Improved Feature Selection by Incorporating Gene Similarity into the LASSO

Christopher E. Gillies, Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

Xiaoli Gao, Department of Mathematics and Statistics, Oakland University, Rochester, MI, USA

Nilesh V. Patel, Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

Mohammad-Reza Siadat, Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

George D. Wilson, Radiation Oncology Department and BioBank Department Beaumont Health System, Royal Oak, MI, USA

ABSTRACT

Personalized medicine is customizing treatments to a patient’s genetic profile and has the potential to revolutionize medical practice. An important process used in personalized medicine is gene expression profiling. Analyzing gene expression profiles is difficult, because there are usually few patients and thousands of genes, leading to the curse of dimensionality. To combat this problem, researchers suggest using prior knowledge to enhance feature selection for supervised learning algorithms. The authors propose an enhancement to the LASSO, a shrinkage and selection technique that induces parameter sparsity by penalizing a model’s objective function. Their enhancement gives preference to the selection of genes that are involved in similar biological processes. The authors’ modified LASSO selects similar genes by penalizing interaction terms between genes.

They devise a coordinate descent algorithm to minimize the corresponding objective function. To evaluate their method, the authors created simulation data where they compared their model to the standard LASSO model and an interaction LASSO model. The authors’ model outperformed both the standard and interaction LASSO models in terms of detecting important genes and gene interactions for a reasonable number of training samples. They also demonstrated the performance of their method on a real gene expression data set from lung cancer cell lines.

Keywords: Gene Expression, Gene Ontology, Least Absolute Shrinkage and Selection Operator (LASSO), Regression, Semantic Similarity

DOI: 10.4018/jkdb.2012010101
INTRODUCTION

An important goal of the healthcare industry is personalized medicine, which has the potential to transform healthcare practice. One aspect of the personalized medicine focuses on customizing treatments based on the genetic profile of a patient. In order to achieve this goal, researchers are looking for biomarkers, such as the expression of a gene or group of genes that correlate with treatment outcomes. Biologists use microarrays to read the gene expression of a biospecimen, see (Dubitzky, Granzow, Downes, & Berrar, 2009) for an introduction to microarrays. Some groups are using newer techniques such as RNA-seq to read gene expression (Marioni, Mason, Mane, Stephens, & Gilad, 2008). In this paper, we focus on microarrays; however, we believe our method could be applied in a similar manner to RNA-Seq data. The result of a microarray experiment is a gene expression profile. We represent a gene expression profile by \( x_i = (x_{i1}, \ldots, x_{ip})' \), where \( p \) is the number of genes and \( x_{ij} \) corresponds to the expression of gene, or feature, \( j \). A series of \( n \) microarray experiments yields a matrix \( X \in \mathbb{R}^{n \times p} \) where each row represents a single microarray experiment. In this paper, we assume each row of \( X \) is a biospecimen from a different patient. The gene expression matrix \( X \) can be used for supervised or unsupervised learning. However, we focus exclusively on supervised learning. In supervised learning, we have a vector \( y \in \mathbb{R}^n \) where each \( y_i \) corresponds to a row of \( X \). If \( y_i \) is continuous then this is a regression problem, but if \( y_i \) represents an element from a set of categories or labels then this is called a classification problem. A representation of these notational concepts can be seen.

\[
x_1 = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \quad X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}
\]

A major challenge with gene expression data is \( p \gg n \). In other words, there are many more genes than analyzed biospecimens. Fitting a model with \( p \gg n \) leads to overfitting. When this problem occurs, the model predicts the response variables for training data with high precision, and the response variables for test data with low precision. The small \( n \) and large \( p \) has been studied for a long time in the field of machine learning. There are many approaches to handling the problem, and one approach is feature selection. Sometimes researchers using gene expression data refer to these methods as gene selection techniques, because the features are usually genes. Overall, feature selection techniques are organized into three categories: filter, wrapper, and embedded methods (Saey, Inza, & Larrañaga, 2007). Filters are applied before classification or regression, where they rank genes based on some metric. Genes that fall below some threshold are removed from further analysis. Some filters include: Signal-to-Noise Ratio (SNR) (Golub et al., 1999) and t-statistics (Speed, 2003). Filter methods tend to be dataset specific, so a researcher must be careful not to add bias to the final accuracy. It is usually the case that a research must specify the number of significant genes to include in the model. Wrapper methods attempt to find an optimal subset of genes to achieve high accuracy. These methods have the name wrapper because these techniques wrap around the classification or regression algorithm and call the algorithm as a subroutine. There are two types of wrapper methods: deterministic and stochastic. The deterministic methods incrementally increase or decrease a gene-subset. These methods can be built using forward or backward selection. Deterministic wrappers tend to follow a greedy approach, and they are prone to local optima. Stochastic wrappers use randomization such as genetic algorithms, to combat the issue of converging at a local optimal. Embedded methods incorporate feature selection into the model fitting process. An example of an embedded method is the LASSO (Tibshirani, 1996). In the context of linear regression and least squares regression,