T9 - A Topic Model of Genetic Mutations in Cancer

Peng Xu
Department of Computer Science
University of Alberta
pxu4@ualberta.ca

Anqi Jing
Department of Electrical & Computer Engineering
University of Alberta
ajing@ualberta.ca

Abstract

Survival prediction of cancer patients is significant because it can provide useful information for patients and doctors. Several survival prediction models, like multi-task logistic regression (MTLR) have been proposed by using clinical and tumor information. Since gene expression values can reveal the patient status, we aim to improve the accuracy of survival prediction by using gene expression data of patients. In this paper, Latent Dirichlet Allocation model is employed to extract topic features from high dimensional gene expression data by using different methods. Then, MTLR model is constructed by using the derived topic features and some other features as the input. By adding gene expression information, the improvement of 0.27 on concordance score and 0.13 on log likelihood loss has been achieved over the baseline, which suggests that gene expression information can really help improve the survival prediction.

1 Introduction

1.1 Motivation and Task

Survival predictions of cancer patients are of great significance because patients and doctors need such information to make decision about treatment and end-life-care. Unlike survival analysis, survival prediction focuses on the evaluation on individuals rather than populations, which attempts to predict the accurate patient-specific survival probability within a specific time. In this project, we aim to improve the performance of survival prediction model for breast cancer patients by using gene expression data.

1.2 Related Work

By using machine learning techniques, the survival prediction model can be learned that uses patient-specific features. One of the most common approaches for survival analysis is Cox proportional Hazards. It can be used to discover attributes that are relevant to survival and predict outcomes. However, the Cox model [3] works with the hazard function, and is not designed to fit the individual specific survival time. It is typically designed for the predicting survival probability from a large heterogenous population. Another model, multi-task logistic regression (MTLR) [2] has been proposed for the task of patient-specific predictions. MTLR model can be viewed as a sequence of dependent regressors, which combines multiple local logistic regression models. It can directly work with the survival function and handle censored samples naturally. In this paper, the MTLR model learns more than 2000 breast cancer patients and uses clinical attributes such as clinical assessments and blood tests. The results show that such information can improve the accuracy of survival prediction compared to using cancer site and stage only.

1.3 Our Approach

Marc J. Van et al [4]. identified a gene expression profile that is associated with prognosis in patients with breast cancer. The result shows that the gene expression profile of breast cancer is a clinically meaningful tool. Since gene expression values can reveal the patient status, we aim to improve the accuracy of survival prediction by using gene expression data of patients. The main challenge in our project is that how to handle the gene expression data for each patient. Obviously, we cannot take the gene expression profiles as the input to the survival prediction model directly, because the high dimensionality of gene expression will increase the time complexity and cause overfitting, and also, there exists the useless information in gene expression data.

Several feature reduction methods have been used for the survival analysis model. Latent Dirichlet allocation (LDA) [7], a probabilistic model, is an efficient tool for accommodating heterogeneity, selecting features, and characterizing complex interactions in a high dimensional textual setting. It can identify a set of topics that co-occur frequently across large sets of words, and associated distributions of the topics over words and documents. In our project, LDA is applied to gene expression data by considering each patient as a "document" with "words" characterizing genomic states of the patient. By this way, derived topics and associated distributions over each document can characterize the genomics states of patients, and the distributions can be used as the features to predict the survival of patients. Since MTLR fit well with our task, then MTLR is applied to perform patient-specific survival predictions by using extracted features. Our results show that the performance of predictor is improved by using the gene expression data.

The outline of our paper is as follows. Section 2 briefly review the two models: LDA and MTLR. Section 3 describes our methods on document construction and parameter selection. Section 4 describes the details of experiments we've done, including evaluation metrics, dataset, results and analysis. Finally we summarize our work and potential further work in section 5.

2 Background

Latent Dirichlet Allocation (LDA) as detailed in Blei et al. [1] and Multi-Task Logistic Regression (MTLR) as detailed in Lin H, Baracos V, Greiner R, et al. [2] are heavily used in our project. Here, we give a brief review of these two models.

2.1 Latent Dirichlet Allocation (LDA)

Assume there are N documents indexed by $i = 1, \ldots, N$, each of which consists of N_i words. The vocabulary is the unique set of length V indexed by $v = 1, \ldots, V$, from which the documents' words arise, and is usually taken to be the union of all words over documents. Further assume that there are K latent 'topics' indexed by $k = 1, \ldots, K$, that govern the assignment of words to documents. The process of generating a corpus is as follows:

- Pick a multinomial distribution ϕ_z for each topic z from a Dirichlet distribution with parameter β;
- For each document d, pick a multinomial distribution θ_d from a Dirichlet distribution with parameter α;
- For each word token w in doucment d, pick a topic $z \in \{1, \ldots, K\}$ from the multinomial distribution ϕ_z;
- Pick word w from the multinomial distribution θ_z.

The LDA model is represented as a probabilistic graphical model in Figure 1. Using variational expectation-maximization (EM) algorithm, we can obtain optimal posterior estimates for $\phi_{1:K}$ and $\theta_{1:N}$ upon convergence.

2.2 Multi-Task Logistic Regression (MTLR)

Consider modeling the probability of survival of patients at each of a vector of time points $\tau = (t_1, t_2, \ldots, t_m)$-e.g., τ can be the 60 monthly intervals from 1 month up to 60 months. We can set up a series of logistic regression models for each of these:

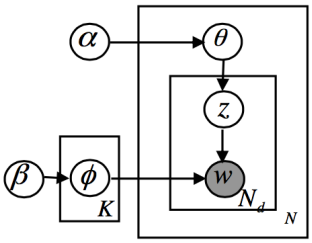

Figure 1: Graphical model representation of LDA. K is the number of topics; N is the number of documents; N_d is the number of word tokens in document d.

$$P_{\vec{\theta}_i}(T \geq t_i | \vec{x}) = \left(1 + \exp(\vec{\theta}_i \cdot \vec{x} + b_i)\right)^{-1}, \quad 1 \leq i \leq m,$$

where $\vec{\theta}_i$ and b_i are time-specific parameter vector and thresholds.

The model encodes the survival time s of a patient as a binary sequence $y = (y_1, y_2, \ldots, y_m)$, where $y_i \in \{0, 1\}$ denotes the status of the patient at time t_i, so that $y_i = 0$ (no death event yet) for all i with $t_i < s$, and $y_i = 1$ (death) for all i with $t_i \geq s$. We denote such an encoding of the survival time s as $y(s)$, and let $y_i(s)$ be the value at its ith position. The probability of observing the survival status sequence $y = (y_1, y_2, \ldots, y_m)$ can be represented by the following generalization of the logistic regression model:

$$P_\Theta(Y = (y_1, y_2, \ldots, y_m) | \vec{x}) = \frac{\exp(\sum_{i=1}^m y_i(\vec{\theta}_i \cdot \vec{x} + b_i))}{\sum_{k=0}^m \exp(f_\Theta(\vec{x}, k))},$$

where $\Theta = (\vec{\theta}_1, \ldots, \vec{\theta}_m)$, and $f_\Theta(\vec{x}, k) = \sum_{i=k+1}^m (\vec{\theta}_i \cdot \vec{x} + b_i)$ for $0 \leq k \leq m$ is the score of the sequence with event occuring in the interval $[t_k, t_{k+1})$ before taking the logistic transform, with the boundary case $f_\Theta(\vec{x}, m) = 0$ being the score for the sequence of all '0's.

Therefore the log likelihood of a set of uncensored patients with survival time $s_1, s_2, \ldots, s_n$ and feature vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ is

$$\sum_{i=1}^n \left[\sum_{j=1}^m y_j(s_i)(\vec{\theta}_j \cdot \vec{x}_i + b_j) - \log \sum_{k=0}^m \exp f_\Theta(\vec{x}_i, k) \right].$$

For handling censored data, this model can handle censoring naturally by marginalizing over the unobserved variables in a survival status sequence $(y_1, y_2, \ldots, y_m)$. For example, suppose a patient with features $\vec{x}$ is censored at time s_c, and t_j is the closest time point after s_c. Then all the sequences $y = (y_1, y_2, \ldots, y_m)$ with $y_i = 0$ for $i < j$ are consistent with this censored observation. The likelihood of this censored patient is

$$P_\Theta(T \geq t_j | \vec{x}) = \frac{\sum_{k=j}^m \exp(f_\Theta(\vec{x}, k))}{\sum_{k=0}^m \exp(f_\Theta(\vec{x}, k)))},$$

where the numerator is the sum over all consistent sequences.

3 Methods

3.1 Document Construction

In our project, how to construct the documents from the gene expressions is one of the main problems. There are several challenges here. First, word occurrences in natural corpus are integer, but

gene expression values are real numbers. Previous study conducted by other students shows that converting the gene expression values directly into word occurrences performs poorly. Second, the distribution of gene expression values is different from the distribution of word occurrences in natural corpus. Finally, it's unclear how to capture the effective information in gene expression values.

In order to capture the effective information in gene expression values, we propose five different document construction methods:

3.1.1 Whisker

A patient's document received a gene word, given by the gene name, if that patient shows extreme expression for that gene. If the gene expression value of one patient is less than lower whisker L or larger than upper whisker U, we say that that patient shows extreme expression for that gene. The formulas of L and U are defined as follows:

$$L = Q_1 - 1.5 \times IQR \quad \text{and} \quad U = Q_3 + 1.5 \times IQR,$$

where Q_1 denotes the first quartile of all patients' expression values for that gene, Q_3 denotes the third quartile, and $IQR = Q_3 - Q_1$.

3.1.2 Bin

One obvious disadvantage of "*Whisker*" method is that every word can only appear at most once in a document, which cannot capture the word occurrences information effectively. In order to capture the word occurrences information, we can split the range of extreme values into N bins and use these bins to count different occurrences of the gene expression.

3.1.3 Bi-Bin

In the "*Bin*" method, we treat high extreme values and low extreme values as the same. In this case, we may lose some useful information during the construction. In "*Bi-Bin*" method, we treat high extreme values and low extreme values separately. In this case, the total number of gene expressions is doubled. From the perspective of LDA, the length of vocabulary V is doubled.

These three methods are illustrated in Figure 2.

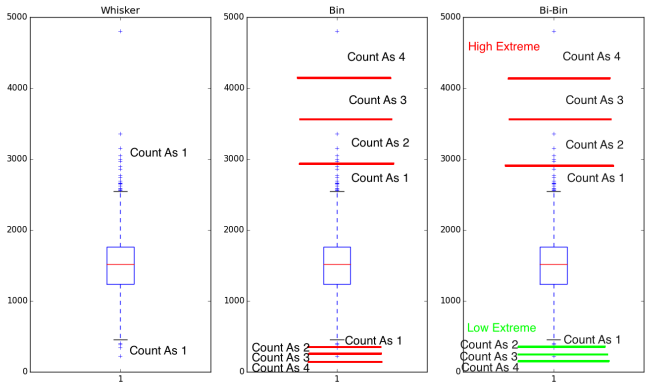

Figure 2: Illustration of "Whisker", "Bin" and "Bi-Bin" for document construction. The red and green solid lines in the figure represent the split of bins. Genes with expression values lie in the bin will be added to the corresponding documents several times shown in the figure.

3.1.4 Fold-Change

Given M healthy samples, we can also utilize the gene expression values of these samples to help construct the documents. The fold change fc for gene v of patient u is defined as follows:

$$fc(u,v) = \log \left| \frac{GE_u(v)}{\sum_{i=1}^{M} GE_i^{healthy}(v)} \right|$$

where $GE_u(v)$ denotes gene expression value for gene v of patient u and $GE_i^{healthy}(v)$ denotes gene expression value of the ith healthy samples for gene v. Then, we can add gene v to patient u's document $\lfloor fc(u,v) \rfloor$ times.

3.1.5 Normal-Bin

After some visualization of the gene expression values, we find that the distribution of the gene expression values are similar to the normal distribution. As a result, we can use normal distribution to hep construct the documents.

To fit gene expression value for a gene with normal distriubtion,

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

we just need to decide the two parameters μ and σ^2 by calculating the mean and the standard deviation of the gene expression value over all samples. Then, we can use standard deviation as bins to count the occurrences of the words. Just like "*Bi-Bin*" method, we also treat two directions separately. Figure 3 shows how this method works.

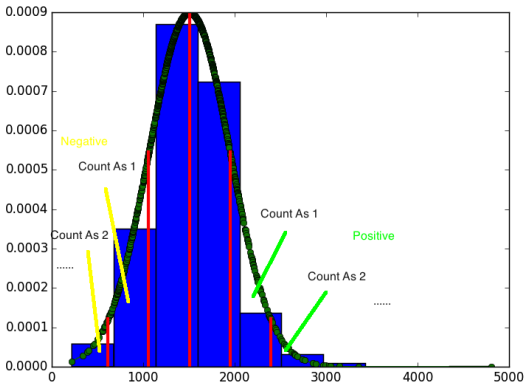

Figure 3: How "Normal-Bin" works. The central red line indicates the mean. Other red lines indicate mth standard deviation from the mean. Genes with expression values lie between the mth std and $m+1$th std will be added to the corresponding documents m times.

3.2 Parameter Selection

Another problem in our project is how to choose the topic number K in LDA. In our project we use 5-fold cross-validation on training part to choose the optimal topic number for each method. By experiments, we find that commonly 10 topics can achieve generally better performance among 10, 20 and 30 topics.

4 Experiments

4.1 Evaluation Metrics

MTLR predictor provides several kinds of measures, including Log likelihood loss, L1 loss, calibrated distribution, and so on. In this project, some metrics are selected to evaluate the performance of the predictor.

4.1.1 D-calibration χ^2-test

D-calibration χ^2-test is a measure to check whether the model can provide believable predictions. For a given model $\hat{\theta}$, the survival curve for a patient x_i is $P_i(t|\hat{\theta}, x_i)$ which is the probability that the event time occurs after time t. Let e_i denotes the event time, which is the death or censored time for uncensored or censored patients, respectively. Assume $p_i(e_i)$ is the predicted probability that each patient x_i died or censored at timee d_i based on model $\hat{\theta}$ for x_i. For any set of patients, $D = \{p_i(e_i)\}$ is drawn from the appropriate distribution. The model θ is distributional calibrated if D is uniform. If the value of χ^2-test is well under 30.14, it suggests that the model is well distributional calibrated [2].

4.1.2 Log Likelihood Loss

One measure used to evaluate the performance of survival prediction is the log likelihood loss. Log likelihood relates to that whether the predictor is "accurate". Let f_r be the facts about patient r and e_r be the "event" time of patient r. The log likelihood of a predictor is

$$\sum_{r \in C} \log P\left(e_r | f_r, \hat{\theta}\right) + \sum_{r \in U} \log p\left(e_r | f_r, \hat{\theta}\right)$$

where C denotes the indices of the censored patients and U denotes the indices of the uncensored patients, and $\hat{\theta}$ is the learned values of parameters. The likelihood for a censored patient is $P\left(e_r | f_r, \hat{\theta}\right) = p(e_r > T | f_r, \hat{\theta})$, where e_r is the censored time. For the uncensored patient, the likelihood is $p\left(e_r | f_r, \hat{\theta}\right) = p(e_r = T | f_r, \hat{\theta})$, where e_r is the death time. After training the model, MTLR tool gives us the log likelihood loss after testing, which is the negative average value of probability of survival times. Lower values means the performance are better.

4.1.3 L1 Log Loss

L1 log loss evaluates the log of the survival times. L1 log loss of a predictor is

$$\frac{1}{n} \sum_{i} |\log t_i - \log p_i|,$$

where p_i is the median of the patient's predicted survival distribution for the i-th patient, and t_i is the true survival time for patient i. It is the mean of the absolute error between logarithm of predicted and true time. One thing to be noted is that L1 Log Loss only deals with uncensored patients.

4.1.4 Concordance Score

Concordance index is one of the most common measure of survival models, which evaluates the probability of concordance between predicted and true survival, shown as the following equation:

$$\frac{1}{m} \sum_{r,s} I[t_r > t_s \;\&\; p_r > p_s].$$

It considers all comparable pairs of patients (r, s). Comparable pairs mean that both patients are uncensored or one censored at time t_r and the other uncensored at an earlier time t_s. The time here is the median value of patients. I is the indicator function, which gives a score of 1 if predicted survival time of patient r is larger than patient s and patient r is actually live longer than patient s. The parameter m denotes the number of comparable pairs. Higher concordance score indicates that the predictor makes better ranking prediction.

4.2 Dataset

We use the data of breast cancer project from The Cancer Genome Atlas (TCGA). We summarize clinical features and gene expression data from 1099 breast cancer patients. Each patient can be

viewed as one sample. One thing should be noted that 993 samples are censored, and 106 samples are uncensored. For each of 1099 patients, clinical information such as age at diagnose, tumor information such as the tumor state, margin status are extracted. There are totally 15 clinical and tumor features for each patient. High throughput measurements of gene expression of each patient are also obtained in the dataset. For each patient, there are totally 20532 gene expression values. As introduced above, LDA model is employed to learn topic features from gene expression data. Moreover, we also got 104 healthy samples which contain the same features as introduced above.

4.3 Results

As introduced above, there are 1099 patients in the dataset. First, we split the dataset such that 80% of the dataset are used as training set and the remaining 20% are used as the testing set. The following figure shows the work flow. Since several sets of topic features are extracted by LDA model based on different document construction methods. In order to select which set of topic features work better, five-fold cross validations are performed on different feature sets, respectively. According to the results of five cross validation, the topic features are selected with best performance. Then we retrain the MTLR model by using the selected topic features and clinical features, and testing are performed on the final model. The Figure 4 shows the overall workflow of our project.

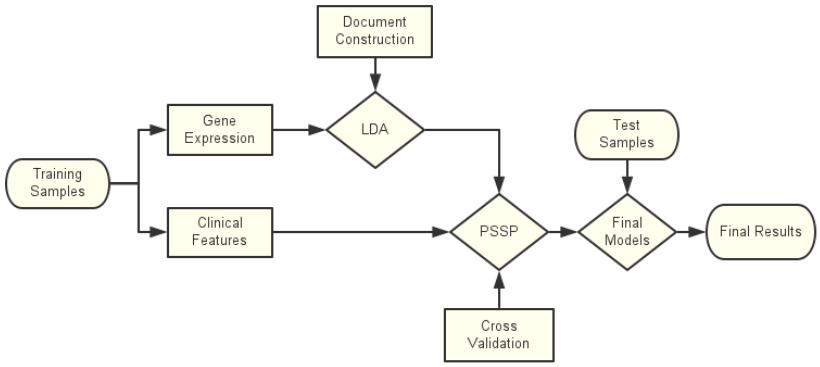

Figure 4: Overall Workflow

4.3.1 Results of five-fold cross validation on different topic feature sets

Five-fold cross validations are performed on six different feature sets, which consist of different topic features extracted by different strategies, and other clinical and tumor features. The result of only using tumor and clinical features is considered as the baseline. Table 1 shows the five-fold cross validation results on different feature sets. The second column presents the results of baseline, and the third column named "*Whisker*" denotes the results on whisker topic features and clinical and tumor features, and so forth.

By comparing against the baseline, we can see the predictions are improved by adding topic features. When Bi-bin topic features are added to the feature space, the predictor gets the best performance than other topic features. The average log likelihood loss is 0.72, lower than other predictors. Moreover, this predictor achieves highest concordance score. Therefore, we selected Bi-bin topics as the optimal topic features. The average value of D-calibration of this predictor is 2.93, which suggests that this predictor is well distributional calibrated.

4.3.2 Test results on Bi-Bin topic features

Based on the five-fold cross validation results, we take the Bi-Bin topics as the optimal topic features. Then the model is retrained on the feature set consisting of topic, clinical and tumor features, after which the testing are performed on the retrained model. Table 3 shows the testing results. The online

Table 1: Five-fold cross validation results on different feature sets

	Baseline	**Whisker**	**Bin**	**Bi-Bin**	**Normal-Bin**	**Fold-Change**
Concordance score	0.48 ± 0.04	0.5 ± 0	0.72 ± 0.05	0.75 ± 0.06	0.5 ± 0	0.71 ± 0.06
Log likelihood loss	0.85 ± 0.04	0.84 ± 0.04	0.85 ± 0.04	0.72 ± 0.05	0.85 ± 0.04	0.74 ± 0.04
L1 log loss	0.66 ± 0.03	0.73 ± 0.05	0.99 ± 0.04	0.89 ± 0.01	0.66 ± 0.04	0.93 ± 0.05
D-calibration χ^2-test	11.19 ± 0.74	11.25 ± 0.58	3.5 ± 0.51	2.93 ± 0.63	11.42 ± 0.7	2.65 ± 0.63

MTLR also gives us results by using Kaplan-Meier [6] method. We can see that MTLR performs better than Kaplan-Meier method.

Table 2: Test results on Bi-bin topics

Measure	**MTLR**	**Kaplan-Meier**
Concordence score	0.65	0.5
L1 Log Loss	0.13	0.14
D-calibration χ^2-test	7.81	2.65

4.4 Analysis

Generally, the "Bi-Bin" achieve the best results, it gains an improvement of 0.27 on concordance score and 0.13 on log likelihood loss over baseline which is really good. But it performs worse than baseline on L1 log loss. The "Bin" and "Fold-Change" also have modest improvements over the baseline. However, the "Whisker" performs almost the same as the baseline, which means that "Whisker" does not work well, and we should use the occurrence information of over expressed genes. The result of "Bin" shows that it does improve the prediction by adding the occurrence information. The "Bi-Bin" considers the extreme low and high values separately, which achieves better result than "Bin". It indicates that low extreme gene expression values contribute to the survival prediction and we should treat the high and low gene expression values differently. However, the result shows that the "Normal-Bin" does not work. We guess there are two reasons. First, the distributions of some gene expression values are quite different from the normal distribution. Second, the threshold of adding genes is not strict. We should select more overexpressed genes, and only add the genes whose expression values are far away from the mean.

5 Conclusion and Further Work

In this report, we first briefly explained our motivation of this project and showed some background materials. Then, we propose five different methods for document construction and illustrate how to choose the optimal topic number. In the experiments, we choose several evaluation metrics to evaluate our results. The best results have quite significant improvements over the baseline method. In the future, we will do more experiments to further improve the results and explore the reason behind them.

6 Acknowledgements

We would like to express our appreciation to Dr. Russ Greiner. Thanks for his time and efforts on guiding the whole process of our project. Also, we would like to thank Luke Kemar for being our co-coach and providing some suggestions.

7 References

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003, 3: 993-1022.

[2] Lin H, Baracos V, Greiner R, et al. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*. 2011: 1845-1853.

[3] David, Cox R. "Regression models and life tables (with discussion)." *Journal of the Royal Statistical Society* 34 (1972): 187-220.

[4] Van De Vijver, Marc J., et al. "A gene-expression signature as a predictor of survival in breast cancer." *New England Journal of Medicine* 347.25 (2002): 1999-2009.

[5] Cancer Genome Atlas Network. "Comprehensive molecular portraits of human breast tumours." *Nature* 490.7418 (2012): 61-70.

[6] Efron, Bradley. "Logistic regression, survival analysis, and the Kaplan-Meier curve." *Journal of the American Statistical Association* 83.402 (1988): 414-425.

[7] Ye, Shuyun, John A. Dawson, and Christina Kendziorski. "Extending Information Retrieval Methods to Personalized Genomic-Based Studies of Disease." *Cancer informatics* 13.Suppl 7 (2014): 85.