Data Fusion for Predicting Breast Cancer Survival

Linbailu Jiang, Yufei Zhang, Siyi Peng
Mentor: Irene Kaplow

December 11, 2015

1 Introduction

1.1 Background

Cancer is more of a severe health issue than ever in our current society. As severe as it is lethal in general, there are many factors that may affect a patient's survival. It is not easy to find a clear pattern to predict the survival outcome of a cancer patient, which could be a complex process involving different biologic conditions. Based on previous studies, many features are considered to potentially affect a survival event – cancer type, age at diagnosis, treatment pathway, time since diagnosis, and some specific genome patterns that may relate to the progression of the cancer disease. While some of these features, such as cancer type and age at diagnosis, are relatively explicit and have more direct relations with the survival outcome, some other features, such as gene mutations and methylation levels, seem more complicated and require more effort to analyze the potential interactions among these features. In this paper, we mainly focus on the pathological conditions of the patients without examining different treatments of cancer.

1.2 Goal and Outline

In this paper, we want to understand a patient's survival rate given his/her genome pattern and time interval since diagnosis. Our ultimate goal is to predict the survival rate of a patient with breast cancer changing over time based on some related genomic data.

Our first step is to understand how genomic data is influenced by different factors, such as gene expression for different RNA, copy number value, and the methylation level for different DNA. By doing feature correlation analysis and feature selection in genomic data, we identify cancer-related genes and their targets. The most challenging task here is to distinguish the "driver" mutations, as a subset that truly contributes to the tumor's progression, from a large number of neutral "passenger" mutations that characterize the cancer. Based on some previous studies, we guess that a support vector machine method might be helpful during this process.

At the second stage, we have relatively fewer features for genomic data that may affect the cancer survivals, so it could be easier for us to conduct a merge based on the patient ID in the survival dataset and the sample ID in the genomic dataset. This would help us to relate all the genomic information to the survival results so we could combine them with other potential features and start to train our model. After that, we would estimate and compare the performance of the models by using some cross validation methods.

1.3 Data

The data of our project is from NIH(National Institutes of Health) Project, and we obtained them from Professor Olivier Gevaert. The data are all pre-possessed, log-transformed, and well-separated based on cancer type into 11 dataset.

The whole dataset is comprised of two parts. First part is a dataset of patient ID, cancer type, time since diagnosis ("TimeToLastContactOrVisit" in days), and his/her survival status (normally 0=alive, 1=dead). Second part are some large lists about genomic data. Each cancer type has one large list. In each large list, there are 3 datasets which separately contain gene expression data, copy number data, and methylation data. All the data are pre-processed so there can be some negative values in these charts.

The first genomic data we used in this project is the gene expression data. Basically, for each sample, a high (large positive) gene expression value for a specific gene code means the information encoded in this gene has been highly "interpreted", while a low value indicates that most of its information has been "hidden". The second genomic data is called copy number variations data. This data represents the structural variation of a specific gene. A larger copy number implies that the gene might be duplicated so that it becomes more than the normal number while a smaller copy number denotes a deletion in a specific region so that the number is less than the normal number. During these years, methylation also becomes an important concept in the cancer research, so our third dataset records the pattern of methylation for each sample. An aberrant DNA methylation pattern, such as hyper-methylation or hypo-methylation pattern, can usually be associated with many types of human malignancies.

In this project, due to time limit, we would only focus on breast cancer(BRCA). The sample sizes of the BRCA dataset is relatively large compared with others, so we believe it's more likely to get valuable results when training on this dataset.

2 Methods

2.1 Feature Selection

For most of our datasets, the numbers of features are much larger than the numbers of sample (p » n), which can cause some difficulties when training the data. For instance, the sample size of the gene expression dataset for breast cancer is only 985, while the corresponding feature size is 16020, which is much larger than the sample size here. A potential problem that can be caused by this is overfitting. To avoid this problem, the first thing needed to be handled is to reduce the number of variables in these datasets.

The first step we did here is to find variance of each feature and get rid of the features with low variance. For example, the variance range of gene expression features for breast cancer is from 0.03 to 25.50 and we find most of these features have relatively low variances. We assume that the low-variance predictors have less predictive power (which is not always true, we just use this simple method to obtain a first impression of the data), so we tried removing some predictors with small variance and use 10-fold cross validation to see how this process may affect the model performance. Table 1 shows the model performance based on different variance thresholds.

# of Selected Features Based on the Threshold	5909	1673	598	78	13
Variance Threshold	1	3	5	10	15
SVM	0.110037	0.109963	0.109976	0.109975	0.110061
Naive Bayes	0.352112	0.347705	0.332234	0.269524	0.121026

Table 1: Error estimations of SVM & Naive Bayes using 10-fold cross validation

As we see, for an SVM model, there's no obvious improvement by reducing feature size;

however, for a Naive Bayes model, the smaller the number of variables is, the better accuracy is achieved. One possible explanation for this is that the actual model (with raw data) violates the conditional independence assumption of the Naive Bayes model, which means some genes may work together, and their expression values can probably be highly correlated. The accuracies for both SVM and Naive Bayes model are close to 90 percent, however we don't think these models are actually good at this point since we found that all of them tend to predict 0(alive) rather than 1(dead), and the corresponding ROC curves imply that the models are uninformative.

To solve this problem, we decided to improve our feature selection process and use a more reliable method to pick important features. According to previous studies, training logistic regression models on each feature separately and ranking them by lowest CV error can be a good method to find valuable features. We picked the top 200 genes with the best performance on individual training, and then ranked them again by using cox model to reduce the final feature size to be 20 (details in next section).

2.2 Survfit and Cox Model

After significantly reducing the number of feature variables, we then combined all the important features to the survival dataset. One important point to notice is the feature variable "TimeToLastContactOrVisit". It indicates the number of days from a patient was first diagnosed breast cancer to his/her last visit date. If the patient is dead, this variable represents the duration to death; if not, this variable becomes a censored data, as we don't know actually what the patient's current status is. We only know this patient was alive x days after the beginning of study (x = "TimeToLastContactOrVisit"). In previous part, to simplify the model, we treated the "TimeToLastContactOrVisit" as a continuous feature variable and trained SVM models on it. However, the performance of these models seems unfavorable, which may be explained by us not handling this term in a right way. An alternate method to deal with it is to treat this term as a part of response instead of a predictor. We will discuss why it could be important to correctly handle this term later when we analyze our sample results.

Based on previous study, a common way to combine the timeline and status of an event is to create a new type object, usually called "Surv" or survival object, which is a small data matrix that contains comprehensive information of an event. In this project, we used cox model to fit the survival data. The cox model has the form:

$$\begin{aligned}\lambda(t|X) &= \lambda_0(t)\exp(\beta_1 X_1 + ... + \beta_p X_p) \\ &= \lambda_0(t)\exp(X\beta')\end{aligned}$$

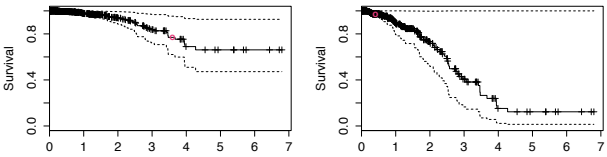
We used functions "Survfit" and "coxph" in the survival package in R to train the models with cox model and then fit the survival object.The next problem we met in this part was the large time cost of fitting cox models on hundreds of features. To improve the efficiency of our models, we ran cox models on each feature individually and pick the top 20 cox-ranked genes to be our final features in our model. After training a cox model on the 20 features, and predicted a survival curve for each test sample and calculate the cross validation error for our model.

3 Results and Discussion

3.1 Interpreting Survival Curves

Different from other types of response, survival objects cannot be directly compared with the original test data. Therefore, we need to find a way to transform our predictions back into two-level (0/1) factors to cal-

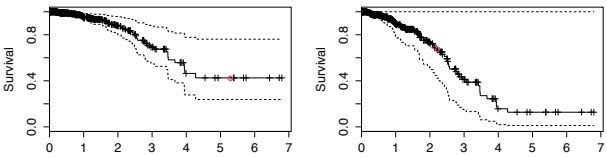
culate model errors. Figure 1,2 and 3 are several examples of survival curves that can help to illustrate this transform process. Each plot represents one patient's survival curve.



(a) Ex1: Event 0 on the 1321th Day (b) Ex2: Event 0 on the 189th Day

Figure 1: Survival Curves for test samples with Event 0(alive)

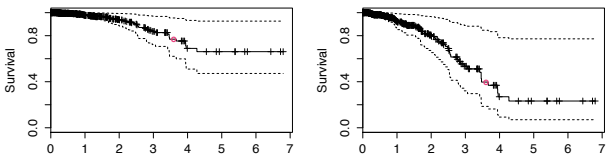
Figure 1 compares two examples with status 0 at last contact. Although the two survival curves look very different, they can have the same event status based on different timelines. The survival curve of the 1st sample seems much better than the curve of the 2nd sample, so we may predict that the survival rate of the 1st person is higher than the 2nd person. However, after considering the time information of the last contact, it's easy to find that the survival rate of the 1st person on the 1321th Day is around 0.75 while the survival rate of the 2nd person on the 189th Day is around 0.98, which is higher than the survival rate of the 1st person in this case. This example clearly illustrates that "TimeToLastContactOrVisit" is a crucial term that can affect our predictions to a great extent.



(a) Ex3: Event 1 on the 1920th Day (b) Ex4: Event 1 on the 825th Day

Figure 2: Survival Curves for test samples with Event 1(dead)

Figure 2 compares two exapmles with status 1 at last contact. Again, it's clear that two survival curves are very different, where the first patient's survival curve is much flatter and higher than the second one. However, since we are looking at different timeline, the two patients are both predicted dead at their timeline, respectively.



(a) Ex1: Event 0 on the 1321th Day (b) Ex5: Event 1 on the 1365th Day

Figure 3: Survival Curves for test samples with Events 0 and 1

Figure 3 gives us an example of two patients with similar timelines and different survival curves. Since at the 1321th day, the first patient's survival rate is around 0.7, while the second patient's is around 0.4, we predict the first patient alive and the second patient dead.

From the example of three comparisons of graphs, we can see that both survival curve and and timeline determine how we predict the living condition of a patient.

3.2 Comparing Models on Different Datasets

Breast Cancer Model Trained on:	Gene Expression	Copy Number	Methylation
Top 20 Cox-ranked Genes	83.6%	52.3%	65.3%
Top 40 Cox-ranked Genes	85.2%	54.1%	69.6%
Top 60 Cox-ranked Genes	86.1%	54.8%	70.0%

Table 2: Accuracy of BRCA Models on Different Datasets & Feature Size

From the model comparison result, we can see that both gene expression data and methylation data have a decent prediction accuracy on the patient's survival rate. Top 20 Cox-rankedg genes are good enough to make rather accurate predictions.

4 Conclusion

From various method, we found that treating both "TimeToLastContactOrVisit" and "event"(the survival status) as a survival object and fitting a cox model to it is a good approach to train and predict the survival status of cancer patients. It has much higher accuracy on predicting the patients' survival status than simply treating the whole problem as a classification model and implementing support vector machine or Naive Bayes model. Both gene expression and methylation dataset work well as feature variables, and using the top 20 Cox-ranked features is enough accurate to make good predictions.

5 Future Work

Currently, we set the threshold as 0.5 in the cox model to predict the cancer patient's survival status. In the future, we could raise the threshold so that we will increase the specificity while not decreasing the sensitivity too much.

Moreover, we hope to find a way to combine gene expression and methylation data and use combined dataset to have a better prediction on cancer patient's survival status.

Other than breast cancer, we could broaden our research on other cancer types as well.

References

[1] https://en.wikipedia.org/wiki/Gene_expression

[2] https://en.wikipedia.org/wiki/Methylation

[3] Magali Champion, Olivier Gevaert,Multi-omics data fusion for cancer data

[4] https://en.wikipedia.org/wiki/Proportional_hazards_model

[5] Geaghan M, Cairns MJ (2015). "MicroRNA and Posttranscriptional Dysregulation in Psychiatry". *Biol. Psychiatry* **78** (4): 231-9.

[6] Zaidi SK, Young DW, Choi JY, Pratap J, Javed A, Montecino M, Stein JL, Lian JB, van Wijnen AJ, Stein GS (October 2004). "Intranuclear trafficking: organization and assembly of regulatory machinery for combinatorial biological control". *J. Biol. Chem.* **279** (42): 43363-6.

[7] Hegde RS, Kang SW (July 2008). "The concept of translocational regulation". *J. Cell Biol.* **182** (2): 225-32.

[8] https://en.wikipedia.org/wiki/Copy-number_variation

[9] https://en.wikipedia.org/wiki/Copy_number_analysis