An Expectation-Maximization Based Machine Learning Strategy to Predict Recurrence Time 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

Abstract

Machine learning methods are promising in predicting the time when breast cancer is likely to recur in patients after surgery. Most of previous works have been trying to handle the challenge of right censored data. In this project, we proposed an improved EM algorithm called EM-Weighted LSE and illustrated its performance with Wisconsin Prognostic Breast Cancer (WPBC) dataset. An expected, complete log-likelihood function was proposed to utilize both recur and non-recur data, as well as a new definition of error. Ten-fold cross validation results shown that EM-Weighted LSE outperformed two baseline methods (Gaussian Linear Model and Recurrence Surface Approximation) in terms of errors for all data, non-recur data, as well as recur data respectively. The results indicated that EM-Weighted LSE is applicable in handling right-censored data, and is able to predict an accurate recurrence time for breast cancer patients.

1 Introduction

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.

To predict the time-to-recur (TTR) after surgery of breast cancer, one of the major challenges 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.

RSA is a linear programming technique which tries 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. We use RSA as a baseline method in our project, the other baseline method is Gaussian Linear Model, which outputs expected TTR as a weighted sum of features. Only recur data is used to construct the model. It assumes the error of TTR follow a Gaussian distribution with mean zero and a constant variance.

In this project, we proposed an algorithm called EM-Weighted LSE to make full use of both recur and non-recur data as well as handle the right censored response. To full use of both recur and non-recur data, we proposed an expected, complete log-likelihood based on a distribution of

$y = \log(TTR) - \log(DFS)$, as well as a new definition for the error. The expected, complete log-likelihood served as an object function in the first step of EM algorithm. Bayesian information criterion (BIC) [6] are used to select an optimal subset of features via step-wise selection method. Then we use Expectation-Maximization (EM) algorithm to update non-recur data and make a prediction for new samples. Wisconsin Prognostic Breast Cancer (WPBC) data set [7]) was used to illustrate our approach. Ten fold cross validations are used to calculate the errors and compare our proposed with baseline methods.

2 Method

To make full use of labeled and unlabeled data, we maximize log likelihood of Gaussian Linear Regression model using both labeled and unlabeled data. In section 2.1, we define the complete log likelihood function and its expectation to be maximized, as well as a new error definition that accommodates to it. Section 2.2 shows that the optimization problem in section 2.1 can be rewritten as Weighted Least Squares Estimation (Weighted LSE) which is a special case of least square estimation when the variances of the observation can be unequal. The output from Weighted LSE can be used as an input of the next iteration in EM algorithm. Section 2.3 explains how to utilize EM approach to our problem setting.

Table 1: List of notations.

Symbol	Type	Description
k	scalar	the number of samples
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 for each feature
TTR	variable	time-to-recur ($t^r = TTR^*$ is for log-scaled)
DFS	variable	disease-free survival time ($t^n = DFS^*$ is for log-scaled)

2.1 Likelihood and Error

For each patient, we observe either TTR or DFS: TTR for recurrent patients, and DFS for non-recurrent patients. Note that the domain of Gaussian Linear Regression Model is one dimensional real space, while the observations are positive real values. Therefore, we transform the original observations (TTR or DFS) into log scale such that $TTR^* = \log(TTR)$ and $DFS^* = \log(DFS)$. For non-recurrent patients, we can only observe DFS^* and obtain $\widehat{TTR^*}$ from Gaussian Linear Model. Therefore the error we actually obtain for non-recurrent patients is ϵ^* instead of the error ϵ which we want to minimize. For non-recurrent patients, we use TTR^* to indicate the (non-observed, log-scaled) true time-to-recur. Then, for non-recurrent patients we define:

$$y = TTR^* - DFS^*, \quad \epsilon = TTR^* - \widehat{TTR^*}, \quad \epsilon^* = \widehat{TTR^*} - DFS^*$$

Now we can write the error ϵ in terms of y and ϵ^*:

$$\epsilon = TTR^* - \widehat{TTR^*} = (TTR^* - DFS^*) - (\widehat{TTR^*} - DFS^*) = y - \epsilon^*$$

Therefore, the complete log-likelihood using both recurrent and non-recurrent patients is:

$$\begin{aligned}\log \mathcal{L}(\beta, y) &= -\frac{\sum_{i \in Recur} \epsilon_i^2}{2\sigma^2} - \frac{\sum_{i \in Non\ Recur} \epsilon_i^2}{2\sigma^2} - \frac{m}{2}\log(2\pi\sigma^2) - \frac{n}{2}\log(2\pi\sigma^2) \\ &= -\frac{\sum_{i \in Recur}(t_i^r - \widehat{t_i^r})^2}{2\sigma^2} - \frac{\sum_{i \in Non\ Recur}(y_i - \epsilon_i^*)^2}{2\sigma^2} - \frac{m+n}{2}\log(2\pi\sigma^2)\end{aligned}$$

Noting that we cannot observe y, by taking expectation on the complete log-likelihood in terms of the conditional distribution of y given DFS^* we get:

$$\begin{aligned}\log \mathcal{L}(\beta) &= \mathbf{E}_{y|t^n}(\log \mathcal{L}(\beta, y)) \\ &= -\frac{\sum_{i \in Recur}(t_i^r - \widehat{t_i^r})^2}{2\sigma^2} - \mathbf{E}_{y|t^n}\left[\frac{\sum_{i \in Non\ Recur}(y - \epsilon_i^*)^2}{2\sigma^2}\right] - \frac{m+n}{2}\log(2\pi\sigma^2) \\ &= -\frac{\sum_{i \in Recur}(t_i^r - \widehat{t_i^r})^2}{2\sigma^2} - \int_y \frac{\sum_{i \in Non\ Recur}(y - \epsilon_i^*)^2}{2\sigma^2} f(y|t^n)dy - \frac{m+n}{2}\log(2\pi\sigma^2)\end{aligned}$$

Here we assume the conditional distribution of y does not depends on features. Then $f(y|t^n)$ can be obtained by:

$$\begin{aligned}f(y|t^n) &= \mathbf{P}[t^r - t^n = y | t^r > t^n] \\ &= \mathbf{P}[t^r = y + t^n | t^r > t^n] \\ &= \lim_{\Delta t \to 0} \frac{\mathbf{P}[t^n + y < t^r \leq t^n + y + \Delta t]}{\mathbf{P}[t^r > t^n]} \\ &= \frac{h(\exp(t^n + y))}{S(\exp(t^n))} \\ &= \frac{h(DFS \times \exp(y))}{S(DFS)}\end{aligned}$$

where $h(\cdot)$ is the hazard function; $S(\cdot)$ is the survival function. The hazard function $h(\cdot)$ can be estimated by smooth estimate of the hazard function proposed by K. Hess and Gentleman (2014, [8]); and the survival function $S(\cdot)$ can be estimated by Kaplan-Meier estimator. [9]

The error definition used by Mangasarian et al. [2] is over simplified since they set the errors to be zero for any predictions smaller than observed TTR. However it is highly possible that the real recur time is slightly ahead of the observed TTR. Therefore we proposed a more strict error definition based on the expected, complete log-likelihood function. Prediction error of recur data is defined as an average of squared differences between TTR^* (observed) and $\widehat{TTR^*}$ (predicted). Note that true value of TTR^* is located between the last times of visiting hospital. Average range between two consecutive visits, c, is estimated as average range between two consecutive observed time, t. Therefore, when predicted TTR is located within c left from the observed TTR^*, it is considered to be a reasonable prediction with an error set to be zero. The new error of recur data can be expressed as follows:

$$error_r^* = \sum_{i=1}^{m} \frac{\left(t_i^r - \widehat{t_i^r}\right)^2 \mathbb{1}_{\left\{t_i^r - \widehat{t_i^r} > c \text{ or } t_i^r - \widehat{t_i^r} \leq 0\right\}}}{m}$$

Prediction error on non-recur data is defined as an average of squared differences between TTR^* (true, but not observed) and $\widehat{TTR^*}$ (predicted). Since TTR^* is not observed for non-recurrent patients, we replace ϵ with y and ϵ^*. Its weighted sum in terms of y would be prediction error on non-recur data:

$$error_n^* = \sum_{i=1}^{n} \sum_{y} \frac{\left(y_i - \left(\widehat{t_i^r} - t_i^n\right)\right)^2}{n} f(y|t_i^r)$$

Since the sum of squared error defined above is log-scaled, we use inverse-transform to get an error with original scale such that $error = \exp(\sqrt{error^*})$. Therefore, the total prediction error under the original scale is defined as a weighted average of the prediction error on recur and non-recur data:

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

2.2 Parameter Estimation via Weighted Least Squares

In this section, we will shows that the optimization problem in section 2.1 can be re-written as Weighted Least Squares Estimation (Weighted LSE) which is a special case of least square estimation when the variances of the observation can be unequal. To calculate the expected error on non recur data, we consider y as a discrete variable and take a weighted average of the errors on its distribution instead of solving integral on a continuous variable. Therefore, for enough number of $y \in 1, 2, \cdots, n_y$, the expected, complete log-likelihood function can be re-written as:

$$\begin{aligned}\log \mathcal{L}(\beta) &= -\sum_{i \in Recur} \frac{(t_i^r - \widehat{t_i^r})^2}{2\sigma^2} - \sum_{y=1}^{n_y} \sum_{i \in Non\ Recur} \frac{(y - \epsilon_i^*)^2}{2\sigma^2} f(y|t^n) + (constant) \\ &= -\sum_{j=1}^{m+n\times n_y} \frac{\left(t_j^{(new)} - X_j\beta\right)^2}{2\sigma^2} w(j) + (constant)\end{aligned}$$

where

$$t_j^{(new)} = \begin{cases} t_j^r & \text{if } j \text{ is recur data (i.e. } g(j) = j) \\ t_i^n + y(j) & \text{if } j \text{ is non-recur data such that } g(j) = i \end{cases}$$

$$w(j) = \begin{cases} 1 & \text{if } j \text{ is recur data (i.e. } g(j) = j) \\ f(y(j)|t_i^r) & \text{if } j \text{ is non-recur data such that } g(j) = i \end{cases}$$

$$X_j = \begin{cases} X_j & \text{if } j \text{ is recur data (i.e. } g(j) = j) \\ X_i & \text{if } j \text{ is non-recur data such that } g(j) = i \end{cases}$$

$$g(j) = \begin{cases} j & \text{if } j \text{ is recur data} \\ i & \text{if } j \text{ is non-recur data such that } j \in \{m + n_y(i - m - 1) + 1, \cdots m + n_y(i - m - 1) + n_y\} \end{cases}$$

$$y(j) = m + n_y(i - m - 1) + j$$

Therefore the expected, complete log-likelihood is same as a log-likelihood function for a linear regression model with response as $t_j^{(new)}$ and explanatory variable as X_j where $j = 1, \cdots, m + (n \times n_y)$ under a heterogeneous variance distribution assumption on the error such that $\epsilon_j \sim \mathcal{N}\left(0, \sigma^2/w(j)\right)$. Note that we make n_y copies of each recurrent sample, assign a new response for each copy by adding a different value of $y \in \{1, \cdots, n_y\}$ to its old response, and give a weight as a function of y. Therefore, we have a closed form of the optimization problem in section 2.1 for Gaussian Linear Regression Model, as a weighted Gaussian Linear Regression Model has.

2.3 Expectation-Maximization

In this section we explain how to utilize EM approach to our problem setting. In the E-step, the probability distribution of the unknown data can be given by:

$$\begin{aligned}&\mathbf{P}(Z|X, \beta^{old}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Z - X\beta^{old})^2}{2\sigma^2}\right) \\ \Rightarrow\ &\mathbf{E}(Z|X, \beta^{old}) = X\beta^{old} \cdot I(Z > DFS) + DFS \cdot I(Z \leq DFS)\end{aligned}$$

Where $Z = \ln(TTR)$ for non-recurrent patients; $I(\cdot)$ is an indicator function.

In M-step, the expected log-joint-likelihood can be given by:

$$\begin{aligned}\mathbf{E}_{Z|X,\beta^{old}}\left[\sum_{i=1}^{m+n}\ln P(Z_i,X_i|\beta)\right] &= \mathbf{E}_{Z|X,\beta^{old}}\left[\sum_{i=1}^{m+n}\ln \mathbf{P}(Z_i|X_i,\beta)\mathbf{P}(X_i|\beta)\right] \\ &= \mathbf{E}_{Z|X,\beta^{old}}\left[-\frac{\sum_{i=1}^{m+n}(Z_i-X_i\beta^{old})^2}{2\sigma^2}+\ln\frac{1}{\sqrt{2\pi\sigma^2}}+\sum_{i=1}^{m+n}\ln\mathbf{P}(X_i)\right]\end{aligned}$$

Since the distribution of features X is independent of β, we have $P(X|\beta) = P(X)$. The last two terms would be a constant in terms of β, therefore the following optimizations are equivalent:

$$\begin{aligned}\beta &\leftarrow \underset{\beta}{\operatorname{argmax}}\,\mathbf{E}_{Z|X,\beta^{old}}\left[\sum_{i=1}^{m+n}\ln\mathbf{P}(Z_i,X_i|\beta)\right] \\ &\leftarrow \underset{\beta}{\operatorname{argmax}}\,\mathbf{E}_{Z|X,\beta^{old}}\left[-\frac{\sum_{i=1}^{m+n}(Z_i-X_i\beta)^2}{2\sigma^2}\right] \\ &\leftarrow \underset{\beta}{\operatorname{argmin}}\sum_{i=1}^{m+n}\left(\mathbf{E}_{Z|X,\beta^{old}}[Z_i]-X\beta\right)^2\end{aligned}$$

Note that Leibniz integral rule tells us that in the last two steps, the objective functions are identical. In addition, as shown the algorithm is essentially minimizing the difference between prediction from previous and current predictions, and we impose the truth of non-recurrent patient to the seemingly self-proven process by disallowing the prediction to be smaller than DFS.

The scheme of the simple EM algorithm would be:

(i) Train the linear regression model on the TTR of recur patients to obtain the initial parameters $\beta^{(0)}$;

(ii) Predict the TTR of non-recur patients. If the predicted TTR is larger than DFS, then replace the DFS with the predicted time; otherwise use the DFS directly as the possible TTR;

(iii) Obtain new parameters $\beta^{(n)}$ by re-training the model on all patients in the training dataset with their guessed TTR (predicted or the DFS).

(iv) Go to step (ii) until convergence.

We also designed an improved EM algorithm called EM-Weighted LSE (Weighted LSE is described in section 2.2). The improved version minimize the errors with the more sophisticated error definition described in section 2.1. Therefore theoretically the new objective function should be more sensitive to non-recurrent errors. The scheme of the EM-Weighted LSE would be similar to the simple EM algorithm, except that in the first iteration we use Weighted LSE to train the model to accommodate the new error definition. The output from Weighted LSE can be used as an input of the next iteration in the original EM algorithm. Note that from the next iterations we will switch back to the original linear model because all the patients will have either an observed TTR or guessed TTR by then, and therefore the error can be measured directly by the difference between prediction and observation/guess.

3 Results

3.1 Imputing Missing Data

The Wisconsin Prognostic Breast Cancer (WPBC) data set [7] 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. The data has four missing values in *Lymph Node Status*. The missing values are imputed via K-nearest neighborhood (KNN, [10, 11]) method with $K = 10$.

3.2 Feature Selection

Features used in our approach are selected based on the simple Gaussian Linear Model. Bayesian information criterion (BIC) [6] are used for select an optimal subset of features via step-wise selection method. The selected features are shown as follows:

radius2, radius3, texture2, perimeter1, perimeter2, perimeter3, area1, area2, smoothness1, smoothness2, compactness2, compactness3, concavity1, concavePoints1, concavePoints2, concavePoints3, symmetry1, fractalDimension2, fractalDimension3

In the list, "1" represents mean value, "2" represents variance, and "3" represents extreme value.

3.3 Comparison of Expectation-Maximization vs. other models

Due to the limitation on the number of examples, ten-fold validation was used to quantitatively evaluated the performance of the models. Total 198 patients were randomly assigned to 10 groups, and in each iteration the models will be tested on one group and trained on other groups. Therefore the validation result is a reflection of the real performance of the models. The prediction errors from ten-fold validation for linear regression, RSA, simple EM, and EM-weighted LSE are shown in Table 2.

Table 2: Comparison of errors in different models

Model	Error for all	Error for non-recur	Error for recur
GLM	1.2735	1.2928	1.2675
GLM (BIC)	1.2283	1.2599	1.1267
RSA	1.1416	1.1132	1.2328
EM	1.1267	1.0842	1.2632
EM-wLSE	1.1094	1.0757	1.2175
EM (BIC)	1.1093	1.0720	1.2294
EM-wLSE (BIC)	1.1058	1.0730	1.2113

In order to satisfy the assumption of linear regression, the observations (TTR and DFS) are log-transformed so that the difference between prediction and true values are actually difference of logarithms (see section 2.1). Therefore the prediction errors are essentially the "absolute" ratio of prediction to true value or the inverse depending on which value is larger (i.e. the absolute ratio is always above 1). We trained two Gaussian Linear Models by using recurrent patients to serve as a baseline for comparison. The models marked by BIC only use 18 features listed in section 3.2.

In general, a model is better if its error is more close to 1. All EM methods achieved lower error/higher prediction accuracy than simple linear regressions and RSA on all patients. It indicates that the extra information added by unlabeled data is beneficial. Note that our error definition penalize large TTR predictions on non-recurrent patients in contrast to zero penalty used by RSA [2], therefore smaller non-recur errors do **not** imply that the models tend to predict large TTR. In addition, improved EM performed better than simple EM, and BIC models performed better than models using all features. It indicates that the selected features could be the most important ones to describe the development of breast cancers.

4 Conclusion

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 project, we proposed EM-Weighted LSE to handle this problem. To full use of both recur and non-recur data, we proposed an expected,

complete log-likelihood based on a distribution of $y = \log(TTR) - \log(DFS)$, as well as a new definition for the error. In the new error definition, we penalize large TTR prediction on non-recurrent patients based on the expected survival time of a patient after DFS. The stricter error definition enables a more accurate objective function to optimize, and results in the improvement of prediction of non-recurrent patients. The log-likelihood function is used as an objective function to be maximized in general setting of Gaussian Linear Regression Model. Features are selected based on BIC. [6] The output from the model is used as an input of the EM algorithm. EM algorithm updates non-recur data and make a prediction for new samples. Our approach is illustrated using Wisconsin Prognostic Breast Cancer (WPBC) data set [7]. The errors are calculated based on the definition proposed in section 2.1.

The choice of linear regression as the fundamental method used as both baseline and in EM is because RSA assumes the linear relationship between features and TTR. More importantly, the flexible nature of linear regression allowed us to integrate it with our probabilistic error definition, which led to the improved version of EM. Moreover, linear regression does not have the problem of local optima, which could be a risk if we deploy complicated non-linear models, e.g. Neural Networks.

Ten-fold cross validation results showed that EM-Weighted LSE performed better than a baseline Gaussian Linear Model and RSA in terms of errors for all patients, non-recur patients, as well as recur patients respectively. The results supported that EM-Weighted LSE is able to make more accurate predictions of recurrence time for breast cancer patients. Several potential important prognostic factors were also identified through feature selection. In addition, the idea of EM-Weighted LSE is applicable to various fields with censored data that is not limited to clinical data. Actually the breast cancer data we used is rather a small dataset, but the proposed method is promising in future applications on larger datasets.

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

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

[8] S original by Kenneth Hess and R port by R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*, 2014. R package version 1.2.6.

[9] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[10] 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.

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