Data Mining and Mathematical Programming
October 10-13, 2006

Optimization Models for Logical Analysis of Data

Peter L. Hammer
hammer@rutcor.rutgers.edu
RUTCOR
Rutgers University
640 Bartholomew Road
Piscataway, NJ 08854-8003
USA

Abstract

The Logical Analysis of Data (LAD) is a combinatorics and optimization-based methodology for extracting information from a dataset. Given a set of n-vectors, the central problem of LAD is to build a relatively simple algebraic expression of a function, which maps the given $m+$ "positive" and $m-$ "negative" n-vectors (also called "observations") into the values $+1$ and -1, respectively. LAD has been introduced in the late 80s, and has been successfully applied to the solution of numerous problems arising in medicine, biology, biomaterial design, finance, economics, etc.

We shall present a basic outline of the LAD methodology, emphasizing some specific optimization problems occurring in the basic steps of LAD: (1) *support set identification* — i.e. recognition of subsets of variables which are sufficient for recognizing the positive or negative character of any vector in the dataset; (2) *binarization* — i.e. the transformation of the original numerical variables into a larger set of 0-1 variables, preserving the key information content of the original dataset; (3) *pattern generation* — i.e. the enumeration of sub-cubes of the binary cube containing only positive, or only negative vectors; (4) *model formation* — i.e. determination of optimal subsets of patterns "covering" all the given observations in the dataset, and (5) *construction of a discriminant* — i.e. of a nonlinear polynomial in the binarized variables associated to the dataset, the sign of which indicates the positive or negative nature of the binarized form of an arbitrary n-vector.

Discriminants are the basic *classification tool* of LAD, used for recognizing the positive or negative character of new observations. The major medical application of classification is the development of diagnostic systems based on clinical, genetic, proteomic, or other data.

Discriminants represent also the basic *risk assessment tool* of LAD, used for recognizing the risk levels of negative observations to turn into positive ones. The major medical application of risk assessment is the development of *prognostic systems*.

The lecture will be devoted to the presentation and discussion of optimization models occurring at each of the 5 stages of LAD. It will be seen that, besides linear and quadratic optimization problems, many models occurring in LAD are variants of set covering problems, which frequently have quadratic — or more generally, nonlinear — objective functions. Because of the very high number of variables in these problems, the application of exact methods has to be preceded by heuristic procedures aimed at eliminating inessential variables.

Applications will be presented both to benchmark problems publicly available on the Web and to new datasets obtained from collaborating field experts.

Joint work with Tibérius O. Bonates.