Breast Cancer Prognosis

Catherine Lu and JJ Liu
{cglu, jsquared}@stanford.edu

December 14, 2012

1 Introduction

An accurate breast cancer prognosis, or breast cancer survivability prediction, is important as it often guides the treatment course of action, ability to claim additional financial support from the government, actions of the patient and family, and more [1]. Predicting breast cancer survivability is commonly done using clinical features. TNM staging, the globally accepted standard used to describe cancer, was devised more than 60 years ago and only looks at three features: size of the tumor, number of regional lymph nodes with cancer, and the spread of cancer to other parts of the body. With the advent of affordable genomic sequencing and acceleration of findings in molecular biology in the past decade, molecular features may be practical to improve breast cancer prognosis.

Molecular diagnostics for cancer therapy decision-making have shown initial promising clinical results. This has lead to a flood of published reports of signatures predictive of breast cancer phenotypes, and several molecular diagnostic tests for cancer therapy decision-making have gained regulatory approval in recent years [2, 3]. However, there is no consensus for the most accurate computational methods and models to predict breast cancer survivability. In addition, it is unclear that incorporating molecular data as a complement or replacement for traditional clinical diagnostic tools adds any value [4]. Therefore, it is necessary to objectively assess whether genomic data currently provides value beyond traditional clinical diagnosis tools.

To aid in efforts to solve this problem, we predicted breast cancer survivability with machine learning techniques as part of the DREAM Breast Cancer Prognosis Challenge. The ultimate goal of the challenge is to objectively compare many computational algorithms through providing a common training dataset in an effort to find the best features for breast cancer prognosis. The dataset provided contains standard clinical measurements in addition to genomic information, thus allowing genomic information to be compared with standard clinical features.

2 Objective

Our goal was to predict survival for each individual. We had two approaches: predicting a discrete survival status based on time since diagnosis and other features, and predicting a continuous survival time based on all features.

3 Data

Breast cancer sample data is made available through the DREAM challenge from the METABRIC data of 1,000 breast tumor samples used in a previous study [2], where data origin and preprocessing is explained in detail. We further process the data by discarding samples with missing values, and are thus left with 931 samples.

3.1 Survival

There are two indicators of survival: time from breast cancer diagnosis to last follow-up and status of the patient (alive or dead) at last follow-up time. Survival data is right-censored, since patients may be alive at the end of the study or lost to follow-up.

3.2 Gene Expression

Gene expression is generated using molecular profiling platforms, described in full detail in another study [2]. The genes used as training features are narrowed to a list of 9 suggested by the DREAM challenge and previous literature. We used two estrogen pathway genes (ER and PR), two human epidermal growth factor 2 receptor amplicon genes (HER2 SNP6 and GII), and five immune response genes (CXCL10, STAT1, GBP1, GZMA, and CD19).

3.3 Clinical Annotations

In addition, we have the following clinical annotations, the classic features used for breast cancer prognosis:

Feature	Metric	Description
Age	Years	Age of patient at diagnosis
Treatment	NONE, HT, RT, CT, HT/RT, HT/CT, RT/CT, CT/HT/RT	HT: hormone therapy RT: radiation therapy CT: chemotherapy
Lymph nodes positive*	0, 1, 2, 3	Number of lymph nodes found with cancer 0: no nodes 1: 1-3 nodes 2: 4-9 nodes 3: over 9 nodes
Size*	0, 1, 2, 3	0: 0-20 mm 1: 21-50 mm 2: over 50 mm 3: Direct extension to chest wall or skin
Grade	0, 1, 2	0: Nottingham score 3-5 1: Nottingham score 6-7 2: Nottingham score 8-9 The score is a semi-quantitative measure of three histopathological characteristics seen under a microscope by a pathologist.
Estrogen Receptor Immunohistochemistry (ER IHC)	+,-	Presence of ER from IHC protocol

*Used in standard TNM classification of breast cancer

4 Measuring Performance

4.1 Predicting Survival Status

We initially build machine learning models that predict the patient's status (dead or alive) based on all other features. We measure performance using 3-fold cross validation accuracy in addition to a data set accuracy for training and predicting on the same, entire data set.

4.2 Predicting Survival Time

Next, we predict survival time of the patient. However, we do not have survival time for all patients; the data is highly skewed and right-censored. Patients may drop out of the study at any point or still be alive by the end of the study.

With a data set of only 931, it is extremely important to still use all of the training data. Two patients' survival times can be ranked not only if both have uncensored survival times but also if the uncensored time or one is smaller than the censored survival time of the other. One of the most commonly used performance measures for survival models is the concordance index (CI) [5]. CI is the fraction of all pairs of subjects whose predicted survival times are ordered correctly across all patients. A CI of 1 indicates perfection prediction accuracy, while a CI of 0.5 is as good as a random predictor.

Hence, we measure performance using 3-fold cross validation (3-fold CV) for CI in addition to CI for training and predicting on the same, entire data set.

5 Development and Results

5.1 Predicting Survival Status

We used patient status as the target variable and all other features as the input features. We used the R Caret package, which provides a library for a number of machine learning models, to write and run different algorithms.

5.1.1 Results

First, we used the K-Nearest Neighbor algorithm to classify our data based on the closest feature training samples. We use a k-value of 2, to see if there were any underlying relationships among features for patients based on status. However, our 3-fold CV accuracy was low (0.519).

We then tried 5 supervised learning models. None of them performed better than 0.556 for 3-fold CV, though running and predicting on the entire data set gave values ranging from 0.693 to 0.716. The models were overfitting the data and were not representing the relationships between the features accurately.

In particular, the Gradient Boosting Model (GBM), an ensemble learning method which uses multiple weak prediction models to form a single model in a stage-wise fashion, resulted in the most overfitted model.

Algorithm	3-Fold CV	Data Set
K-Nearest Neighbor	0.519	0.597
Multinomial	0.549	0.694
Linear Discriminant Analysis	0.542	0.693
Generalized Linear Models	0.545	0.694
Linear Support Vector Machines	0.556	0.697
Generalized Boosted Model	0.541	0.716
Cox Proportional-Hazard Regression*	0.702	0.706
Random Survival Forest*	0.813	0.812

*Concordance Index

Out of the standard machine learning approaches, linear SVM performed slightly better than the rest, possibly because it did not overfit the data as much as other models.

It is interesting to note that Linear Discriminant Analysis (LDA) performed approximately the same as Generalized Linear Models (GLM), even though LDA is a more simple model than GLM. LDA finds a linear combination of our clinical features which characterizes the patient survival status. We also used GLM, a generalization of ordinary linear regression models that allow for response variables that do not follow a normal distribution, because our response variables do not necessarily follow a normal distribution, but instead could follow a distribution more similar to a log-odds model due to our prediction of status as a Bernoulli variable.

5.2 Predicting Survival time

We then predicted survival time using all features as input and the CI as the measurement of model performance. The outputted survival models compute the time it takes for death to occur according to the features.

5.2.1 Cox Proportional-Hazard Regression with Akaike Information Criterion

Proportional hazard (PH) models are the standard for studying the effects of features on survival time distributions. A hazard function $\lambda(t)$ measures the instantaneous rate of death at time t.

The PH model assumes there is a multiplicative effect of the features on the hazard function:

$$\lambda(t|x) = \lambda_0(t)e^{(w^T x)} \tag{1}$$

where $\lambda(t|x)$ is the hazard function with features x, $\lambda_0(t)$ is the baseline hazard function when $x = 0$, w is the vector of unknown parameters, and $e^{w^T x}$ is the relative hazard function.

The Cox Proportional-Hazard [6] approach estimates weight w by leaving the baseline hazard function unspecific and maximizing the likelihood:

$$L(w) = \prod_{T_i \text{ uncensored}} \frac{e^{w^T x_i}}{\sum_{T_j \geq T_i} e^{w^T x_j}} \tag{2}$$

where T_i is survival time of patient i.

After this estimation, we trained using weighted linear regression. In order to avoid overfitting, we use Akaike Information Criterion (AIC) on the features passed to the Cox model. The AIC is a measure of the relative goodness of fit of a statistical model, often described as a tradeoff between bias and variance or between model accuracy and complexity. We first find the corresponding AIC values, and selected the model that minimizes information loss.

We obtained a 3-fold CV CI of 0.702, comparable to the CI of 0.812 for training and predicting over the entire data set.

5.2.2 Random Survival Forest

The Random Survival Forest (RSF) algorithm [7] is an ensemble tree method for the analysis of right censored survival data. More specifically, the algorithm performs the following:

1. Draw B bootstrap samples from the original data, where each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data).

2. Grow a survival tree for each bootstrap sample. At each node, randomly select p variables. Then, split the node with the candidate variable which maximizes survival difference between daughter nodes.

3. Grow the tree to full size.

4. Calculate a hazard function (HF) for each tree, and average to obtain the ensemble HF.

Based on the size of our data, we ran a RSF algorithm with the number of trees to grow to 1000. We use the logrank splitting rule, which splits tree nodes by maximization of the log-rank test statistic.

We obtained a 3-fold CI of 0.813, which is also comparable to the CI of 0.812 for training and predicting over the entire data set.

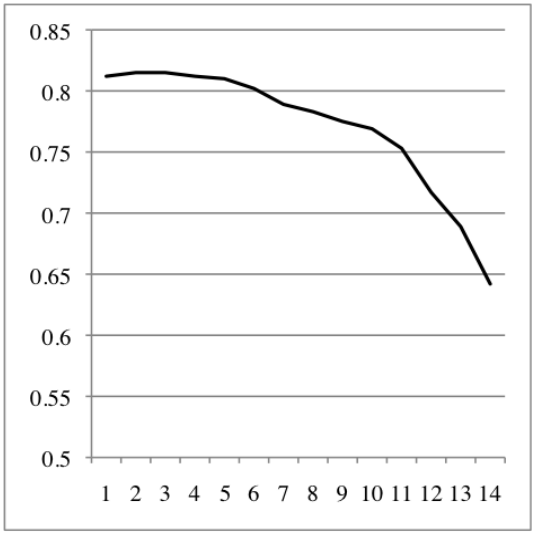
6 RSF Analysis

We chose the RSF model, the best performing model, to gain insights into relationships among features.

6.1 Feature Selection

We determined which features contributed most to the learning using backward search feature selection.

Features ommitted (cumulative)	3-fold CV CI
None ommitted	0.812
ER IHC status	0.815
ER expression	0.815
Grade	0.812
HER2 SNP6 state	0.810
GBP1 expression	0.802
CD19 expression	0.789
Treatment	0.783
CXCL10 expression	0.775
GZMA expression	0.769
PR expression	0.753
Size	0.717
GII	0.689
Lymph nodes posititve	0.642
Age	N/A (all omitted)



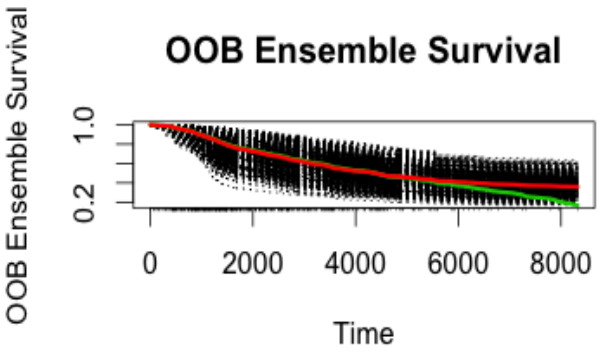
The best 3-fold CV CI was achieved by taking all features except for EHR IHC status and ER expression. EHR IHC status appears to lower the CI and ER expression does not add any value.

6.2 Ensemble Analysis

The following figure shows the ensemble survival function for each patient. The thick red line is overall ensemble survival, and the thick green line is Nelson-Aalen estimator. The Nelson-Aelen, often used to give an idea of the survival rate shape, is given by the equation:

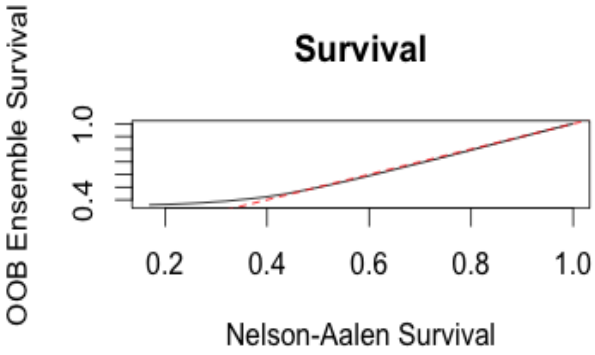
$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{3}$$

where d_i is the number of deaths at t_i and n_i is the total number of patients alive at t_i.



Note that the overall ensemble survival begins to deviate from the Nelson-Aelen estimator at later times.

The second figure below shows the same relationship, where it is shown that RSF tends to predict higher survival probabilities when survival proportions in the data set are low.



7 Discussion

Breast cancer prognosis presents an important challenge with many real life implications. In this paper, we have described our use of various machine learning approaches to the complex problem of predicting breast cancer survivability rate, with the data provided through the DREAM Breast Cancer Prognosis Challenge.

Our results indicate that it is difficult to create accurate standard machine learning models for predicting patient survival status. Survival data has many unique properties. The standard machine learning models did not have any notion of a hazard function for determining patient survival status. Instead, it found unreal relationships that solely existed in the unique data set, which was seen from the large difference in accuracy between 3-fold CV and accuracy from training and predicting on the data set (which were also quite low).

On the other hand, the two models that predicted hazard functions seemed to do quite well, though it is difficult to compare due to the different model performance measurements. It appears that both the Cox and RSF models capture the relationship among features and survival outcome, as seen in almost identical values between the 3-fold CV CI and CI from training and predicting on the data set.

From feature analysis, we learned that at least for the RSF model, age at diagnosis was the best feature predictor. In addition, eliminating two features (estrogen receptor copy number and estrogen receptor gene expression) in the model lead to a slightly higher 3-fold cross validation score than with all features.

From RSF ensemble analysis, we saw that RSF seemed to perform better at predicting either patients with less time since diagnosis or when there is higher probability of survival, or both. Therefore, RSF combined with another algorithm that performs well in these conditions may produce even better results.

This work has limitations and could be improved in three major ways. First, we should examine all genes available in the data set and, using feature selection, find the most predictive genes. Second, we should modify our regular machine learning models to predict the Cox hazard function to give each model the right-censored data relationship that exists. It is not necessarily that RSF is the best predictor of survival out of the algorithms we have used. Third, we should run our algorithms on more data. To do so, we should modify our algorithms to impute or skip missing features without discarding the entire training example and use publicly available data sets.

References

[1] R. Henderson, M. Jones, J. Stare, "Accuracy of Point Predictions in Survival Analysis," Statistics in Medicine, 2001.

[2] L. J. vant Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature, vol. 415, no. 6871, pp. 530536, Jan. 2002.

[3] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med., vol. 351, no. 27, pp. 28172826, Dec. 2004.

[4] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. GrŁf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, M. Group, A. Langerd, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Brresen-Dale, J. D. Brenton, S. Tavar, C. Caldas, and S. Aparicio, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature, 2012.

[5] V. C. Raykar, H. Steck, and B. Krishnapuram, "On Ranking in Survival Analysis: Bounds on the Concordance Index," NIPS, 2007.

[6] J. Fox, "Cox Proportional-Hazards Regression for Survival Data", Appendix to "An R and S-PLUS Companion to Applied Regression", 2002.

[7] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random Survival Forests," The Annals of Applied Statistics, 2008.