Predicting Kidney Cancer Survival from Genomic Data

Christopher Sauer, Rishi Bedi, Duc Nguyen, Benedikt Bünz

Abstract

Cancers are on par with heart disease as the leading cause for mortality in the United States. In particular, Kidney Renal Clear Cell Carcinoma (KIRC) has an approximate five-year mortality rate of 53%. Cancer mortality, however, is highly variable, and the side effects of current standard of care treatment regimens are severe and costly. Survival indications and the course of action pursued are largely determined by the stage and the grade of the cancer, metrics with varying predictive values. We apply supervised machine learning techniques to predict mortality using genetic mutations and gene expression, achieving maximum predictive accuracy of 97.2%. Additionally, we explore various feature selection methods, noting the cost constraints that become relevant when conducting expression microarrays and SNP genotyping on increasingly large numbers of genes. We identify 27 particularly notable genes mutually identified by multiple feature selection metrics. Finally, we consider unsupervised learning techniques to search for distinguishable genetic subtypes with significantly differing survival outcomes.

1 Introduction

Current survival indications of cancer are largely tied to discrete measures of disease progression (stage and grade), metrics with varying predictive value for actual prognosis. There is value in increasing prognosis accuracy, in terms of both patient lifestyle decisions and selection of treatment. Rather than relying on clinical prognostic indicators, there is significant recent evidence in the literature that superior survival predictions can be made from applying a statistical approach to genetic indicators, in numerous malignancies. We seek to apply similar approaches to KIRC, relying on supervised machine learning techniques that the univariate analyses that dominate existing literature [8, 9] . Recent advances, particularly with regard to breast cancer, have additionally shown promise in the quest to find structure in mutation clusters [7]. We thus apply unsupervised machine learning techniques here to attempt to distinguish between survival outcome groups based on clusters of shared mutations.

2 Data Set

The Cancer Genome Atlas (TCGA) is a National Cancer Institute-supervised project to make available genomic and exomic data for various cancer types. All our samples come from their publicly available data for the most prevalent type of kidney cancer, renal clear cell carcinoma. Many different feature types are available on a subset of the data. We selected single nucleotide polymorphism (SNP) mutation data (available for 417 samples) and gene expression data (72 samples) because they likely reflect both the source of the cancer and its current state. All samples had clinical data associated with them, i.e., the patients were tracked over a period of time to determine how long they survived. As with all such data, some patients were alive at the time the data was submitted or were not able to be contacted, but this information was recorded in addition to the number of days we know they survived for. We also used American Joint Committee on Cancer (AJCC) stage and grade criteria as baseline clinical predictors against which to measure the success of our genetic classification model. We therefore attempted to predict whether patients had survived to the point of last contact. In the unsupervised part of the study, we used the standard Kaplan-Meier plot to also take the duration of survival into account, based on the survival data as described above.

3 Learning Setup

Creating our preliminary training set required us to conduct an initial phase of preprocessing. Due to the sporadic nature of the TCGA dataset, one of the challenges that we were met with was normalizing the data in such a way that allowed us to compile data about the same patient across several different

sources. Furthermore, not all the patients had the same depth of information linked to them. For example, the NIH provides Somatic Mutation Data for 417 patients, but Gene Expression data for only 72 patients, overlapping but non-identical subsets.

4 Experimental Results

In this paper, we use the following definition for our accuracy metric: TRUE POSITIVE + TRUE NEGATIVE / TOTAL. Conversely, we define error as FALSE POSITIVE + FALSE NEGATIVE / TOTAL.

4.1 Establishing Baseline Models

In the absence of any feature selection metrics, we attempted to predict boolean survival using the somatic mutation and gene expression datasets separately. As shown below, our best result was 81.8% accuracy, using a Neural Network on gene expression data. This is substantially below comparable published results in the literature, indicating severe overfitting.

Algorithm	Somatic Mutations	Gene Expression
Naïve Bayes	49%	77.78%
Decision Trees	75.0%	69.3%
SVM	74.6%	80.56%
Neural Net	65.0%	81.8%

Table 1: Accuracy of four supervised learning algorithms on the entire datasets.

4.2 Preliminary Feature Selection

Although the efficiency of Naïve Bayes classification and Support Vector Machines allows us to learn on all of the features in each dataset, machine learning using other algorithms (e.g. logistic regression) proved infeasible on our extremely large feature set. Though one potential avenue of addressing this problem was to use established methods of feature selection, we decided to test our own algorithms in order to achieve quick improvements from our baseline results, as well as get a feel for the nature of the data. Our naive feature selection algorithm was choosing the top X single features that alone contributed maximum value.

4.2.1 Gene Expression Feature Selection

Our preliminary attempt at feature selection on the Gene Expression Dataset involved running logistic regression on each feature (gene) individually and ranking the features by lowest LOOCV error. The speed of this algorithm comes from the fact that each regression is only run on a one-dimensional feature space, and the fact that the algorithm only makes one pass through the original feature set. A feature subset would be constructed from the top N number of features on our ranking. This smaller subset allowed us to test not only Naïve Bayes classification and Support Vector Machines, but also logistic regression and regularized logistic regression.

Algorithm	Gene Expression
SVM – top 50	76.39%
SVM – top 100	81.94%
Naïve Bayes – top 50	84.72%
Naïve Bayes – top 100	87.50%
Log. Regression – top 14	79.17%
Reg. Log. Reg. – top 14	84.72%

Table 2: Rudimentary feature selection accuracy on the gene expression dataset.

Though testing error improved as compared to our baseline results, we noticed that this approach does not completely address the problem of overfitting. This is apparent in the relationship between training and testing error as the number of features increases. Figure 1 below shows both training and testing error of Naïve Bayes Classification as the number of genes examined is increased.

As training error decreases, testing error increases– a telling result of overfitting. This trend was consistent across Naïve Bayes, SVM, and logistic regression. Though overfitting was mitigated using regularized logistic regression, it is evident that this preliminary feature selection strategy is choosing extraneous features leading to overfitting. Another problem for the logistic regression was that our feature matrix is very sparse. This leads to singularity issues when running gradient descent. One approach to resolve this issue is to reduce the dimension by finding latent variables using factor analysis, or projecting our entire training dataset onto a lower-dimensional subspace using PCA. However, one of our main goals was

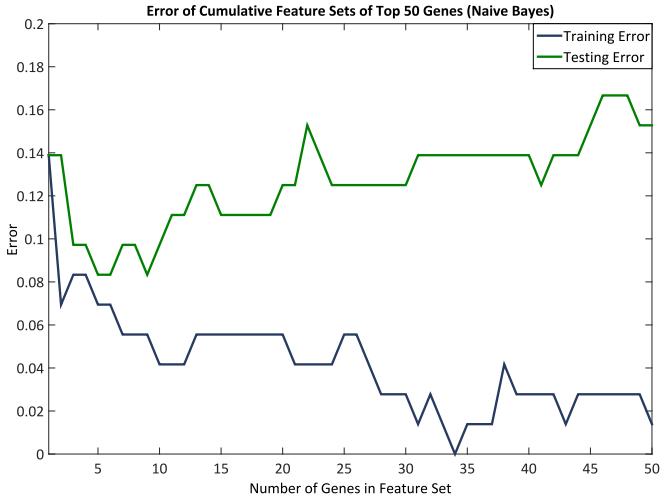

Figure 1: Naïve Bayes Training vs Testing error

to reduce the number of features that need to be measured as there is a physical cost associated with them. Furthermore, although PCA reduces the amount of features examined and increases efficiency, the features that it produces are not selected by predicting power– they are simply lower-dimensional representations of the original data. We thus proceeded using a more robust feature selection on the original feature set.

4.3 Robust Feature Selection

Despite the immediate increase in hit rate achieved through our rapid feature selection approaches, the apparent overfitting led us to pursue more intensive feature selection. We decided on the forward search algorithm in order to choose features that strictly improved testing error. Our forward search terminates once it fails to strictly improve the testing error for 5 consecutive iterations. Each round of the forward search requires evaluating $O(n)$ different models. Given our large feature space had to make some design decisions to get a tractable feature selection. Firstly we only the Naïve Bayes and SVM algorithm were fast enough to be considered. Furthermore, running an $O(m)$ evaluation algorithm, such as LOOCV on each of these models wasn't tractable so we decided to evaluate the models using 10-fold cross validation.

4.3.1 Forward Search on Gene Expression Dataset

Running forward search on the gene expression dataset proved extremely fruitful. In contrast to the results achieved through our preliminary feature selection, both training error and testing error decreased as the number of elements in our feature set increased. On the gene expression dataset, Naïve Bayes forward search selected 5 out of the 16383 genes in the set, and achieved a 97.22% leave-one-out cross validated hit rate. SVM forward search selected a grand total of 3 genes, achieving a 93.1% hit rate.

4.3.2 Forward Search on Somatic Mutations

Seeing the success of forward search on the gene expression we decided to run the same analysis for the somatic mutations data set. Again we could see that using the feature selection we were able to drastically decrease both the training and the testing error. For Naïve Bayes the feature section converged after 39 iterations and had a LOOCV accuracy of 95.7%. Figure 2 plots the training and testing error during the forward selection. Additionally we ran the same forward feature selection using SVMs. Observing only the top 41 genes, the SVM's accuracy was 92.3%. One interesting result was that the genes selected by both the Naïve Bayes and the SVM feature selection were very similar. Concretely over 65% of the genes selected by the Naïve Bayes feaature selection were also seleced by the SVM's feature selection. Furthermore the top 13 genes were not only the same but even in the same order. We further validate the selected features using mutual information analysis in section 4.3.3

These top 13 genes ordered by their selection were: ITGB1, MAP3K2, SPTBN2, RABEP1, C12orf64, SLC12A1, PIPSL, COL17A1, CMA1, OR5P2, CLCN3, FBXL19.

These results are valuable from a clinical standpoint, considering the significant real cost differential in collecting gene expression data for 5 genes or somatic mutation data for 40 genes than it is for over 10,000 genes.

Having a better understanding of which genes to examine could improve practicality of making predictions on prognosis based on gene expression data.

Furthermore, there has been a growing field of research on how to personalize medicine involving the genomic signature of a malignancy[1] .

4.3.3 Mutual Information Heuristics

Given the large time complexity of forward feature selection, we examine the effectiveness of heuristics

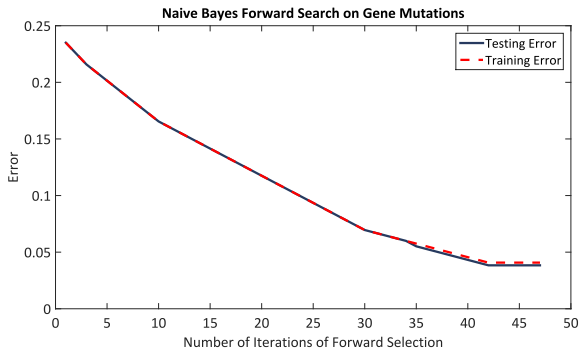

Figure 2: Naïve Bayes Mutations

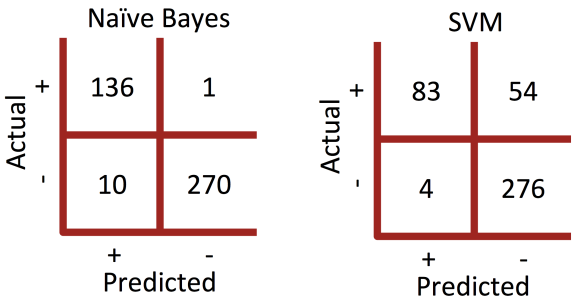

Figure 3: Confusion matrices for Naïve Bayes and SVM with forward search feature selection

that use the mutual information metric to determine important features much more quickly, using the well-studied mRMR metric.

S denotes the feature-set, i, j are any two features in S, and h denotes the output variable. We seek to minimize expression (2), the "redundancy" between two features (i.e, their mutual information), while maximizing included features' "relevance," expression (3), given by a feature's mutual information with the output variable. Expression (1) gives the definition of mutual information, specifically for binary variables (i.e., taking on values of 0 and 1, exclusively).

$$I(i,j) = \sum_{i\in\{0,1\}} \sum_{j\in\{0,1\}} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \quad (1)$$

$$\min W_I, W_I = \frac{1}{|S|^2} \sum_{i,j\in S} I(i,j) \quad (2)$$

$$\max V_I, V_I = \frac{1}{|S|} \sum_{i\in S} I(h,i) \quad (3)$$

Interestingly, 11 of the top 13 genes selected by SVM and Naïve Bayes feature selection were also chosen by the mRMR heuristic. This indicates that forward selection is making intelligent choices by maximizing relevance while minimizing redundancy, but also that mRMR is a computationally efficient way

Algorithm	Somatic Mutations		Gene Expression	
Naïve Bayes	95.7%	+18.1%	97.2%	+19.6%
clinical only	77.6%		77.6%	
SVM	92.3%	+15.6%	93.1%	+16.4%
clinical only	76.7%		76.7%	

Table 3: Forward search feature selection accuracy results, with the red number denoting the added accuracy from including genetic data versus clinical alone.

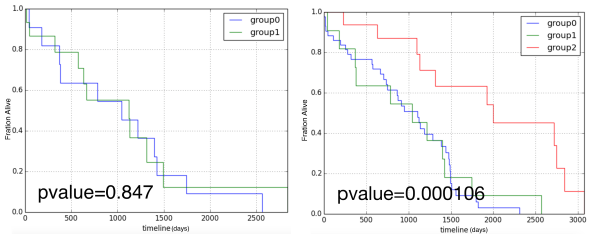

Figure 4: k-means clustering with k={2,3}

to replicate similar results, to within a certain gene depth, past which its selections begin to differ with forward search's.

4.4 Unsupervised Clustering

Having achieved some success in directly predicting clinical outcomes, we decided to also explore whether kidney cancer might have several distinguishable subtypes with differing survival outcomes. This would support the longstanding theory that even within each anatomical region, there are several discrete paths of mutations that can lead to cancer. Each path ought to lead to a different subtype of the cancer that would be distinguishable from other subtypes from differences in both gene expression and mutations.

We first ran k-means clustering on purely the gene expression data with 2, 3, and 4 centroids. Since there are only 72 samples with gene expression information, we are restricted to relatively few centroids to maintain a meaningful number of samples in each group.

After clustering the samples into several groups based solely on gene expression, we compared the survival outcomes of the groups characterized by differences in gene expression. The Kaplan-Meier plot showing the survival differences for each of the three sets of groupings is shown below:

The significance of the difference in survival was determined using the multivariate logrank test. Splitting the data into 3 groups yielded a highly signif-

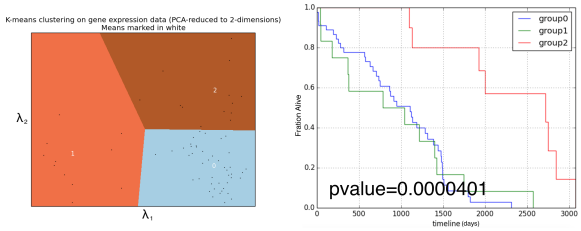

Figure 5: K-means clustering on PCA features and associated Kaplan-Meier survival plot.

icant difference in survival with one group surviving much better then the other two. Some of this is preserved with 4 centroids. Since, again, the groupings were determined only on gene expression, this suggests that there are at least three subtypes of KIRC, with one having significantly better survival outcomes than the others.

To better visualize the results, we next tried first performing PCA to reduce the dimensionality of the data before running k-means and plotting survival differences. The plot of the expression data reduced to two dimensions and the corresponding Kaplan-Meier plot are below. Group numbers are preserved between the two. The differences between groups only became more significant when PCA was performed first.

We attempted to perform similar a similar set of analysis on mutations, but k-means and other clustering algorithms performed poorly on Boolean data. This is because the distances between samples are discrete and the same, so the algorithms tend to produce a single monolithic group with all other groups having size one. Normalizing and performing PCA to reduce the dimensionality of the data failed to correct for this problem, despite the fact that this made the features closer to being real-valued.

5 Conclusion

In this paper we have applied several different machine learning techniques to genomic data from kidney cancer patients. Concretely we attempted to predict the mortality of cancer patients based on the expressions and somatic mutations of their genes. The genomic data is characterized by an extremely high dimensional feature space and a relatively small number of samples. We have shown that in such a setting a plain application of these state of the art algorithms does not result in good and generalizable predictions of the mortality of such patients.

However, if we apply the same algorithms on a specific subset of the genomic data we can successfully provide high accuracy predictions. Our experiments show that such a subset can be several magnitudes smaller than the original feature set. Concretely, by using forward feature selection, we have been able to correctly predict the mortality for over 95% of the patients using only 40 somatic m ation features instead of the original 12,000.

Moreover this extreme reduction of the feature space does not only reduce the risk of overfitting. Measuring a feature has a significant real world cost both in terms of time and money.

Finally we additionally demonstrated that there seem to be inherently different subtypes of kidney cancers. Using unsupervised learning techniques we were able to separate the patients into three groups. The difference in survival probability for each of these groups was highly significant.

In conclusion we have shown that machine learning techniques can be very successfully applied to genomic data for cancer patients.

References

[1] Villarroel, Maria C., et al. "Personalizing cancer treatment in the age of global genomic analyses: PALB2 gene mutations and the response to DNA damaging agents in pancreatic cancer." Molecular cancer therapeutics 10.1 (2011): 3-8.

[2] Chow W-H, Dong LM, Devesa SS. Epidemiology and risk factors for kidney cancer. *Nature reviews.* Urology 2010;7(5):245-257. doi: 10.1038/nrurol.2010.46.

[3] Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics.* 2008;9 Suppl 1:S13.

[4] Loeb LA, Loeb KR, Anderson JP. Multiple mutations and cancer. *Proc Natl Acad Sci USA.* 2003;100(3):776-81.

[5] Nowell PC. The clonal evolution of tumor cell populations. *Science.* 1976;194(4260):23-28.

[6] The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

[7] Cancer Genome Atlas Network. "Comprehensive molecular portraits of human breast tumours." Nature 490.7418 (2012): 61-70.

[8] Gross, Andrew M., et al. "Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss." Nature genetics (2014).

[9] Rossi, Davide, et al. "Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia." Blood 119.2 (2012): 521-529.