Data Fusion for Predicting Breast Cancer Survival

Linbailu Jiang, Yufei Zhang, Siyi Peng

Abstract

- In this project, we want to understand a patient's survival rate given his/her genome pattern and time 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.
- By training logistic regression models and cox models on each feature separately and ranking them by lowest CV error, we largely reduced the number of features in our models.
- We combined the timeline and status of each event,trained cox models, plotted survival curves for test samples,and then calculated test errors to evaluate our models.

Background

- *Cancer is usually considered as one of the most terrifying diseases in our current society:*

As severe as it is lethal in general,there're many *that may affect a patient's survival.*

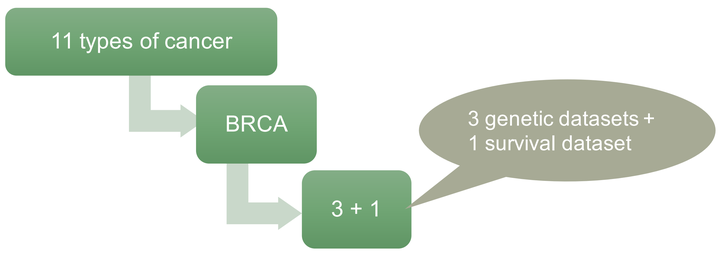
Age,treatment pathway, **genome patterns**…

Is it possible for us to create a model to **predict the survival rate of a patient,**just by analyzing his/her **genetic** data?

Data & Preprocessing

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.



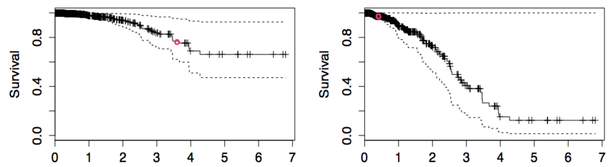
Method Overview

- ***Main Work:***
- **A. Feature Selection:**
- 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 **over-fitting**.
- Method 1: find variance of each feature and get rid of the features with low variance.
- Method 2: train logistic regression models & cox models on each feature **separately** and **ranking them by lowest CV error.**

- **B. Surv Object & Cox Model:**
- Use Surv object to combine the information of both timeline & status of events
- Fit cox models and analyze the survival curves for test data

Results & Discussion

- ***Analyze survival curves***



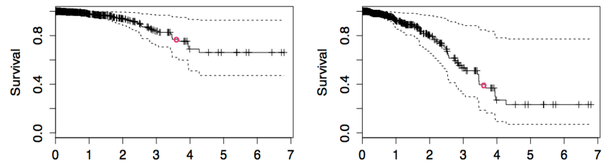
(a) Ex1: Event 0 on the 1321th Day (b) Ex2: Event 0 on the 189th Day

Figure 1: Survival Curves for test samples with Event 0

Both** status are **alive**,but curves look very different. **Why?

*They have **different timeline!***

Results & Discussion



(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

Similar timeline** with **different status.
Person 1 is predicted to be alive on 1321th day & Person 2 is predicted to be dead on 1365th day. Both are consistent with their actual status in test data.

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

Gene expression dataset** seems to be the **most informative!

Method Comparing

Remove low-variance features ☹

Select top cox-ranked features ☺

Treat contact time as continuous features ☹

Treat contact time as part of response ☺

SVM & Naïve Bayes ☹

Cox model ☺

Conclusion & Future work

- ***Conclusion***

Best combination of algorithms:

Select top logistic/cox-ranked features + treat contact time as part of response (in Surv objects) + Cox model

- ***Future Work***

Combine gene expression & methylation datasets to fit models

- ***Acknowledgement***

Irene(Mentor), Professor Andrew Ng, CS 229 Teaching Stuff, Professor Olivier Gevaert,Magali Claire Champion, NIH(National Institutes of Health).