Symbolic Discriminant Analysis for Mining Gene Expression Patterns

Jason Moore
Vanderbilt University Medical School
Program in Human Genetics 519 Light Hall
Vanderbilt University Medical School Nashville, TN 37232-0700
USA
615-343-5852
615-343-8619
moore@phg.mc.vanderbilt.edu
CAMDA00 Dataset 2: Leukemia

Jason H. Moore, Joel S. Parker, Lance W. Hahn

Linear discriminant analysis is a popular multivariate statistical approach for classification of observations into groups because the theory is well described and the method is easy to implement and interpret. However, an important limitation is that linear discriminant functions need to be pre-specified. That is, specific variables need to be selected and added linearly into the model. Only the coefficients are estimated from the data. To address this limitation, we developed symbolic discriminant analysis (SDA) for the automatic selection of gene expression variables and discriminant functions that can take any form. Our SDA approach is inspired by the symbolic regression approach of Koza (1992). We begin by defining the mathematical functions (e.g. +, -, /, *, log, sqrt, etc.) and the list of gene expression variables that could potentially be used as the building blocks for discriminant functions. Symbolic discriminant functions are evaluated by generating discriminant scores for each observation to be classified. The overlap in distributions of discriminant scores between groups is an estimate of the classification error. Class membership for new observations can be predicted from the discriminant score that separates the distributions. To identify optimal symbolic discriminant functions from the near infinite model space, we employed parallel genetic programming for machine learning on 4 processors of a 110 processor Beowulf-style parallel supercomputer. We applied the SDA approach to identifying subsets of gene expression variables and symbolic discriminant functions that can correctly classify and predict types of human acute leukemia. Using a leave-one-out cross-validation strategy, we identified two different combinations of gene expression variables and symbolic discriminant functions that correctly classified 38/38 observations in the first dataset and correctly predicted 33/34 observations in the independent dataset. Genes identified in these two models included adipsin, erythroid beta-spectrin, nucloporin 98, and CD33. These are all genes associated with leukemia. We conclude that the SDA approach provides a powerful alternative to traditional multivariate statistical methods for identifying gene expression patterns. The advantages of SDA include the ability to identify an important subset of gene expression variables from among thousands of candidates and the ability to identify the most appropriate mathematical functions relating the gene expression variables to a clinical endpoint. We anticipate this will be an important methodology to add to the repertoire of approaches for mining gene expression patterns.

Keywords

symbolic discriminant analysis, genetic programming, machine learning, supervised pattern recognition

Tools

We will be begin beta testing our software within the next 6 months.

website

http://phg.mc.vanderbilt.edu