Logistic and LASSO regression models in predicting cancer event with significant analysis of microarray experiments

Chongshu Chen

October 5, 2015

1 Objective

As experimental costs decrease, large scale microarray experiments are becoming increasingly routine, particularly in characterizing the genome wide dynamic regulation of gene expression and time-course information provides valuable insight into the dynamic mechanisms underlying the biological processes being observed. However, a proper statistical analysis of genome data requires the use of more sophisticated tools and complex statistical models. For example, there are problems due to multiple comparisons increased by catering for changing effects. There exists different significance methods for analyzing microarray data to identify differential expressed genes. In this research project, we would like to investigate the logistic predicting model with significant analysis and LASSO penalized regression to detect significance of genes that predict the recurrence of cancer event by using microarray experiments. These significance methods can be applied to the typical types of comparisons and sampling schemes in microarrary data and extend it to more complicated situation in Genomic data analysis. We further evaluate the reliability of proposed methods through using the cross validation techniques.

2 Background

Although Tamoxifen significantly reduces tumor recurrence in certain patients with early stage estrogen receptor positive breast cancer, predictive markers of treatment failure have not been identified. In the paper of "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen", Ma, Wang, eta. generated gene expression profiles of hormone receptor -positive primary breast cancer. They proposed a method that predict the disease-free survival with a two-gene ratio, HOXB13 versus IL17BR, which outperformed existing biomarkers. In 2008, Loi, Haibe-Kains, Desmedt, etl researched on predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. They developed a gene classifier consisting of 181 genes belonging to 13 biological clusters. In the independent set of adjuvantly-treated samples, it was able to define two distinct prognostic groups. Six of the 13 gene clusters represented pathways involved in cell cycle and proliferation. One of the classifier components suggesting a cellular inflammatory mechanism was significantly predictive of response.

3 Description of Data

As studying gene expression profiling of tumors appears to be a promising new strategy for predicting clinical outcome in oncology. The research studies of genome not only become highly valuable to clinician and scientific researchers, but it will also give significant outcomes to patients. As Human Genome Project was completed in 2003, it had estimated that humans have between 20,000 to 25,000 genes. The genome databases become highly available for researchers since the development of genomic technology in 21st

century. Researchers can obtain very rich datasets from various public database resources. These databases are valuable for analyzing the functionality of genes. This project will involved on using data from Gene Expression Omnibus (GEO) dataset with reference series GSE1378. The detail of GDS807 is available in following reference link at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS807. We will download the dataset manually from the NCBI website database. The platform for this data set is GPL1223, a custom-designed spotted human oligonucteotide microarray (designed by Arcturus and fabricated by Agilent). There are gene expression profiles for each of 60 samples of breast cancer tissue (from distinct subjects). The data in the ExpressionSet can be taken to be normalized, in which converted from a two-color array to a single gene expression value. A probe annotation table is available at the GEO entry. There are probeID, controlType and GeneName listed in the dataset. We are only interesting in determine the genes that are experimental. We further create a subset of the ExpressionSet object consisting of all 60 samples, but with only those probes identifies as experimental. Furthermore, we label the featureNames slot of the subsetted ExpressionSet using the appropriate accession numbers from the GDS807 features files.

4 Statistical Methods

4.1 Logistic predicting models

We want to construct a set of biomarkers that will predict recurrence using gene expression levels. Consider the following procedure as fitting a logistic regression model using recurrence/non-recurrence as the outcome and the 5 selected gene expression levels as the predictors. Suppose $x = (x_1, ..., x_k)$ is a vector of gene expression levels for a given subject (here k = 5). We will have the linear predictor term

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1}$$

In a logistic regression model, we assume that the probability of recurrence give gene expression level x is

$$Pr(Y = 1 \mid x) = \phi(\eta).$$

We will evaluate the model by implementing gene selection procedure inside the cross-valuation method. A design of precedure is given as following:
(a) Select columns as the test data, and the remaining data as the training data.
(b) Select the 5 genes according to wilcoxon rank sum test from the training data only.
(c) Evaluate the risk score, the linear predictor term η for subjects making up the test data.
(d) estimate the predictive outcomes under the model for all test set and repeat above steps.

4.2 LASSO models

We further investigate data with a penalized regression method to detect significance of genes that predict the recurrence of cancer event. A simple explanation of the Lasso regression model as giving a set of input measurements $x_1, x_2, ..., x_k$ and an outcome measurement y, the lasso fits a linear model

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{2}$$

subject to minimize $\sum(y_i - \beta^T x_i)^2 + \lambda \sum \mid \beta_j \mid$. It means that the criterion is not only to minimize $\sum(y - \hat{y})^2)$, but also subject to a constraint that $\sum \mid b_k \mid \leq$ B, for some constant value B. The first part of sum is residual sum of squares, which is taken over observations in the dataset. The parameter λ is a tuning parameter. In general, we would include the λ in B. When "B" is large enough, the constraint has no effect and the solution is just the usual multiple linear least squares regression of y on $x_1, x - 2, ...x_k$. When for smaller values of B, some of the coefficients β_j are shrunk to zero. In addition to that, we requires to use cross-validation to estimate the optimal value for λ. Thus, we extend the cross validation procedure with a method of nested cross-validation approach to evaluate robustness of the LASSO model in predicting the disease state for women. The nested cross valuation not only selects the optimal lambda, but also evaluates the accuracy of

prediction as the same time. The basic idea is that we implement a cross validation procedure in an inner loop to choose the optimal λ value, then we make an outer loop function to fit the LASSO model at the optimal λ and validate the error rate with another cross validation procedure. This method would set up strong penalty of a fitting model. We would expect the AUC statistic would be lower than previous cross validation procedure. Therefore, we can measure the true prediction power of LASSO model and determine the "good" genes that have true predicting power.

5 Summary

We will identify the genes that are significantly predicting the occurrence/non-occurrence of cancer events in patients treated with Tamoxifen. We will approach this research objective as a a classification problem. First of all, wee will fit the logistic regression with significant analysis of microarray experiments by pre-selecting the most differential expressed genes and construct the models based on five genes. Then, we will implement the lasso procedure for shrinkage and model selection for shrinking the number of insignificant genes in dataset. Finally, we evaluate the performance the Logistic and Lasso regression models with cross-validation procedures with measurement of AUC (area under curve) statistics. Especially, we will implement the nested cross-validation procedure for Lasso model to evaluate the performance of predictive accuracy on disease state to assess the true performance of predictive model. Hence, we will design the cross-valuation scheme for fitting logistic regression model and fitting the lasso predictive model under optimal tuning parameters to find the significant genes, in which they can be used as a set of classifiers to predict the recurrence of cancer event in patient treated with Tamoxifen.

References

[1] Loi S, Haibe-Kains B, Desmedt C, Wirapati P et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* 9:239

[2] Ma, X., Wang, Z., et al. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 5(6), 607-16.

[3] NCBI GDS807 Dataset http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS807

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.