Proceedings of The National Conference
On Undergraduate Research (NCUR) 2018
University of Central Oklahoma
Edmond, Oklahoma
April 5-7, 2018

Discovering Prognostic Genes for Glioma with Elastic Net and Stability Selection

Shuning Jin
Department of Mathematics and Statistics
University of Minnesota Duluth
1049 University Drive
Duluth, Minnesota 55812 USA

Faculty Advisor: Dr. Thierry Chekouo

Abstract

The primary goal of this research is to identify a subset of genes as prognostic markers for glioma, using mRNA expression of 20501 genes in The Cancer Genome Atlas data. The selected genes will help predict risk for individual glioma patients. We use a multivariate Cox proportional hazards model based on both clinical and genetic covariates. After preliminary filtering, elastic net is used for automatic variable selection. Furthermore, stability selection is adopted to improve feature selection stability, combining subsampling with elastic net. As a result, we get a stable feature set of 22 genes. From 10-fold cross-validation, concordance index is 0.864 and stability is 0.962, indicating the algorithm is robust in prediction and consistent in selection.

Keywords: Elastic Net, Stability Selection, Survival Analysis

1. Introduction

Glioma is a common type of brain tumor. Tumors have four grades by World Health Organization (WHO) classification. High grade glioma has poor prognosis, with median overall survival of 3 years for grade III and 15 months for grade IV. An accurate prediction method of survival can help with personalized treatment. Discovering genomic based biomarkers for glioma prognosis is an active research field.

A main challenge in gene analysis is that the data usually consists of tens of thousands of genes yet only hundreds of patients. This leads to the high dimensional problem where the number of predictors is much bigger than the number of observations, causing overfitting and high variance. Hence, we want to identify a small subset of genes that are most relevant to glioma prognosis, to improve prediction accuracy and interpretability of the model. Least absolute shrinkage operator (lasso) is a powerful method to perform continuous shrinkage and automatic variable selection with L1 penalty[1]. Furthermore, elastic net uses a combination of L1 (lasso) and L2 (ridge) regularization, and takes the advantage of both methods to achieve sparsity and prediction accuracy[2].

In addition to predictive power, another concern is the stability of feature selection (FS), defined as the sensitivity of FS algorithms to variations in the training set. Selecting a stable set of features is significant to the reliability and interpretability of a model. If small perturbations in training samples result in very different feature sets, the resulting genes cannot be established as reliable prognostic markers. Meinshausen and Bühlmann proposed stability selection method, which combines subsampling with FS to improve stability and performance of existing methods[3]. When applying FS on multiple subsamples, selection frequency of each gene is counted and genes with high selection probability are selected. A similar idea has been presented in bootstrap-enhanced lasso (bolasso), which uses bootstrapping and takes features selected by all bootstrap samples or at least 90% samples in a soft version[4]. While those studies originally focus on lasso, the frameworks can be extended to other FS algorithms.

In this research, we use a multivariate Cox proportional hazards model to analyze the relationship between survival

time and risk factors including both clinical and genetic covariates[5]. For gene selection, we adapt stability selection for elastic net. The original approach subsamples only half of the data, which may result in a significant loss of information. Thus, we choose to subsample 90% data. The objective of our work is two-fold: (i) to identify a stable set of genes related to glioma prognosis, and (ii) to produce a model that can predict relative change in risk based on simultaneous effects of important covariates.

In the rest of the paper, we describe the data used in the project in Section 2, and then illustrate our methodology for survival analysis and variable selection in Section 3. The experimental results and discussions are presented in Section 4. We end up with conclusions in Section 5.

2. Data

The glioma data of The Cancer Genome Atlas (TCGA) project is downloaded from the Broad Institute GDAC Firehose via *TCGA2STAT* package in R[6]. The dataset includes clinical and genetic data of 1110 patients from two subtypes: Lower Grade Glioma (LGG) and Glioblastoma Multiforme (GBM), corresponding to grade II, III and grade IV respectively. RNA-Sequencing data from the second analysis pipeline (RNASeqV2) is used, represented by upper quartile normalized RSEM values of 20501 genes. For clinical data, survival time, censoring status, age, gender, and histological stage are used. Only 674 samples have both clinical data and gene expression data.

In preprocessing, eight samples are removed for the following reasons. While histological stage consists of 6 stages, 2 stages are removed as each case has only 1 sample. The 4 remaining stages are astrocytoma (A), oligoastrocytoma (OA), oligodendroglioma (O), and untreated primary gbm (G). Also invalid or missing values in survival time or age are removed. As a result, 666 samples are retained for the following analysis. In addition, clinical variables are transformed. Survival time (time to death or last follow-up) is converted from days to months. Since gender and histological stage are qualitative, they are encoded using indicator variables, with female and astrocytoma serving as referents.

3. Methodology

3.1 Survival Analysis

To analyze the relationship between survival time and covariates, we adopt Cox proportional hazards model:

$$h(t|\boldsymbol{x}_i) = h_0(t)\, exp(\boldsymbol{x}_i{}^T \beta)$$

where $h(t|\boldsymbol{x}_i)$ is the overall risk or hazard at time t for an individual i with risk vector $\boldsymbol{x}_i$, and $h_0(t)$ is a baseline hazard rate. For a single covariate k, the hazard ratio (HR) reveals its effect on overall risk, defined as $HR_k = exp(\beta_k)$. HR > 1 indicates a positive effect, HR < 1 indicates a negative effect, and HR = 1 indicates no effect.

3.2 Variable Selection

As a preliminary step, 4000 genes with top variance are selected, and genes with small variability are filtered out. Also gene expressions are $log_2(x + 1)$ transformed to reduce skewness in data. It is followed by univariate screening, and multivariate selection with elastic net and stability selection.

3.2.1 univariate selection

By fitting the univariate Cox model, we derive the HR and p-value from Wald test for each gene. While HR indicates biological significance, p-value indicates statistical significance. The significant genes are selected based on these two criteria. In addition, the p-values are adjusted by Benjamini-Hochberg procedure to control false discovery rate[7]. The selection thresholds are: HR > 2 or HR < 1/2, and adjusted p-value < 0.01.

3.2.2 multivariate selection

We adopt a multivariate Cox model with elastic net penalty using all gene expressions as covariates, and the overall survival time as the dependent variable. Elastic net is a generalization of L1 and L2 regularization, characterized by a convex penalty:

$$\lambda(\alpha|\beta|_1 + (1 - \alpha)|\beta|^2)$$

We use the efficient implementation of elastic net in *glmnet* package in R[8]. There are two tuning parameters. α is the compromise between lasso and ridge, which is set to 0.8 here. λ controls regularization strength, and is determined dynamically by a nested loop of 5-fold cross-validation.

While many models are based on either clinical data or genetic data, combining both information is promising to improve predictive power. In our model, the explanatory variables include age, histological stage (with 4 levels), gender (with 2 levels) and genetic variables selected from the previous procedure. Especially, age and histological stage are closely related with survival time. We adopt the 'favoring' strategy proposed by De Bin et al.[9], where clinical variables are unpenalized and hence are always retained by the model. All genetic variables are equally penalized.

3.2.3 stability selection

Stability selection is a general framework using subsampling to improve stability and prediction of a given variable selection algorithm. The basic idea is to apply the algorithm to multiple subsamples of original data, and genes with high selection probability are chosen as important features.

Formally, the maximum likelihood estimator of selection probability for gene k is derived as $\hat{\pi}_k = \frac{a_k}{N}$, where N is the number of subsamples and a_k is the number of times when k is selected by the algorithm. The set of stable variables is defined as $\hat{S}_{stable} = \{k: \hat{\pi}_k \geq \pi_{thr}\}$, where π_{thr} is the threshold probability.

Here, we adapt stability selection for elastic net. We use the strategy of random subsampling 90% of training data, and create $N = 100$ replicates. The cutoff π_{thr} is set as 0.5. For downstream prediction, we fit the resulting stable genes and clinical variables into a multivariate Cox model.

3.3 Evaluation Metrics

The algorithms are evaluated based on two criteria: stability of feature selection procedure, and prediction accuracy of the resulting model.

3.3.1 stability

Stability across multiple feature sets $\mathcal{A}$ can be quantified as the average of pairwise similarity[10]:

$$\hat{\phi}(\mathcal{A}) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \phi(s_i, s_j)$$

where M is the total number of sets, and ϕ is pairwise similarity.

Specifically, Jaccard index is used as a measure of similarity between two sets:

$$\phi_{Jaccard}(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|}$$

3.3.2 prediction accuracy

The concordance index (c-index) is a standard measure of prediction performance in survival analysis based on rank correlation, defined as the proportion of concordants[11]:

$$c\text{-}index = Pr(concordant) = \frac{concordant}{comparable\ pairs}$$

A pair of observations is concordant if the shorter observed survival time of the two has the larger predicted risk. The comparable pairs are all possible pairs excluding incomparable pairs, where two uncensored observations have tied survival time, or the censored observation is shorter than the uncensored one.

4. Result and Discussion

We perform data analysis on the TCGA glioma dataset using the statistical software R.

4.1 Evaluation

To evaluate performance of different methods, we derive c-index and stability by 10-fold cross-validation. The stability is calculated across *M = 10* feature sets obtained from each training fold. The results are shown in Table 1.

Our primary method is stability selection with both clinical and genetic variables (STAB-CG), which has c-index of 0.864 and stability of 0.962. For comparison, we also evaluate three other methods: unpenalized Cox regression with clinical variables only (COX-CLIN), elastic net with genes only (EN-GENE), and elastic net with both clinical and genetic variables (EN-CG). All the methods are based on a multivariate Cox model.

EN-CG has higher c-index than both COX-CLIN and EN-GENE. This indicates that combining clinical and genetic variables is beneficial for prediction. Compared to EN-CG, STAB-CG has a slightly higher c-index and significantly higher stability. Enhancing elastic net with stability selection increases prediction accuracy and improves feature selection stability greatly, resulting in a highly consistent and interpretable model.

Table 1. Performance of four methods in 10-fold cross-validation. c-index is represented by mean, with standard deviation in parenthesis.

Method	c-index	Stability
COX-CLIN	0.835 (0.030)	NA
EN-GENE	0.844 (0.033)	0.649
EN-CG	0.854 (0.035)	0.535
STAB-CG	**0.864** (0.034)	**0.962**

4.2 Final Model

We derive the final gene set and model by applying stability selection to the whole dataset.

4.2.1 univariate selection

With 4000 input genes after variance filtering, univariate selection results in 1266 genes, with HR > 2 or HR < 1/2 and adjusted p-value < 0.01. The result is visualized as volcano plot (Figure 1), represented by $-\log_{10}$(adjusted p-value) and $\log_2$(HR).

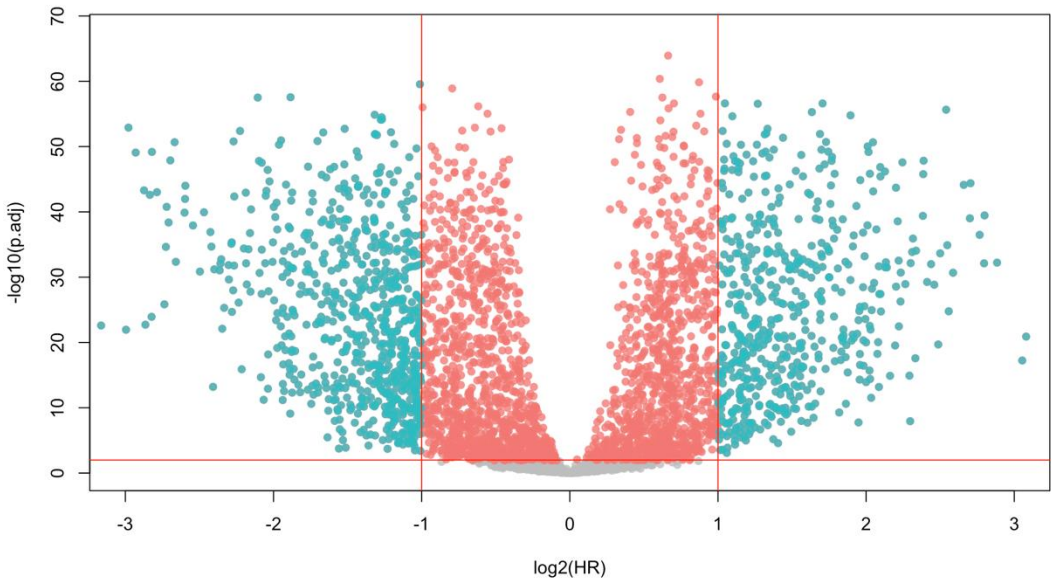

Figure 1. Volcano plot for univariate selection results. The log transformed cutoffs are $|x| \geq 1$ and $y \geq 2$, shown as red lines. The green points are the 1266 selected genes.

4.2.2 stability selection

Elastic net with stability selection results in 22 genes, with selection probability $\hat{\pi} \geq 0.5$. Since the estimated stability of this method is as high as 0.96, the final feature set is robust and reliable. Eventually, we fit a multivariate Cox model with clinical variables and selected genes as covariates. By global likelihood ratio test, the p-value $\leq 2 \times 10^{-16}$.

The statistics of clinical variables and selected genes are summarized in Table 2. Three genes have notably high selection probability: SMC4 (1.0), AGPAT12 (0.96), and ERRFI1 (0.93). The estimated hazard ratios reveal that, SMC4 and ERRFI1 are positively related to risk, and AGPAT12 is negatively related to risk. We conjecture genes with high selection frequency are closely related to glioma prognosis. However, the biological significance of these genes needs more rigorous scientific investigation.

Table 2. Feature set and corresponding statistics. $\hat{\pi}$ is the selection probability in 100 subsamples. The 95% confidence interval (CI) is based on hazard ratio. Gender has 2 levels: male and female, with female as the referent. Stage has four levels: astrocytoma (A), oligoastrocytoma (OA), oligodendroglioma (O), and untreated primary gbm (G), with A as the referent.

Covariate	$\hat{\pi}$	HR	95% CI
Age	NA	1.033	(1.021, 1.046)
Gender (male)	NA	1.267	(0.940, 1.707)
Stage (OA)	NA	0.913	(0.552, 1.509)
Stage (O)	NA	0.767	(0.489, 1.201)
Stage (G)	NA	0.915	(0.584, 1.434)
SMC4	**1.00**	1.218	(0.975, 1.521)
AGPAT12	**0.96**	0.625	(0.455, 0.858)
ERRFI1	**0.93**	1.117	(0.961, 1.298)
AGPAT1	0.87	0.497	(0.287, 0.860)
TTYH3	0.85	1.143	(0.909, 1.436)
GNS	0.72	0.899	(0.651, 1.241)
DDX1	0.72	0.691	(0.482, 0.991)
SMARCB1	0.71	0.980	(0.612, 1.570)
BID	0.70	0.848	(0.631, 1.139)

PTGFRN	0.68	1.048	(0.876, 1.254)
NR1D2	0.67	0.758	(0.593, 0.970)
TXN2	0.65	0.706	(0.462, 1.078)
TCF25	0.63	0.730	(0.431, 1.237)
ARL6IP1	0.62	2.015	(1.397, 2.907)
SFRS5	0.60	0.750	(0.561, 1.004)
PPP2R5C	0.58	0.918	(0.593, 1.422)
NFIB	0.57	0.832	(0.678, 1.022)
SHISA5	0.56	1.364	(0.986, 1.889)
HSPB1	0.54	0.975	(0.782, 1.216)
TM7SF2	0.54	0.882	(0.711, 1.093)
CPD	0.52	1.119	(0.849, 1.475)
TGFBR1	0.52	1.294	(0.981, 1.705)

4.3 subsample number in stability selection

We are interested in investigating how different subsample number N in stability selection affects the stability and c-index. We examine $N \in \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ in 10-fold cross-validation. $N = 0$ (baseline) is equivalent to elastic net without stability selection, and $N = 100$ is used in our aforementioned method. The results are shown in Figure 2.

In general, as N increases, c-index and stability increase. N has a minor effect on c-index, but a great effect on stability. Using large N is computationally expensive in practice. Yet the figure shows precipitous increase occurs at the early stage. $N = 20$ leads to stability of 0.893, which is 67% higher than that of the baseline. Hence, it is feasible to use stability selection to improve variable selection at a relatively small computational cost. Additionally, the computation of independent subsamples can be greatly facilitated by parallel processing.

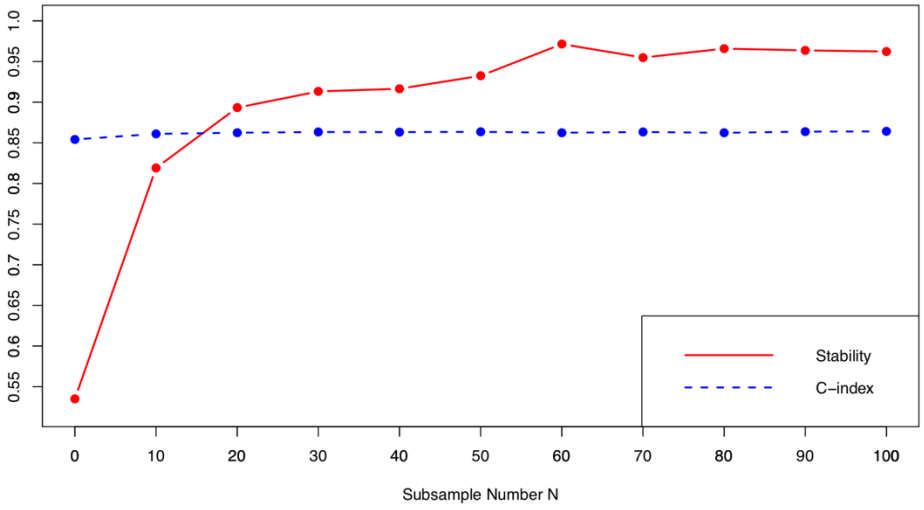

Figure 2. Effect of subsample number N on stability and c-index.

5. Conclusion

By comparing different methods, we find that using genetic and clinical covariates jointly is beneficial for risk prediction. Also, stability selection can greatly improve stability and interpretability of elastic net. Efficient improvement can be achieved with relatively few subsamples. Using elastic net and stability selection, we discover a stable set of 22 genes as prognostic biomarkers. The algorithm has stability of 0.962 and c-index of 0.864 in 10-fold cross-validation, evidencing that the resulting gene set is reliable and the model is robust in prediction. However, the

biological roles of these genes in glioma pathogenesis need further investigation.

For future work, we want to look into the correlations between genes and investigate the grouping effect of this selection procedure. Also, we are interested in examining different settings of hyper-parameters, such as weight α in elastic net, cutoff probability π_{thr} and subsample size in stability selection.

6. Acknowledgements

The author is grateful for the advisement of Dr. Thierry Chekouo. The research is supported by Undergraduate Research Opportunities Program of the University of Minnesota.

7. References

1. Rob Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B* 58, no. 1 (1996): 267.

2. Hui Zou and Trevor Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series B* 67, no. 2 (2005): 301.

3. Nicolai Meinshausen and Peter Bühlmann, "Stability Selection," *Journal of the Royal Statistical Society Series B* 72, no. 4 (2010): 417.

4. Francis R. Bach, "Bolasso: Model Consistent Lasso Estimation through the Bootstrap" (presentation, International Conference on Machine Learning, Helsinki, Finland, July 5-9, 2008).

5. David R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society Series B* 34, no. 2 (1972): 187.

6. Ying-Wooi Wan, Genevera I. Allen, and Zhandong Liu, "TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R," *Bioinformatics* 32, no. 6 (2016): 952.

7. Yoav Benjamini and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B* 57, no. 1 (1995): 289.

8. Jerome Friedman, Trevor Hastie, and Rob Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software* 33, no. 1 (2010): 1.

9. Riccardo De Bin, Willi Sauerbrei, and Anne-Laure Boulesteix, "Investigating the Prediction Ability of Survival Models Based on Both Clinical and Omics data: Two Case Studies," *Statistics in Medicine* 33 (2014): 5316.

10. Sarah Nogueira and Gavin Brown, "Measuring the Stability of Feature Selection" (presentation, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Riva del Garda, Italy, September 19-23, 2016).

11. Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark, "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors", *Statistics in Medicine* 15 (1996): 370.