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.

1. INTRODUCTION

DBMSs act as a repository of biomedical data sets, including gene expression data sets. In this work we address the problem of of Bayesian variable selection for linear regression in microarray data sets using a DBMS. Benefits of data analysis inside a DBMS include fast data access speed, flexible querying and increased data security. Moreover, the data sets used in our project were already stored on a DBMS. Variable selection, the search of best subsets of variables predicting a target variable, is a signficantly hard problem when the number of dimensions is high (thousands). Since the search space is large, a brute force search approach is infeasible. Instead, a promising approach from modern statistics is to solve this problem through a Bayesian approach using Markov Chain Monte Carlo (MCMC) methods [1, 5]. In our study, an optimized algorithm is applied to identify the variable subsets that are best predictors of survival time of brain cancer patients.

November 1, 2013, San Francisco, CA, USA.

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 ($1 < k \ll n$). 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

[1] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[2] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[3] J.Fan and J.Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*, 70:849–911, 2008.

[4] J.-M. Marin and C. P. Robert. *Bayesian core: A practical approach to computational Bayesian statistics*. Springer, 2007.

[5] M. Navas, C. Ordonez, and V. Baladandayuthapani. Fast PCA and bayesian variable selection for large data sets based on SQL and UDFs. In *Proc. ACM KDD Workshop on Large-scale Data Mining: Theory and Applications (LDMTA)*, 2010.

[6] C. Ordonez. Statistical model computation with UDFs. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(12):1752–1765, 2010.

[7] P. Roth, J. Wischhusen, C. Happold, P. A. Chandran, S. Hofer, G. Eisele, M. Weller, and A. Keller. A specific miRNA signature in the peripheral blood of glioblastoma patients. *Journal of neurochemistry*, 118(3):449–457, 2011.

[8] S. Srinivasan, I. R. P. Patric, and K. Somasundaram. A ten-microRNA expression signature predicts survival in glioblastoma. *PLoS One*, 6(3):e17438, 2011.