Predicting Breast Cancer Survival from Genomic Data

Changmao Li, Han He, Yunze Hao, and Caleb Ziems

1. Project Goal and Hypothesis

The goal of this project is to train machine learning classifiers that can predict the 10-year survival of breast cancer patients. Our first class of model (**M1**) will use features from the highly refined, low-dimensional *Clinical Patient Data*. Our second class of model (**M2**) will use features from the high-dimensional *CNA*, *Gene Expression* and *Mutations* data. We hypothesize that our best **M2** classifier will outperform all **M1** classifiers.

2. Exploratory Data Analysis and Preprocessing

We will be considering the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset, which contains clinical as well as genomic information for each patient [1]. Much of the data contains missing values. We specifically indicate methods for dealing with these values in the following subsections. The data analysis and preprocessing will all be done using pandas and scikit-learn's preprocessing library.

Clinical Patient Data

The Clinical Patient Data contains entries for 2,509 patients, each containing 17 explanatory variables and three response variables. We see that 14 of the 17 explanatory variables come from discrete categories including the genomic classification of the cancer (ER/PR/HER2), the estrogen receptor status, and a binary indication of chemotherapy treatment. In our preprocessing state, it will be necessary to use dummy variables for each category.

The remaining three explanatory variables contain continuous numerical data from the Nottingham prognostic index (NPI), the patient's age at diagnosis, and a count of the observed cancer-positive lymph nodes. We will use these variables in a regression model, so we will standardize them first. We will also need to consider the correlations between variables. For instance, NPI and lymph node count have a Pearson correlation of 0.55.

The three response variables provide the overall survival time or the time a patient was last seen alive, their status (either living or dead), and whether breast cancer was the cause of their death. We summarize the response into a binary classification: either the patient died from their cancer before 120 months or they lived past 120 months. We ignore patients who were last seen or had died of other causes at time less than 120 months. Removing unlabeled samples and semi-supervised learning are two ways to work around. After removing these, we are left with 1,469 patients with binary response categories. We take care to observe that the data somewhat imbalanced, with 390/1125 = 0.35 positive examples. Since 41.5% of our samples are removed, leading to the lack of data. It will be interesting for a semi-supervised framework to exploit those unlabeled data. More specifically, we plan to employ some practical semi-supervised learning algorithms, including self-training, multi-view learning, to make full use of our data.

We now turn to missing values. Each column in the Clinical Patient Data contains anywhere from 11 to 745 missing values with a median number of 529. We will need to decide if these values are missing completely at random (MCAR). If the MCAR condition is satisfied, we can safely drop the missing values without biasing the data. This will bring us to 1,125 patients.

Gene Expression

This table provides gene expression data for 1,903 patients. Gene expression is given as the abundance of mRNA for that gene, and we have data on 24,367 genes. We see that the number of features is much greater than the number of observations. Therefore high variance and overfitting are both concerns. We will rely on the L1 regularization for automatic feature selection in the regression models. When λ is large enough, many solution coefficients will be zero. We will also consider XGBoost and other decision tree models, which handle overdetermined data well. And to deal with missing values, we can use imputation. This time, there are only 11 patients with missing values.

Mutations

Mutation in gene can lead to significant consequence including null mutations, abnormal protein product, etc. Germline mutations in *BRCA1/BRCA2*, are known to considerably increasing the risk of occurring breast cancer and ovarian cancer in familial cases [2]. To include mutations in our model while avoid creating too many dimensions, it is reasonable to pick one representative mutation out of mutations occurred on a same gene from a same patient. This protocol significantly reduced the total dimensions while only compromising minor amount of information.

CNA

The CNA data gives us the number of copies for 22,543 genes across 2,174 patients. A copy number of -2 indicates both copies have been deleted, and -1 means one has been deleted, and 0 means the patient has both copies. However, a 1 indicates a patient with more than two copies and a 2 indicates one many more than two copies. For this reason, we cannot interpret these values as linearly ordered. Instead, they can be treated as categories with one-hot encodings. This entire table will then be inner joined on patients with mutations and gene expressions.

3. Methodology, Model Selection, and Tuning

Since our clinical data is primarily categorical, our choice of model is restricted. For example, our discrete categories will not work well in a K-nearest-neighbors approach. We are choosing to use a classification approach, and so we will consider many off-the-shelf algorithms implemented in scikit-learn, including decision trees, random forests, and gradient boosting, etc. We will then implement four advanced classification algorithms for handling the high-dimensional genomic data and we will compare these approaches for the **M1** and **M2** models.

L1-Regularized Logistic Regression

Caleb will be implementing L1-Regularized Logistic Regression. This algorithm is a well-known solution to over-fitting with underdetermined data, and we selected it especially for its interpretability. The implementation we will be using is based on Lee et al. (2006), which solves the L1-constrained optimization problem with LARS in iterative steps [3].

XGBoost

Yunze Hao will be implementing a pipeline based on XGBoost . XGBoost is a scalable end to end sparsity-aware implementation of gradient boosted decision trees with great speed and performance. The implementation is based on its paper[4].

Multilayer perceptron with grasshopper optimization

Changmao will be implementing multilayer perceptron with grasshopper optimization. Multilayer perceptron is a kind of simple fully connected neural network which has multiple layers. Grasshopper optimization is a kind of non-gradient optimize methods which has high probability to find more optimal value than gradient based algorithms. The implementation is based on [5].

Semi-supervised learning

In our dataset, 41.5% patients were last seen or had died of other causes at time less than 120 months, which can be considered as unlabeled data. Instead of ignoring them, **Han He** will explore semi-supervised learning algorithms to improve classification rate. Starting from self-training, we will implement some classical semi-supervised algorithms, then report and analyze the performance gain.

4. Validation

Our validation metric will be the Receiver-Operating Characteristic Area-Under-Curve (ROC-AUC). In doing so, we compare the true positive and false positive rates, and we expect to achieve areas greater than 0.5, which is random guessing.

5. Timeline

We will begin the exploratory data analysis and preprocessing in the upcoming week, aiming to finish by 2/28. We will then begin applying baseline methods before spring break on 3/8, and we will run our initial validations by 3/15. We will dedicate the next week to our stretch algorithms from 3/18-3/25. After that, we will work on debugging and consider modifications by 4/1. Then from 4/1-4/10, we will finish up these algorithms and begin final validations. Finally, we will begin the report by 4/10, and finish by 4/22. The last week (4/22-4/29) will be spent on presentations.

References

[1] Curtis, C., Shah, S. P., Chin, S., Turashvili, G., Rueda, O. M., Dunning, M. J., . . . Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403), 346-352. doi:10.1038/nature10983

[2] P Lux, Michael & A Fasching, Peter & W Beckmann, Matthias. (2006). Hereditary breast and ovarian cancer: Review and future perspectives. Journal of molecular medicine (Berlin, Germany). 84. 16-28. 10.1007/s00109-005-0696-7.

[3] Lee, S., Lee, H., Abbeel, P., & Ng, A. Y. (2006). Efficient L1 Regularized Logistic Regression. AAAI.

[4] Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.

[5] Heidari, Ali Asghar & Faris, Hossam & Aljarah, Ibrahim & Mirjalili, Seyedali. (2018). An Efficient Hybrid Multilayer Perceptron Neural Network with Grasshopper Optimization. Soft Computing. 10.1007/s00500-018-3424-2.