Bayesian Variable Selection for Linear Regression in High Dimensional Microarray Data

Wellington Cabrera  
University of Houston  
Houston, TX 77204, USA

Carlos Ordonez  
University of Houston  
Houston, TX 77204, USA

David Sergio Matusevich  
University of Houston  
Houston, TX 77204, USA

Veerabhadran  
Baladandayuthapani  
MD Anderson Cancer Center  
University of Texas  
Houston, TX 77030, USA

ABSTRACT
Variable selection is a fundamental problem in Bayesian statistics whose solution requires exploring a combinatorial search space. We study the solution of variable selection with a well-known MCMC method, which requires thousands of iterations. We present several algorithmic optimizations to accelerate the MCMC method to make it work efficiently inside a database system. Our optimizations include sufficient statistics, variable preselection, hash tables and calling a linear algebra library. We present experiments with very high dimensional microarray data sets to predict cancer survival time. We discuss encouraging findings, identifying specific genes likely to predict the survival time for brain cancer patients. We also show our DBMS-based algorithm is orders of magnitude faster than the R statistical package. Our work shows a DBMS is a promising platform to analyze microarray data.

2. SUMMARY
Linear regression is a statistical model that describes a linear relationship between a scalar dependent variable and a set of explanatory variables (or independent variables). We focus on the case when the number of explanatory variables, p, is very high (thousands) and the number of data points, n, is small (a few hundred). There are two main steps in our algorithm: Variable preselection and an optimized Gibbs sampler. We integrate both steps into the DBMS through User Defined Functions (UDFs). Variable preselection is an effective strategy for dealing with high dimensionality problems [2, 3]. In order to reduce the size of the original problem we process the initial data set, preselecting d out of the p variables that are most correlated to the dependent variable (n < d < p). The resulting reduced data set is loaded into RAM.

The second step of the algorithm is a Gibbs sampler with several optimizations. This is an MCMC method that produces an approximation of the posterior probabilities of the model parameters, based on an informative Zellner’s G prior [4]. We also introduce a prior on the vector of the k selected variables, favoring parsimonious models (1 < k ≪ n). We use the data summarizations described in [6] to accelerate the posterior calculations. This process is further accelerated by noticing that many models are frequently repeated, therefore storing the computed probabilities on a hash table saves a considerable amount of time. Further acceleration is achieved by discarding low frequency variables after the burn in period, that is, variables that appear less than a predetermined number of times in the selected models. We also integrate LAPACK into the DBMS to improve the accuracy and performance of matrix inversions.

Experimental results show consistent marginal probabilities for the most frequent variables across experiments. Our algorithm found some of the top markers (dimensions) that had been previously described in the biomedical literature as important in determining survival time of cancer patients [7, 8]. When comparing performance with the R package, our algorithm is up to 100 times faster, depending on d. Our algorithm performs each iteration in the worst case time complexity T = O(dnk^2) + O(dk^3), but the average case is much better. We have successfully integrated this optimized algorithm with SQL queries and UDFs.
3. REFERENCES


