Developing a Molecular Prognostic Predictor of a Cancer based on a Small Sample

Inyoung Kim*, Sunho Lee†, Sun Young Rha*, and Byungsoo Kim‡

Abstract

One important problem in a cancer microarray study is to identify a set of genes from which a molecular prognostic indicator can be developed. In parallel with this problem is to validate the chosen set of genes. We develop in this note a K-fold cross validation procedure by combining a 'pre-validation' technique and a bootstrap resampling procedure in the Cox regression. The pre-validation technique predicts the microarray predictor of a case without having seen the true class level of the case. It was suggested by Tibshirani and Efron (2002) to avoid the possible over-fitting in the regression in which a microarray based predictor is employed. The bootstrap resampling procedure for the Cox regression was proposed by Sauerbrei and Schumacher (1992) as a means of overcoming the instability of a stepwise selection procedure. We apply this K-fold cross validation to the microarray data of 92 gastric cancers of which the experiment was conducted at Cancer Metastasis Research Center, Yonsei University. We also share some of our experience on the 'false positive' result due to the information leak.

Keywords: Bootstrap, K-fold cross validation, Over-fitting, Pre-validation, Stepwise selection.

*Cancer Metastasis Research Center in Yonsei University
†Department of Applied Mathematics in Sejong University
‡Department of Applied Statistics in Yonsei University

1 Introduction

One of the important applications in the cancer microarray experiment is to develop a molecular prognostic indicator of the cancer when clinical covariates including survival times are available. The first step toward constructing the indicator is to select the optimal subset of informative genes. The next step would be validation of the newly developed indicator in an independent test set. When the clinical covariates including the survival times are available for cancer patients, the Cox proportional hazard regression can be used for screening the informative genes. Once a set of informative genes is chosen, a further step of reducing the number of covariates in the Cox regression needs to be done. A means of reducing the number of covariates, particularly for the small sample, would be the bootstrap model selection which combines the bootstrap method with the stepwise selection procedure (Sauerbrei and Schmacher, 1992).

The performance of the prognostic indicator is best assessed by applying the rule created in the training set to an independent test set. When the sample size is not large enough to have separate training and test sets, the cross-validation is the choice method. Tibshirani and Efron (2002) proposed the 'pre-validation' procedure as a means of comparing the microarray predictor to an existing clinical covariates in the logistic regression. This pre-validation technique can be applied when we develop a prognostic indicator of a cancer based on the cross validation.

2 K-fold cross validation approach

We develop a K-fold cross validation as follows by combining the pre-validation and the bootstrap-model selection;

Step 1: Divide the whole patients into K groups, S_1, S_2,$\cdots$, S_K, with balanced sample sizes and equal proportions of the cancer relapse.

Step 2: Let S_{-1} be the group of whole sample minus the patients in S_1. From S_{-1}, we draw a random sample, with replacement, of same size with S_{-1}. For screening the informative genes we identify in the bootstrap sample individual genes whose expression levels are highly correlated with the relapse time using univariate Cox regressions. With this informative gene set we perform a stepwise selection to further reduce the number of genes.

Step 3: We iterated Step 2 for 100 times.

Step 4 : From 100 bootstrap samples we select genes based on the frequencies of being selected in the final stepwise selections. Let G_1 denote the set of these genes.

Step 5: Estimate a Cox proportional hazard model based on genes in G_1 and calculate risk scores of patients in S_1.

Step 6: Repeat steps 2-5 for $S_{-2},\cdots,S_{-k}$.

All the patients can be sorted and grouped according to their risk scores. The differences of survival curves among groups may validate the newly developed molecular prognostic indicator.

3 Application

We have 7-year follow-up data for 92 cancer patients. These include survival times, stage, TNM status, differentiation, age, sex and curative operation status. For these 92 patients cDNA microarray experiments were performed using a common reference design. The cDNA microarray contains about 17,000 human genes. 54 patients experienced relapse and 38 didn't have relapse. We performed the 9-fold cross validation by following the algorithm described in the previous section. Based on the estimated risk scores we divided the whole patients into 4 groups using quartiles. We conducted a log-rank test to test a difference among survival curves and found there was no evidence that the gene expression profile predicts the survival times of the gastric cancer.

Before we reach this conclusion we had a few trial and errors which we would like to share in this talk. We randomly divided the whole data set into three classes of training and test sets with 2:1 for the sizes of training and test sets. We restricted ourselves to a fixed set of 24 genes which were selected from 92 patients based on the univariate Cox regression. We then fixed this set of genes for the further cross-validation steps. Using these genes we performed the bootstrap-model selection and then estimated risk scores of patients in the test sets. Unlike the previous result we obtained a strong evidence that the gene expression profile predicts the survival time. We believed that this 'false positive' result was due to the information leak at the initial step of fixing the informative genes.

REFERENCES

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11, 2093-2109.

Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, Vol. 1: No. 1, Article 1 (http://www.bepress.com/sagmb/vol1/iss1/art1).