight into fundamental biological processes. One major challenge is to identify computationally efficient analyses that extract the most informative and unbiased components from the entire microarray data set. Recently, we presented (Androulakis 2005) an approach combining optimization and machine learning techniques (i.e., decision trees) for selecting minimal sets of maximally informative genes. Despite the significant success that we demonstrated, a number of issues remain. Primarily, the implicit definition of the classification model renders the overall solution cumbersome but also makes extensions, such as analysis of robustness and creation of multiple sets of informative features, complicated.

We will discuss our current work towards the development of a generalized framework based on a mixed-integer optimization formulation coupling optimization-based classification models and feature selection. We model the feature selection via the appropriate use of binary variables. Recently, (Street 2005) discussed an interesting optimization model (OC-SEP) for the derivation of multi-category oblique decision trees. The foundation of this algorithm is the formation of an orthogonality-based separation vector, which approximates the overall distance of the elements in each class vector, a given subset of the examples, above or below a given separating hyperplane. Each successive separating surface is the solution of the nonlinear optimization problem (OMCT) (Street 2005). This formulation has the advantage of building, recursively, arbitrary linear separators between the data. The objective attempts to make sure that for each class the majority of the points are either above or below the plane, and the trivial solution is avoided. The advantage of this formalism is the closed form representation of the classification problem renders the problem amenable to extensions. However, a number of limitations exist, and our work's purpose is to address them. Specifically, we seek to address the following issues:

1. The (OMCT/OC-SEP) formulation, as is, does not perform any feature selection,

2. The formulation needs to be solved recursively to construct the classifier and can not be solved as one single problem to build an equivalent classifier,

3. The formulation does not have the ability to construct an ensemble of classifiers in a systematic way,

4. The formulation does not account for modeling complexity,

5. The formulation is highly non-linear, and

6. The formulation is non-convex.

We present our novel extensions of this framework for selecting informative genes as the solution of a single, large-scale integer optimization problem and present our modeling approach for addressing all the aforementioned modeling limitations (items 1-4 in the list mentioned previously). Furthermore, we will:

1. Suggest numerous reformulations aimed at removing some of the critical problems associated with the original formulation (item 5 in the list mentioned previously), and

2. Explore the possibilities offered by global optimization methods already implemented in commercially available packages [e.g., BARON in GAMS (Sahinidis 1996)] to address the issues related to the possible non-convexities (item 6 in the list mentioned above).

We propose possible algorithmic decompositions for solving the large-scale optimization problem and explore distributed branch-and-bound implementations (Androulakis and Floudas 1999) for the solution of this very demanding computational problem. We test our methodology using an extended library of publicly available microarray data sets.

References

Androulakis, I. P. (2005). "Selecting maximally informative genes." Computers & Chemical Engineering 29(3): 535-546.

Androulakis, I. P. and C. A. Floudas (1999). Distributed Branch and Bound Algorithms for Global Optimization. Parallel Processing of Discrete Problems. P. M. Pardalos. New York, Springer-Verlag. 106: 1-36.

Sahinidis, N. V. (1996). "BARON: A general purpose global optimization software package." Journal of Global Optimization 8(2): 201-205.

Street, W. N. (2005). "Oblique multicategory decision trees using nonlinear programming." Informs Journal on Computing 17(1): 25-31.