CS 229 Project Final Report, Autumn Quarter, 2012-2013

Analysis for Association between Genomic Information and Survival Rate of Glioblastoma Multiforme

Yingying Bi, Weiruo Zhang, Jieying Luo

1. Introduction

In order to combat cancer, more and more efforts have been made in genome studies, as cancer is likely to be caused by errors in DNA that let the cells grow uncontrolled. Glioblastoma multiforme (GBM) is the most aggressive malignant brain tumor, and it has a very low survival rate of about one year [1]. It is significant, if some relationship can be found between the survival rate of GBM patients and the genome information. The Cancer Genome Atlas (TCGA) project collects tumor tissue donated by cancer patients. The TCGA data contains clinical information, genomic characterization data and high-throughput sequencing analysis of the tumor genomes [2].

In this project, we make use of TCGA gene expression data and clinical data. The TCGA microarray gene expression data contains 538 samples, and each contains 22278 genes indicating gene expression changes. In the clinical data, there are also 538 samples, and each sample gives the survival days and censoring information of a patient. 421 samples out of 538 samples are deceased patients.

We use the gene expression data as features and formulate the problem such that we can predict the survival time of a patient. This problem is challenging in the following aspects [3]. Firstly, sample size is relatively small, which makes it difficult to identify genes associated with survival from thousands of genes exhibiting expression changes. Secondly, due to the nature of high-throughput microarray technologies, many gene expressions may be random noises, which make them poor indicators. Thirdly, different subtypes of cancer may exist, which adds further complexities to identify the important gene signatures [4].

In the current literatures, several computational methods to approach this problem have been reported. Firstly, feature selection methods are applied to select important indicators from the gene expression data. For example, Bair et al. used Cox proportional hazards model to select genes [4], and Zhang J. et al. made use of differential gene expression clustering to select differentially expressed gene indicators [3]. Secondly, after the feature selection, computational models such as SVM, supervised principal components, partial least squares, various regression methods etc. are used to predict the survival.

In our project, we followed the standard approach in the current literatures. In the preliminary step, we selected the high variance genes out the 22278 genes. We first applied feature selection on those high-variance genes and then tried different learning methods, including SVM and Cox regression.

2. Methods and Results

1) SVM

SVMs are capable of dealing with a large number of input variables with a slight increase in computation complexity, and thus it is suitable for the analysis of high dimensional gene expression microarray data. They can also integrate classification and feature selection in a single consistent framework. Therefore, linear SVM is chosen as the classification algorithm.

We started approaching this problem with a simple two-classes classification model. According to biomedical statistics, most GBM patients die within one year. Therefore, we started by dividing the patient samples into long survival if they have days-to-death greater than 365 days and short survival if they have days-to-death smaller than 365 days. We have noted that there are only 12 samples are within the range of 355-375 days which may be problematic as they are close to the threshold.

(A) Feature Selection

We use feature-ranking techniques to perform feature selection. A fixed number of top ranked features can be selected for further analysis or to design a classifier. We apply the SVM method of recursive feature elimination for gene selection. The procedure is implemented as follows [5]:

```
(1) Start: ranked feature set $R = [\,]$; selected feature subset $s = [1, \cdots, d]$;
(2) Repeat until all features are ranked:
    a) Train a linear SVM with features in set $S$ as input variables;
    b) Compute the ranking scores for features in set $S$ with certain ranking criterion $c_i$
    c) Find the feature with the smallest ranking score: $e = arg\min_i c_i$;
    d) Update: $R = [e, R], S = S - [e]$ ;
(3) Output: Ranked feature list $R$.
```

Two ranking criteria are applied.

Criterion 1 (SVM-RFE)[5]: $c_i = w_i^2$. This objective function in linear SVM is $J = \frac{1}{2}||w||^2$. Using this criterion of w_i^2 corresponds to removing the features whose removal changes the objective function the least.

Criterion 2 (R-SVM)[6]: $c_i = w_i(\mu_i^+ - \mu_i^-)$. μ_i^+ and μ_i^- are the mean gene expression values in each class (long and short survival). The contribution of each gene can be assessed according to its contribution in separating the two classes. The difference of the two class means in the decision function is: $S = \sum_{i=1}^n w_i\mu_i^+ - \sum_{i=1}^n w_i\mu_i^- = \sum_{i=1}^n w_i(\mu_i^+ - \mu_i^-)$. The criterion for ranking the genes is $r_i = w_i(\mu_i^+ - \mu_i^-)$.

(B) Two Classes SVM

Two rankings of genes are obtained using Criterion 1 and Criterion 2. For each ranking, two versions of cross validation (CV) have been implemented to design the optimal classifier and find the most correlated gene subsets, namely, k-fold cross validation (k=10) and hold-out cross validation (70% for training and 30% for testing). To reduce variability of CV estimate, the hold-out CV is run multiple times and an average is computed. To testify our cross validation, we also run the same process with randomized survival data and CV errors are around 0.5 as expected. The results are summarized in Table I, Figure 1 and Figure 2.

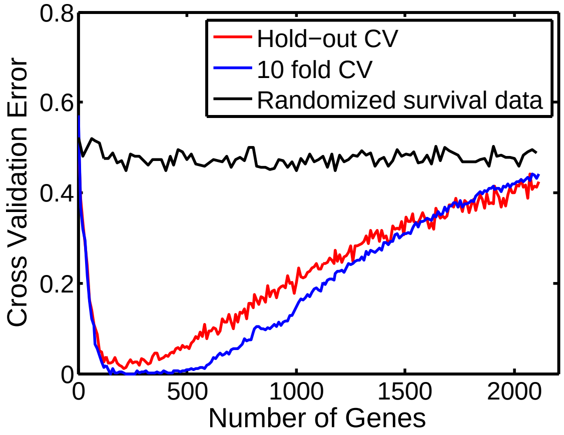


Figure 1: Cross validation results using Criterion 1.

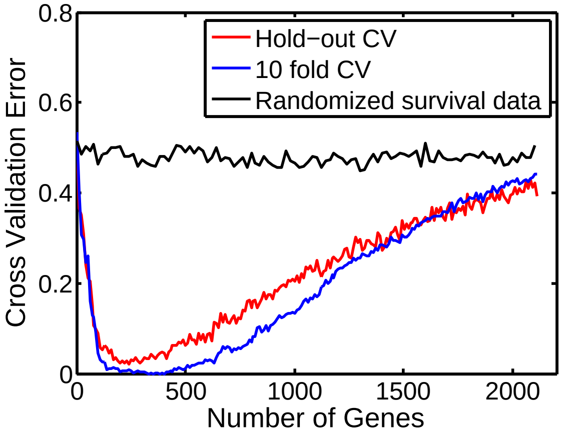


Figure 2: Cross validation results using Criterion 2.

Table I: Comparison of Different Feature Selection Methods

	All high variance genes	Criterion 1	Criterion 2
Number of genes	2113	211	241
Cross validation error	0.4323	0.0110	0.0213

Criterion 1 produced an optimal subset of 211 genes, while Criterion 2 produced an optimal subset of 241 genes. They have 143 genes in common. We find that cross-validation prediction performances of R-SVM and SVM-RFE are nearly the same, and they both have very small CV errors, which demonstrates that the feature selection algorithms worked well. Criterion 1 can offer even smaller CV error.

(C) Multiple Classes SVM

To make the prediction more clinically meaningful, multiple classes SVM is also investigated. The distribution of survival days in our dataset is shown in Fig. 3. We divide the survival days into several classes such that each class has a certain number of data and different classes are as separated as possible.

Three classes ([0 250), [250, 525), [525, $+\infty$)) and four classes ([0 195), [195, 525), [525, 725), [725, $+\infty$)) are studied. SVM is used and the feature selection method is the same as in section 2.1.(A). As the prediction becomes more complicated, the cross validation error increases with multiple classes (Fig. 4). In the case of three classes, the minimum cross validation error is 0.135 with 161 features, which is an acceptable accuracy.

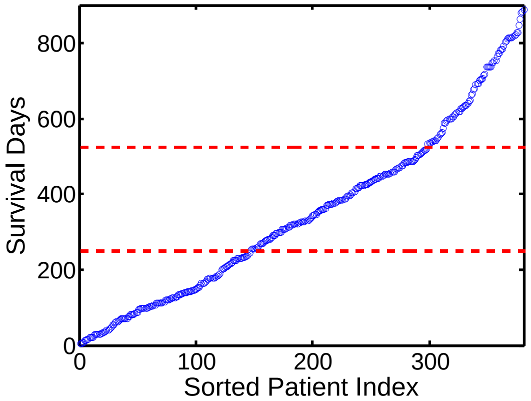


Figure 3: Distribution of patients' survival days (sorted). Dashed red line denotes classification for three classes.

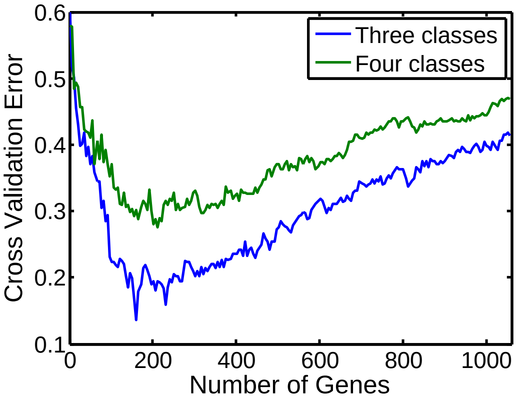


Figure 4: Cross validation results of three classes and four classes.

2) Clustering

Because GBM may have different subtypes, simply classifying the patients according to the survival days may not be biologically meaningful [4]. Different subtypes may have different gene signatures but share wide overlap region in the survival days. Classifying patients based on the survival days solely cannot differentiate subtypes and may result in inaccurate predictions.

We use "k-means" clustering method to first cluster the data into four clusters, and then apply SVM with multiple classes. To evaluate the quality of clustering, the Silhouette value is calculated (Fig. 5). The average silhouette value is 0.1640 , which is comparable to the reported value for clustering gene expression data (0.18 in [7]). Each cluster is trained separately using SVM with four classes. The performance of prediction is greatly improved compared to the case of multiple-class SVM without clustering (Fig. 6). Also the genes selected for each cluster are different, which is an indication of potential different subtypes.

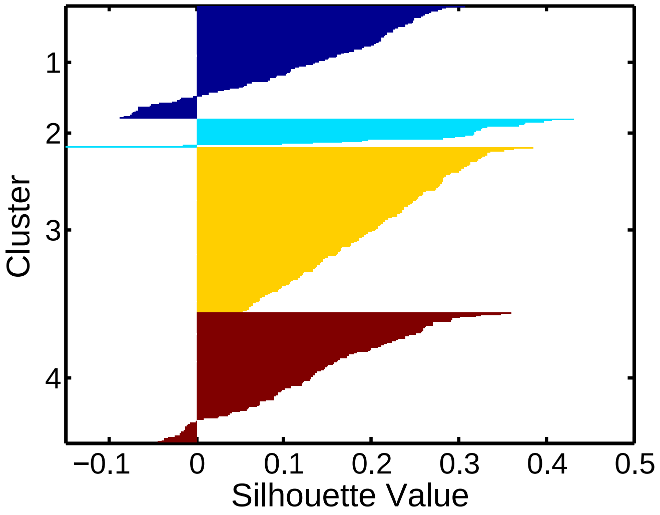


Figure 5: Silhouette value of four clusters.

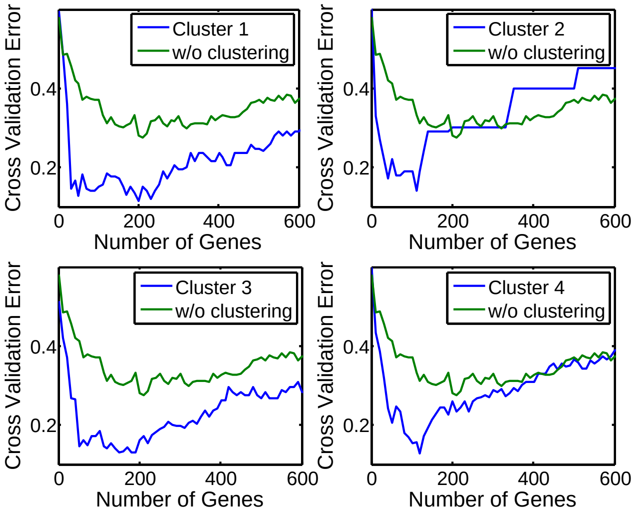


Figure 6: Comparison of CV w/ and w/o clustering.

3) Cox Proportional Hazard Model

One issue with SVM model is that it does not take the censoring information. If a patient is still living, we do not have their number of days to death. Thus, those samples are not included in the SVM prediction. To solve this problem, we use Cox proportional hazard model to take into account the information of patients who are still living and turned it into an inferential problem.

Let $t = (t_1, t_2, \cdots t_n)$ be the times of observation of the n samples, and in the regression, x_i is a vector of gene expressions as predictors$(x_{i,1}, x_{i,2}, \cdots x_{i,n})$ and δ_i is the status indicator (δ_i is 1 if the patient is dead and δ_i is 0 if right-censored) [8]. The Cox model assumes $h_i(t) = h_0(t)\exp(x_i^T\beta)$, where $h_i(t)$ is the hazard for patient i at time $t, h_0(t)$ is the baseline hazard, and β is a fixed length vector. Therefore, we can make the inference using the partial likelihood:

$$L(\beta) = \prod_{i=1}^{n}(\exp(x_{j(i)}^T\beta) / \sum_{j \in R_i} \exp(x_j^T\beta))$$

where R_i is the set of indices j. Using ℓ_2 penalties to maximize the partial likelihood, we can obtain the value for β. We use the R package "glmnet", implemented by Simon N et al. [8], which reduces the partial likelihood maximization problem into a repeatedly solving penalized weighted least squares problem:

$$\beta = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} w(\eta)_i (z(\eta)_i - x_i^T\beta)^2 + \lambda(1-\alpha)\beta_k$$

10-fold cross-validation is used to get the minimum λ as shown in Figure 7.

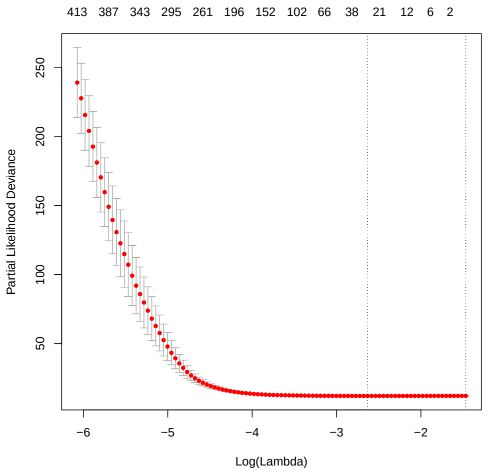


Figure 7: Calculated λ. Each dot represents a λ value, and the left vertical bar indicates the minimum CV error. The top of the plot is the number of coefficients.

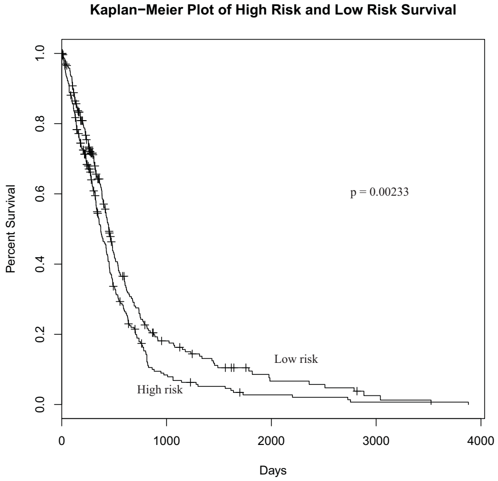


Figure 8: Kalpan-Meier curve of high risk survival rate and low risk survival rate. p $\leq$ 0.05 is biologically significant.

After obtaining β, we use Kaplan-Meier curves in the cross validation [9], as shown in Figure 8. Let b denotes the vector of regression coefficients and x denotes the vector of gene expressions. If $b_i = 0$, then the gene i is not included in the prediction. For k-fold cross validation, $b^{(k)}$ is the vector of regression coefficients based on training set $(D - D_k)$. If a sample j in testing set D_k has predictive index $b^{(k)}x_j$ which is larger than the median of $b^{(k)}x_i$ of the training set, then the sample j in testing set is assigned to higher risk group, otherwise, it is assigned to lower risk group. Higher risk means lower survival rate for the samples in the group.

We calculated the p value to test the statistical significance between the high risk KM curve and low risk KM curve using R package "survival" survdiff function. We obtained p value of 0.00233 (p = 0.05 as the significance) which indicates that the two groups are differentiated and the predictions for survival is significant.

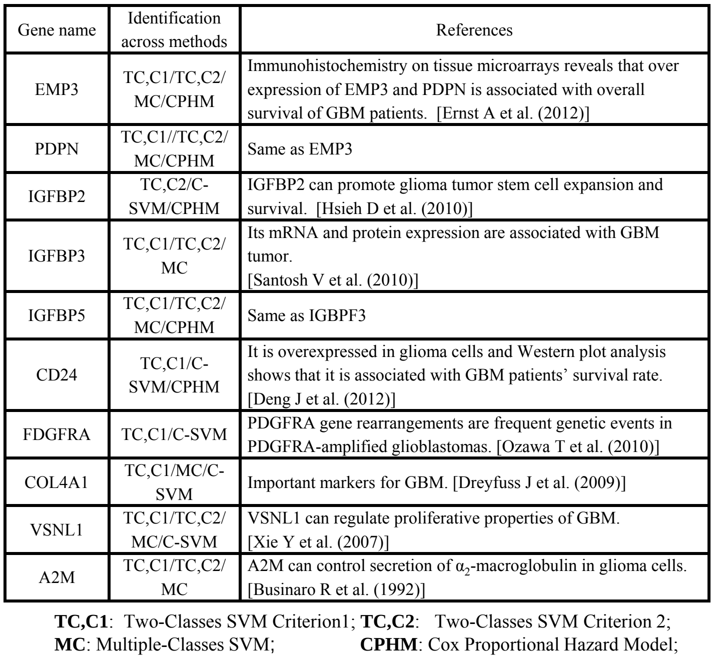
3. Discussion

We have tried two models to approach this problem, namely classification and inference. In the classification model, two-classes model and multiple-classes model can both produce reasonable cross-validation errors. Multiple-classes model has higher errors, but it is more accurate in association with survival in the sense that it separates the classes of samples more than only two-classes. By using clustering to first divide the samples, the results can be further improved. We find several important genes which have higher ranking in the clustering result than that of the multiple-classes SVM result. For example, IGFBP3 ranked 200 in the multiple-classes SVM result, but it is ranked 6^{th} in cluster 1. This further indicates that GBM may have several different subtypes, and each may have different set of gene signatures.

In the inference model, Cox proportional hazard model makes use of the censoring data and finds fewer gene signatures than that of classification model. By comparing the inferential results of high risk survival and low risk survival, it is clearly that the two groups are differentiated.

We compare the results of the three methods, and some interesting genes are identified as shown in Table II.

Table II: Important Identified Gene Signatures



Gene name	Identification across methods	References
EMP3	TC,C1/TC,C2/ MC/CPHM	Immunohistochemistry on tissue microarrays reveals that over expression of EMP3 and PDPN is associated with overall survival of GBM patients. [Ernst A et al. (2012)]
PDPN	TC,C1//TC,C2/ MC/CPHM	Same as EMP3
IGFBP2	TC,C2/C-SVM/CPHM	IGFBP2 can promote glioma tumor stem cell expansion and survival. [Hsieh D et al. (2010)]
IGFBP3	TC,C1/TC,C2/ MC	Its mRNA and protein expression are associated with GBM tumor. [Santosh V et al. (2010)]
IGFBP5	TC,C1/TC,C2/ MC/CPHM	Same as IGBPF3
CD24	TC,C1/C-SVM/CPHM	It is overexpressed in glioma cells and Western plot analysis shows that it is associated with GBM patients' survival rate. [Deng J et al. (2012)]
FDGFRA	TC,C1/C-SVM	PDGFRA gene rearrangements are frequent genetic events in PDGFRA-amplified glioblastomas. [Ozawa T et al. (2010)]
COL4A1	TC,C1/MC/C-SVM	Important markers for GBM. [Dreyfuss J et al. (2009)]
VSNL1	TC,C1/TC,C2/ MC/C-SVM	VSNL1 can regulate proliferative properties of GBM. [Xie Y et al. (2007)]
A2M	TC,C1/TC,C2/ MC	A2M can control secretion of α_2-macroglobulin in glioma cells. [Businaro R et al. (1992)]

TC,C1: Two-Classes SVM Criterion1; **TC,C2**: Two-Classes SVM Criterion 2;
MC: Multiple-Classes SVM; **CPHM**: Cox Proportional Hazard Model;

4. Conclusion

To summarize, we have tried both classification model and inference model to the problem. We find that there are clearly some association between the gene expression and GBM survival information. There are also some important gene signatures identified which are proved experimentally to be associated with survival.

Acknowledgements

We would like to thank Prof. Andrew Ng and TAs of CS 229 for guidance and suggestions. We would also like to thank Prof. David Dill's Lab for providing data.

References

[1] C. W. Duarte *et al.*, "Expression Signature of IFN/STAT1 Signaling Genes Predicts Poor Survival Outcome in Glioblastoma Multiforme in a Subtype-Specific Manner," *PLoS ONE*, vol. 7, p. e29653, 01 2012.

[2] TCGA data portal, https://tcga-data.nci.nih.gov/tcga/.

[3] J. Zhang, B. Liu, *et al.*, "A systems biology-based gene expression classifier of glioblastoma predicts survival with solid tumors," *PLoS ONE*, vol. 4, p. e6274, 07 2009.

[4] E. Bair and R. Tibshirani, "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data," *PLoS Biol*, vol. 2, p. e108, 2004.

[5] I. Guyon *et al.*, "Gene Selection for Cancer Classification Using Support Vector Machines," *Mach. Learn.*, vol. 46, pp. 389–422, Mar. 2002.

[6] X. Zhang, X. Lu, Q. Shi, *et al.*, "Recursive Sample Classification and Gene Selection Based on SVM: Method and Software Description," *BMC Bioinformatics*, vol. 7, pp. 389–422, Apr. 2006.

[7] Roel G.W. Verhaak and Katherine A. Hoadley and others, "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98 – 110, 2010.

[8] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent," *Journal of Statistical Software*, vol. 39, Mar. 2011.

[9] R. M. Simon *et al.*, "Using Cross-Validation to Evaluate Predictive Accuracy of Survival Risk Classifiers Based on High-Dimensional Data," *Bioinformatics*, vol. 12, May 2011.