A Machine Learning Strategy for Prediction of Time to Recurrence of Breast Cancer

Fen Pei
University of Pittsburgh
fep7@pitt.edu

Seo-Jin Bang
Carnegie Mellon University
seojinb@andrew.cmu.edu

She Zhang
University of Pittsburgh
shz66@pitt.edu

1 Background

Recurrence time of breast cancer refers to the time when breast cancer is likely to recur in patients after the removal of tumor by surgery. An accurate prediction of prognosis for a patient could help a lot in choosing appropriate treatments immediately after the surgery. Besides the costly surgical biopsy method [1], applying machine learning methods to do prediction is highly promising in this field.

In this project, we plan to develop an applicable machine learning method to predict the time-to-recur (TTR) after surgery of breast cancer. One major challenge of the work is to handle the right censored data. Right censored data is composed of two outcomes: recurrence indicator and observed time. The recurrence indicator classify patients into two groups: recurrent ($= 1$) and non-recurrent ($= 0$) patients. If the disease is recurred, TTR will be observed for the patient. However, if the disease is not recurred until the end of the study or it is not possible to follow up a patient, then the only time observed is his/her last check-up time, which is called disease-free survival (DFS).

Previous works employed Recurrence Surface Approximation (RSA) [2], neural networks [3, 4], and genetic programming [5] to make predictions on such dataset. Some of which are described in detail in the section 2. In contrast, our plan is to develop an Expectation-Maximization algorithm to approach this problem. Generalized linear models and RSA were chosen as the baseline methods and the results are presented in this report.

2 Related work

2.1 Recurrence Surface Approximation

Mangasarian et al. [2] approached the problem with a linear programming technique called Recurrence Surface Approximation (RSA). The detailed mathematical explanation is provided in the method section as it is one of the methods we chose as the benchmark. Briefly and intuitively, the method tried to fit a hyper plane to the data by utilizing two types of bounds: the recurrent time for recurrent patients as the upper bounds, and disease-free survival for non-recurrent patient as the lower bounds.

The method was tested on 187 patients with breast cancer (a subset of the data on which our methods were tested) via leave-one-out cross validation technique, and only the predictions known to be wrong were deemed as errors, i.e. only the late predictions of TTR and the early predictions of DFS were accounted for in the validation error. In general RSA performed better than a simple least one-norm program and a variant of its own called poolRSA on all patient data.

2.2 Neural Network

Neural network models were also reported to study the data [3, 4]. Due to the limitation of paragraphs, here we only introduce the first study. They used a neural network with one hidden layer

and ten output units. Each output unit corresponded to the probability of survival within one year, and the entire range was from 0 to 10 years. The TTR was predicted at time t if the t-th output unit is the first one indicating a DFS probability less than 0.5.

The model was evaluated using ten-fold cross-validation. Two criteria were used: the ability of the models to separate favorable and unfavorable prognoses; as well as the accuracy of the predicted recurrent rates (see descriptions below).

Good vs. Poor Prognoses. The survival probability vs. time curves of patients predicted to have recurrence in the first five years (poor prognoses) and those predicted to have recurrence at time points greater than five years (good prognoses) were plotted individually. A statistically significant difference ($p < 0.001$) between these two curves were found, indicating the model was able to distinguish good prognoses from poor prognoses.

Predicted vs. Actual Group Survival. The Kaplan-Meier estimated survival curve for the entire WPBC training set was compared with their predicted DFS rates. The results showed that there were no significant differences ($p = 0.2818$) between the predicted curve and estimated survival curve.

3 Baseline method

Two different baseline approaches are applied to predict time to recurrence: Linear Regression and Recurrence Surface Approximation. The following symbols are used to define the methods throughout this report. (See Table 1)

Table 1: List of notation.

Symbol	Type	Description
k	scalar	the number of s
m	scalar	the number of recurrent patients
n	scalar	the number of non-recurrent patients
X	vector (length k)	$X = [x_1, \cdots, x_k]^T$ where $x_1, \cdots, x_k$ are the value of each feature
M	matrix (size $m \times k$)	feature matrix of recurrent patients s.t. each element $M_{il} = x_l^{(i)}$ where $x_l^{(i)}$ is the value of l-th feature of recurrent patient i
N	matrix (size $n \times k$)	feature matrix of non-recurrent patients s.t. each element $N_{jl} = x_l^{(j)}$ where $x_l^{(j)}$ is the value of l-th feature of non-recurrent patient j
D	matrix (size $(m+n) \times k$)	feature matrix of all patients s.t. $D = \begin{bmatrix} M \\ N \end{bmatrix}$
t_r (TTR)	variable	time-to-recur
t_n (DFS)	variable	disease-free survival time
t	variable	observed time s.t. $t = \min(t_r, t_n)$ t indicates either time-to-recur or disease-free survival time

3.1 Generalized Linear Regression model

Generalized Linear Regression model assumes features act multiplicatively on the function of time-to-recur t_r. This report fits two types of models that use different distributional assumptions. (See Table 2) Gaussian Linear Model specifies the log-transformed time-to-recur, $\log(t_r)$, as an addictive form of the features with Gaussian noise ϵ. Similarly, Gamma Generalized Linear Model [6] specifies the log-transformed mean of time-to-recur, $\log \mathbf{E}[t_r]$, as an addictive form of the features where

t_r follows Gamma distribution with mean $\exp(\beta_0 + \beta X)$ and variance $\theta \exp(\beta_0 + \beta X)$ where θ is a scale parameter.

Effect of features is specified with regression coefficients, β_0 and β, where β_0 is scalar and β is a vector with length k. The distributional parameters are defined in terms of features and regression coefficients. The regression coefficient parameters β_0, and β are estimated by maximizing log-likelihood function. The solution of of Gaussian Linear Model has a closed form while the solution of Gamma Generalized Linear Model is obtained via Iteratively Reweighted Least Squared method. [6] Note that Generalized Linear Regression model only uses the recur data which only have non-censored information to train the model.

Akaike information criterion (AIC) [7] and Bayesian information criterion (BIC) [8] are used for select an optimal subset of features via step-wise selection method. Likelihood Ratio test between two models constructed with AIC and BIC is performed to select a model to give a better fit of data.

Table 2: List of linear regression models with different distributional assumptions

Name	**Model**	**Distributional Assumption**
Gaussian Linear Model	$\log(t_r) = \beta_0 + \beta X + \epsilon$	$\epsilon \sim \mathcal{N}(0, \sigma^2)$ $\sigma^2 \in \mathbb{R}^+$ variance
Gamma Generalized Linear Model	$\log(\mathbf{E}[t_r]) = \beta_0 + \beta X$	$t_r \sim \Gamma(k, \theta)$ $k \in \mathbb{R}^+$ shape $\theta \in \mathbb{R}^+$ scale
Exponential AFT	$\log(t) = \beta_0 + \beta X + \epsilon$	$t \sim Exp(\lambda)$ $\lambda \in \mathbb{R}^+$ rate
Weibull AFT	$\log(t) = \beta_0 + \beta X + \epsilon$	$t \sim Weibull(\lambda, k)$ $\lambda \in \mathbb{R}^+$ scale $k \in \mathbb{R}^+$ shape
Log-normal AFT	$\log(t) = \beta_0 + \beta X + \epsilon$	$t \sim \mathbf{ln}\mathcal{N}(\mu, \sigma^2)$ $(\Leftrightarrow \epsilon \sim \mathcal{N}(0, \sigma^2))$ $\mu \in \mathbb{R}$ mean $\sigma^2 \in \mathbb{R}^+$ variance
Log-logistic AFT	$\log(t) = \beta_0 + \beta X + \epsilon$	$t \sim \mathbf{ln}Logistic(\mu, s)$ $(\Leftrightarrow \epsilon \sim Logistic(0, s))$ $\mu \in \mathbb{R}$ mean $s \in \mathbb{R}^+$ scale

3.1.1 Accelerated Failure Time model

Accelerated Failure Time (AFT) model [9] is alternative form of linear regression model that can handle the censored data. The AFT model specifies the log-transformed time (i.e. minimum value between time-to-recur and disease-free survival), $log(t)$, as an addictive form of the features. Four types of distributional forms are assumed for the noise ϵ, which implies different distributional forms of t: Exponential, Weibull, Log-normal, and Log-logistic. The parameters of each distribution are defined in terms of features and regression coefficients, β_0 and β. (See Table 2) Like as Generalized Linear Regression Model, effect of features is specified with regression coefficients where β_0 is scalar and β is a vector with length k. Note that AFT model uses both recur and non-recur data while Generalized Linear Regression model only use the recur data. Therefore, in order to make full use of both recur and non-recur data, regression coefficient parameters β_0, and β of AFT model will

be obtained maximizing partial-likelihood function [10] such that:

$$\mathbf{L} = \prod_{i=1}^{m} \lambda\left(t_r^{(i)}\right) S\left(t_r^{(i)}\right) \times \prod_{j=1}^{n} S\left(t_n^{(j)}\right) \tag{1}$$

Here $S(t)$ is *survival function* that is defined as:

$$S(t) = \mathbf{P}[T \geq t] = 1 - F(t) = \int_t^{\inf} f(x)dx$$

where $f(t)$ is probability density function of time, t. It gives the probability of being survived just before the time t. Also, $\lambda(t)$ is *hazard function* that is defined as:

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbf{P}[t \leq T < t + dt | T \geq t]}{dt}$$

It gives the instantaneous rate of occurrence of the event.

Feature selection for each model is performed same as Generalized Linear Regression model where both AIC and BIC are used and compared to find the best fit.

3.2 Recurrence Surface Approximation

RSA optimizes the following function:

$$\begin{aligned} min_{w,y,z,v} \frac{e^T y}{m} + \frac{e^T z}{n} + \delta \frac{e^T v}{m} \\ s.t. -v \leq Mw - t_r \leq y \\ -Nw + t_n \leq z \\ v, y, z \geq 0 \end{aligned}$$

y is a vector of the violations for each prediction, which is simply the time difference between predictions and real time if the prediction is later than time-to-recur. v is also a vector accounting for the violation if the predicted time is earlier than time-to-recur. This is to prevent the predictions to be too early, but it is multiplied by a small number δ to make the boundary weaker. z is the vector of violations when the prediction is earlier than disease-free survival. e is just a vector of ones.

3.3 Prediction error

The error definition used by Mangasarian et al. [2] is over simplified in the sense that the error of early predictions of TTR were not accounted for. Therefore we modified the error definition as follows.

Prediction error on recur data is defined as a weighted mean of absolute differences between TTR (observed) and TTR (predicted). Note that true value of TTR is located between the time of last visit and the observed TTR. Average range between two consecutive visits, c, is estimated as average range between two consecutive observed time, t. Therefore, the weight would be 0 if predicted TTR is located within c left from observed TTR, otherwise 1. Therefore prediction error on recur data is:

$$PE_r = \frac{\sum_{i=1}^{m} \left| t_r^{(i)} - \widehat{t_r^{(i)}} \right| \mathbb{1}_{\left\{ t_r^{(i)} - \widehat{t_r^{(i)}} > c \text{ or } t_r^{(i)} - \widehat{t_r^{(i)}} \leq 0 \right\}}}{m}$$

Prediction error on non-recur data is defined as a weighted mean of absolute differences between DFS (observed) and TTR (predicted). Note that TTR is always larger than DFS but TTR is not observed for non-recurrent patients. Therefore, the weight would be 0 if the predicted TTR is larger than observed DFS, otherwise 1. Therefore prediction error on non-recur data is:

$$PE_n = \frac{\sum_{j=1}^{n} \left| t_n^{(j)} - \widehat{t_r^{(j)}} \right| \mathbb{1}_{\left\{ t_n^{(j)} > \widehat{t_r^{(j)}} \right\}}}{n}$$

The total prediction error is defined as a weighted average of the prediction error on recur and non-recur data:

$$PE = \frac{(error_r \times m) + (error_n \times n)}{m + n}$$

4 Experiment

4.1 Data

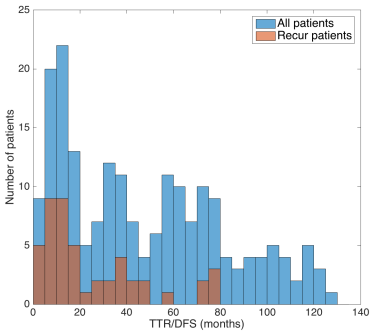

Figure 1: Histogram of the recur and all data, x-axes represents observed time t and y-axes represents the number of patients.

The Wisconsin Prognostic Breast Cancer (WPBC) data set [11] contains 198 (47 recurrence / 151 non-recurrence) patients and 32 features taken from breast cancer microscopic instances. Of those 30 features obtained from a digitized image of a fine needle aspirate of a breast mass and two additional features about tumor size and lymph node status. As shown in Figure 1, observed time t of all 198 patients varies from 0 to 125 months while TTR (t_r) of 47 recurrent patients varies within 78 months. Both distributions are left skewed.

The data has four missing values in *Lymph Node Status*. The missing values are imputed via K-nearest neighborhood (KNN, [12, 13]) method with $K = 10$. For each samples, it finds the 10 nearest neighbors using a Euclidean metric, then imputes the missing values by averaging values of *Lymph Node Status* of its neighbors.

4.2 Result

4.2.1 Generalized Linear Regression model

Gaussian Linear Model and Gamma Generalized Linear Model was applied to the data. The likelihood ratio test ($\alpha = 0.05$) to compare the goodness of fit of models said that the data are more likely under the model with BIC step-wise selection than the model with AIC step-wise selection. Therefore, we choose the reduced model obtained from the BIC step-wise selection. Table 3 is a list of features selected via step-wise selection method using BIC. Prediction error of the model is evaluated as defined in section 3.3 (See Table 5)

4.2.2 Accelerated Failure Time model

AFT models under different distributional assumptions was applied to the data. The likelihood ratio test ($\alpha = 0.05$) to compare the goodness of fit of models said that the data are more likely under the model with AIC step-wise selection than the model with BIC step-wise selection. Therefore, we choose the reduced model obtained from the AIC step-wise selection. Table 3 is a list of features selected via step-wise selection method using AIC. Prediction error of the model is evaluated as defined in section 3.3 (See Table 5). A function `survreg()` in R-package named *survival* [14, 15] is used to fit the models.

4.2.3 Recurrence Surface Approximation

Because we will use a different error definition than what Mangasarian et al. [2] was using, we tried to reproduce their results and convert the errors once the reproduced results were acceptable. We

Table 3: Selected features via the BIC step-wise selection method

	Features
Gaussian Linear Model	*radius2 + radius3 + texture2 + perimeter1 + perimeter2 + perimeter3 + area1 + area2 + smoothness1 + smoothness2 + compactness2 + compactness3 + concavity1 + concavePoints1 + concavePoints2 + concavePoints3 + symmetry1 + fractalDimension2 + fractalDimension3*
Gamma Generalized Linear Model	*radius2 + radius3 + texture2 + perimeter1 + perimeter2 + perimeter3 + area1 + area2 + smoothness1 + smoothness2 + compactness2 + compactness3 + concavity1 + concavePoints1 + concavePoints2 + concavePoints3 + symmetry1 + fractalDimension2 + fractalDimension3*
Exponential AFT	*radius1 + radius3 + texture2 + perimeter1 + area1 + area2 + area3 + compactness1 + compactness2 + symmetry2 + symmetry3 + lymphNodeStatu*
Weibull AFT	*radius1 + radius3 + texture2 + perimeter1 + area1 + area2 + area3 + compactness1 + compactness2 + symmetry2 + symmetry3 + lymphNodeStatus*
Log-normal AFT	*radius1 + radius3 + perimeter3 + area1 + area2 + area3 + smoothness2 + compactness1 + compactness2 + compactness3 + concavity1 + symmetry1 + symmetry2 + symmetry3 + lymphNodeStatus*
Log-logistic AFT	*radius1 + radius3 + perimeter3 + area1 + area2 + area3 + smoothness2 + compactness1 + compactness2 + compactness3 + concavity1 + symmetry1 + symmetry2 + symmetry3 + lymphNodeStatus*

implemented RSA and tested it on the data of 198 patients. The results were summarized in the Table 4. Note that We arbitrarily set $\delta = 0.1$, since it was not reported in the paper.

Table 4: Prediction errors of RSA via cross validation

	All	**Recur**	**Non-recur**
Mangasarian et al.	18.3	13.0	19.9
Leave-one-out	27.9	9.9	33.4
Ten-fold	28.4	11.6	33.4

4.2.4 Comparison of prediction accuracy

We calculated the prediction error of each aforementioned model based on the error definition in section 3.3. The cutoff for early predictions c was set to 1, because the estimated average range between two consecutive visits is $\hat{c} = 1.32$. As summarized in Table 5, Generalized Linear Model achieved a good prediction accuracy on the recur data, but a poorer but acceptable accuracy on the non-recur data; AFT models achieved an excellent accuracy on the non-recur patients but an extremely poor accuracy on the recur data; and the predictions from RSA were acceptable in general but the error on all patients was the poorest.

5 Summary

How to handle censored data properly to serve the purpose of improving prediction accuracy is the question we attempt to tackle in this project. In this midterm report, we applied several fundamental methods, e.g. Generalized Linear Models and AFT models, to serve as benchmarks. We also

Table 5: Prediction errors via 10-fold cross validation

	All	**Recur**	**Non-recur**
Gaussian Linear Model	30.9	5.3	38.6
Gamma Generalized Linear Model	30.6	5.2	38.3
Exponential AFT	28.2	104.2	1.8
Weibull AFT	24.8	92.2	1.8
Log-normal AFT	21.3	75.6	2.6
Log-logistic AFT	21.0	73.6	2.8
RSA	32.1	22.1	35.1

reproduced a reported method called RSA which was designed to address the issue [2]. We then compared the models by evaluating the selected features and prediction performance in the cross validation.

Generalized Linear Models were trained only on the recur data, therefore it was expected that they predicted much better on recurrent patients than other models including RSA (as shown in section 4.2.4). However, what is interesting is that Generalized Linear Models also showed similar performance on the non-recur data as compared to RSA, which was trained on both recur and non-recur data. A naive conclusion could be drawn that the censored data compromised the accuracy of RSA instead of improving it. Nonetheless, note that the RSA we used was a reproduction based on what was described by Mangasarian et al., therefore it may not be the optimal performance of RSA. In any case, our task remains the same, which is to explore a better usage of the censored data.

AFT models were more likely to predict TTR longer than other models. It led AFT to have higher prediction error on the recur data while lower prediction error on the non-recur data compared to other models. This is because AFT models estimated regression coefficients by maximizing the partial likelihood function (see equation (1)), not by minimizing training prediction error. Therefore, we will remove AFT model from our list.

Tumor type/grade, tumor size and lymph node status are reported to be important factors of breast cancer prognosis [16]. The first 30 features represent cellular nuclei size, shape and texture, which are related to tumor type. The Generalized Linear Models only selected features of cellular nuclei size, shape and texture, whereas AFT models selected cellular nuclei size, shape, texture and lymph node status, but none of the models selected tumor size.

References

[1] Pedro Ferreira, Nuno A Fonseca, Inês Dutra, Ryan Woods, and Elizabeth Burnside. Predicting malignancy from mammography findings and surgical biopsies. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 339–344. IEEE, 2011.

[2] Olvi L Mangasarian, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[3] W Nick Street, Olvi L Mangasarian, and William H Wolberg. An inductive learning approach to prognostic prediction. In *ICML*, pages 522–530. Citeseer, 1995.

[4] Ioannis Anagnostopoulos and Ilias Maglogiannis. Neural network-based diagnostic and prognostic estimations in breast cancer microscopic instances. *Medical and Biological Engineering and Computing*, 44(9):773–784, 2006.

[5] Simone A Ludwig and Stefanie Roos. Prognosis of breast cancer using genetic programming. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 536–545. Springer, 2010.

[6] John A Nelder and R Jacob Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[7] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec 1974.

[8] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[9] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[10] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[11] M. Lichman. UCI machine learning repository, 2013.

[12] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[13] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays, 1999.

[14] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

[15] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

[16] Mary Cianfrocca and Lori J Goldstein. Prognostic and predictive factors in early-stage breast cancer. *The oncologist*, 9(6):606–616, 2004.