Survival Prediction of Breast Cancer Patient from Gene Methylation Data with Deep LSTM Network and Ordinal Cox model

Isabelle Bichindaritz, Guanghui Liu, Christopher Bartlett

Department of Computer Science, State University of New York at Oswego, New York, USA
{ibichind, guanghui.liu, cbartle3}@oswego.edu

Abstract

Survival analysis has currently become a hot topic because it has been proven to be useful for understanding the relationships between patients' variables and covariates (e.g. clinical and genetic features) and the effectiveness of various treatment options. In this study, we study survival analysis of breast cancer patient with gene methylation data and clinical data. We propose a novel method for survival prediction using bidirectional LSTM network and ordinal Cox model. First, gene methylation expression data and clinical data are merged and filtered. To reduce the gene expression features dimension, a weighted gene co-expression network analysis (WGCNA) algorithm is used to obtain cluster eigengenes. Then, the eigengenes serve as input features for a machine learning network. We build a cox proportional hazards model for survival analysis and use LSTM method to predict patient survival risk. We use the leave-one-out method for cross validation and the concordance index (C-index) to evaluate the prediction performance. Stringent cross-validation tests on the benchmark dataset demonstrates the efficacy of the proposed method, which achieves very competitive performance with existing state-of-the-art methods.

Introduction

Breast cancer is one of the most omnipresent diseases in today's US Healthcare system (Group, 2017). Gene methylation influence on cancer has been introduced with great success. Methylation of CpG sites is an epigenetic regulator of gene expression that usually results in gene silencing. Extensive perturbations of DNA methylation have been noted in cancer, causing changes in gene regulation that promote oncogenesis (Suzuki et al., 2012). Consequently, to explore the utility of methylation analysis for cancer diagnosis, we analyzed gene methylation expression of tumors from patients with breast cancer to identify potential cancer-specific methylation genes. One long-term goal of cancer research is to be able to identify prognostic factors that affect patients' survival time, which in turn allows clinicians to make early decisions on treatment. The clinical phenotype of breast cancer is quite diverse, ranging from slow-growing localized tumors to aggressive metastatic disease. Therefore, prognostic markers play a crucial role in stratification of patients for personalized cancer management, which could avoid either overtreatment or undertreatment. For instance, patients classified into a high-risk group may benefit from closer follow-up, more aggressive therapies, and advanced care planning (Yu et al., 2016).

Cox proportional hazard model (Lin et al., 1993) is among the most popular survival prediction models. Recently, based on the Cox model, several regularization methods have been proposed in the literature. The least absolute shrinkage and selection operator Cox model (LASSO-COX) (Shao et al., 2018) apply lasso feature selection method to select components that are related to cancer prognosis. Random survival forest (RSF) (Ishwaran et al., 2008) computes a random forest using the log-rank test as the splitting criterion. It computes the cumulative hazards of the leaf nodes and averages them over the ensemble. Cox regression with neural networks using a one hidden layer multilayer perceptron (MLP) (Xiang et al., 2000) was proposed to replace the linear predictor of the Cox model. It was showed that some novel networks were able to outperform classical Cox models (Amiri et al., 2008). DeepSurv (Katzman et al., 2016) is a Cox proportional hazards deep neural network and a survival method for modeling interactions between a patient's features and treatment effectiveness in order to provide personalized treatment recommendations (Katzman et al., 2018). DeepSurv is developed upon Cox proportional assumption with a cutting-edge deep neural network model.

Although much progress has been made using above approaches, the prediction performance of the existing methods is still far from satisfactory, and there still exits much room for further improvement. In addition, the above methods assumed that the survival information of one patient is independent from another, and thus miss the strong ordinal relationships between the survival times of different patients. Motivated by all these consideration, we thus propose a survival analysis method with deep LSTM networks (Graves et al., 2005) using ordinal Cox model to predict a breast cancer patient's survival risk from gene methylation data. We demonstrate the importance of gene methylation signatures and the efficacy of the proposed method. In data preprocessing, we use weighted gene co-expression network analysis (WGCNA) algorithm (Langfelder and Horvath, 2008) to cluster genes into co-expressed eigengenes. Below, we describe these steps systematically and evaluation results.

Materials and Methods

Benchmark Datasets

Survival data are included in the main clinical file downloaded from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), which provides an extensive collection of genomic and clinical outcome data for large cohorts of patients of more than 30 types of cancers. The main files contain patients' clinical annotations and information. In our case, two clinical variables are used: Overall Survival Status (1 if the patient deceased, 0 if he/she is living at the time of the last follow-up) and Overall Survival (Months), which represents the number of months between diagnosis and date of death or last follow-up. In clinical data, patients with missing follow-up were excluded. The gene expression data are the normalized methylation data. The file used was downloaded from FireHose (Deng et al., 2017). Gene methylation expression data and clinical data were merged and filtered to keep matching. Consequently, the benchmark dataset including 779 patients with 20106 genes was obtained. The gene methylation and clinical characteristics for the selected patients are summarized in Table 1.

Table 1. Methylation gene and clinical characteristics.

Characteristics	Summary
Patient no.	779
Gene no.	20106
Survival status	
Living	677
Deceased	102
Follow up (months)	0.03-282.69
Age (years)	
Range	26-90
Median	58.08

Gene Co-expression Clustering: WGCNA

As for the gene methylation expression data, we firstly use WGCNA algorithm to cluster genes into co-expressed modules, and then summarize each module as an eigengene. This algorithm yields 12 co-expressed gene modules (features).

System Algorithm Flow Chart

Figure 1 shows the algorithmic process of our proposed method. There are three stages, including the gene co-expression cluster stage, the bidirectional LSTM network stage and the COX model stage. In the gene co-expression clustering stage, gene methylation expression data could be reduced in terms of feature dimension. WGCNA algorithm is used to cluster genes. So, twelve eigengenes are obtained and will serve as input features for the machine learning network. Secondly, the bidirectional Long-Short-Term-Memory (biLSTM) method is proposed to predict patient survival risk. Finally, a novel ranking loss function for the deep cox proportional hazard model is built for survival analysis to ensure that the ordinal relationship among the survival time of different patients can be preserved.

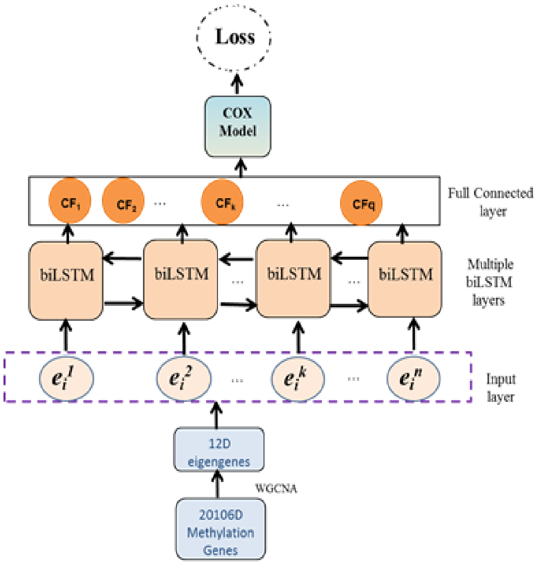

Figure 1. Schematic diagram of the proposed method

Cox Model and Loss Function

In survival analysis, prediction of the time duration until a certain event occurs is the goal of survival analysis and the death of a cancer patient is the event of interest in our study. Cancer patients can be divided into two categories i.e., censored patients and non-censored patients. For censored patients, the death events were not observed for them during the follow-up period, and thus their genuine survival times are longer than the recorded data; while for non-censored patients their recorded survival times are the exact time from initial diagnosis to death. We use a triplet (x_i, t_i, δ_i) to represent each observation in survival analysis, where x_i is the feature vector, t_i is the observed time, and δ_i is the censoring indicator. Here, $\delta_i = 1$ or $\delta_i = 0$ indicates a non-censored or censored instance, respectively. The negative log partial likelihood function of the Cox model is defined as follows (Farzindar and Kashi, 2019; Sy and Taylor, 2000):

$$l(\alpha) = \sum_{i=1}^{n} \delta_i \left(\alpha^T x_i - \log \sum_{j \in R(t_i)} \exp(\alpha^T x_j) \right) \qquad (1)$$

where $\alpha^T x_i$ is called the survival function, in which α can be estimated by minimizing its corresponding negative log partial likelihood function; n denotes the number of patients, $R(t_i)$ denote the set of all individuals at risk at time t_i, which represents the set of patients that are still under risk before time t_i.

Although we could use the above Cox model to directly make survival prediction, it does not take the ordinal survival information between different patients (e.g., the survival time for patient A is longer than that for patient B) into consideration. In the hazard ratio based model, if the survival time for patient i is shorter than that for patient ($i-1$), the hazard risk of patient i will be zero and the hazard risk of patient ($i-1$) will be ($x_i + x_{i-1}$). By utilizing the above ordinal relationship indicated by the Cox model, we design a ranking loss function to capture the ordinal survival information among different patients as follows:

$$J(\sigma(D_i)) = \sum_{i=1}^{n} \delta_i \left(\alpha^T \sigma(D_i) - \log \sum_{j \in R(t_i)} \exp(\alpha^T \sigma(D_j)) \right) \qquad (2)$$

Where D_i is individual with event times at t_i, and is part of $R(t_i)$; $\sigma(D_i)$ is the risk set at ordinal time, and when $t_i < t_{i-1}$, there are $\begin{cases} \sigma(D_{i-1}) = x_i + x_{i-1} \\ \sigma(D_i) = 0 \end{cases}$.

Experimental Results and Discussions

In this part, we assess the performance of the proposed method and carry out experiments on the training set through leave-one-out cross validation. Specifically, we firstly use WGCNA algorithm to cluster genes and obtain 12 eigengenes. Then the Cox proportional hazards model is built on the clustered eigengene features in the training set. After that, the median risk score predicted by the cox proportional hazards model is used as a threshold to split patients into low-risk and high-risk groups. Finally, we test if these two groups have distinct survival outcomes using Kaplan-Meier estimator and rank test. The survival curves are drawn by applying different methods.

Comparison with Different Survival Prediction Methods over Cross-validation Test

We compare the prediction effects of our proposed method with four machine learning methods: RSF (Ishwaran et al., 2008), LASSO (Shao et al., 2018), MLP (Amiri et al., 2008), and DeepSurv (Katzman et al., 2016). The concordance index (C-index) (Mayr and Schmid, 2014) is used to evaluate the prediction performance. C-index quantifies the fraction of all pairs of patients whose predicted survival times are correctly ordered. For the sake of fairness, we carry out the same feature set in all cross validation tests.

Table 2 Performance comparison among different survival prediction methods by the measurements of Concordance Index (C-index)

Methods	C-index
The proposed method	0.6330
DeepSurv	0.6122
MLP	0.6090
LASSO	0.6032
RSF	0.5449

Table 2 summarizes the performance comparisons between the proposed method, DeepSurv, MLP, Lasso, and RSF by the measurements of C-index. From Table 2, we find that the cross validation of the proposed method on the standard training set is better than the other four methods. Compared with the methods: RSF, LASSO, MLP, and DeepSurv, the C-index of the proposed method is improved by 8.81%, 2.98%, 2.40% and 2.08%. As can be seen from Table 2, firstly, the prognosis power of the regularized Cox models (i.e., RSF and LASSO) is inferior to the other deep model based methods (i.e., MLP and DeepSurv). This is because the deep model can better represent gene features than the hand-crafted low-level features. Secondly, the proposed biLSTM method can achieve higher C-index values than the comparing methods, which demonstrates the advantage of LSTM that can represent the heterogeneous patterns of sequential methylation data.

Survival Stratification Prediction

Another important task in survival analysis is to stratify cancer patients into subgroups with different predicted outcomes, by which we can develop personalized treatment plans during cancer disease progression. The median risk score method is used in the training set as a threshold to stratify patients in the test set into low-risk and high-risk groups, and then test if these two groups have significantly different survival time using the log-rank test. Better prognosis prediction performance comes with smaller p-value from the log-rank test. We show the stratification performance of different prediction methods in Figure 2.

As shown in Figure 2, the proposed prediction method achieves significantly superior stratification performance (lower p-value) when compared with the other methods (RSF, LASSO, MLP, and DeepSurv) on gene methylation datasets. This is because our proposed model considers both the ordinal characteristics and the heterogeneous patterns in survival analysis. Thus its prognostic power is effectively improved.

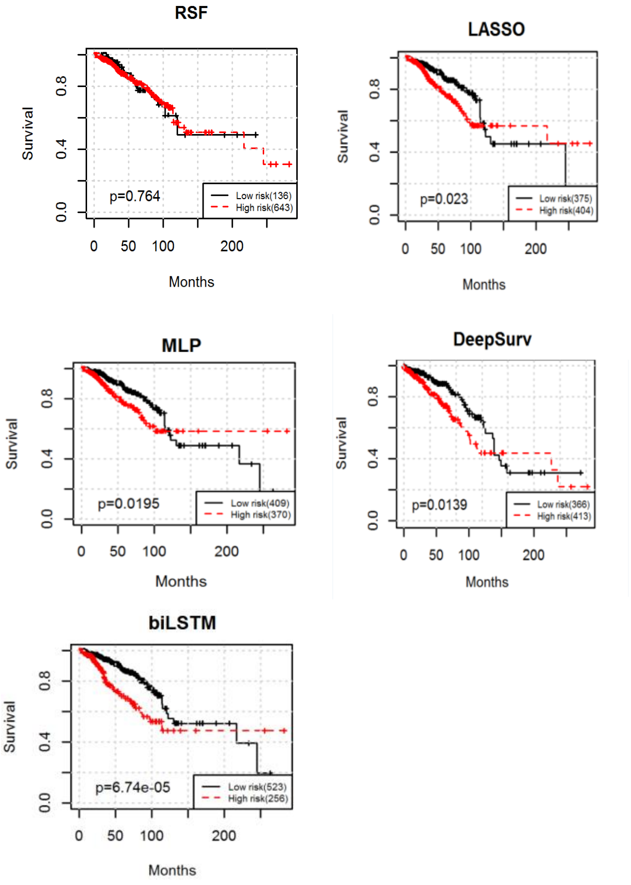

Figure 2. The survival curves by applying different methods.

Conclusion

In this study, we have developed a survival prediction framework for breast cancer patients, in which we take patients' ordinal survival information into consideration. Leave-one-out cross-validation experiments on the gene methylation expression data and clinical data were carried out. Experimental results demonstrate the superiority of the proposed method over the existing RSF, Lasso, MLP, and DeepSurv predictors. The good performances of the proposed method come from the use of the combined bidirectional LSTM predictor and ordinal information.

References

Amiri, Z.; Mohammad, K.; Mahmoudi, M.; Zeraati, H.; and Fotouhi, A. 2008. Assessment of gastric cancer survival: using an artificial hierarchical neural network. *Pak J Biol Sci*, 11(8):1076-84.

Deng, M.; Brägelmann, J.; Kryukov, I.; Saraiva-Agostinho, N.; and Perner, S. 2017. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database*, 2017.

Graves, A.; Fernández, S.; and Schmidhuber, J. 2005. idirectional LSTM networks for improved phoneme classification and recognition. International Conference on Artificial Neural Networks, 2005, pp. 799-804. Springer.

Group, U.C.S.W. 2017. United States cancer statistics: 1999–2014 incidence and mortality web-based report. *Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*.

Farzindar, A.A.; and Kashi, A. 2019. Multi-Task Survival Analysis of Liver Transplantation Using Deep Learning. *The Thirty-Second International Flairs Conference*, 2019.

Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; and Lauer, M.S. 2008. Random survival forests. *The annals of applied statistics*, 2(3):841-860.

Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2016. Deep survival: A deep cox proportional hazards network. *stat*, 1050:2.

Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24.

Langfelder, P.; and Horvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Lin, D.Y.; Wei, L.-J.; and Ying, Z. 1993. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557-572.

Mayr, A., and Schmid, M., 2014. Boosting the concordance index for survival data. *Ulmer Informatik-Berichte*:26.

Shao, W.; Cheng, J.; Sun, L.; Han, Z.; Feng, Q.; Zhang, D.; and Huang, K. 2018. Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 648-656. Springer.

Suzuki, H., Maruyama, R., Yamamoto, E., and Kai, M., 2012. DNA methylation and microRNA dysregulation in cancer. *Molecular oncology*, 6(6):567-578.

Sy, J.P.; and Taylor, J.M. 2000. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227-236.

Tomczak, K.; Czerwińska, P.; and Wiznerowicz, M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68.

Xiang, A.; Lapuerta, P.; Ryutov, A.; Buckley, J.; and Azen, S. 2000. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243-257.

Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., and Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7:12474.