PBAE : 암 생존 분석을 위한 PATHWAY 기반 오토인코더 모델

벨무루간 아레쉬 발라지, 최철웅, 김경백
전남대학교 인공지능융합학과
e-mail: arreshvnass@gmail.com, sentilemon02@gmail.com, kyungbaekkim@jnu.ac.kr

Pathway Based Auto-Encoder (PBAE) Model for Cancer Survival Analysis

Velmurugan Arresh Balaji, Chulwoong Choi, Kyungbaek Kim
Dept. Artificial Intelligence Convergence, Chonnam National University

요 약

In 2020, the COVID-19 pandemic dominated the global consciousness, directing health policy and research efforts. Glioblastoma multiforme (GBM), the most common malignant brain tumor, universally carries a poor prognosis. Despite aggressive multimodality treatment, the median survival is ~18-20 months, depending on molecular subgroups. Developing cancer patient's treatment in not only important, by using clinical and genomic data we have to understand and improve the survival prediction of cancer patients through biological complex mechanisms, is more vital in these days. Looking forward, defining the predictive value and clinical utility of detecting resistance mechanisms in 'real time', and to increase the speed, efficiency and accuracy of predictive biomarker analysis, will bring the field closer to a cost-effective and, notably, patient-centric reality. In this thesis, we proposed a biologically interpretable pathway-based autoencoder model named PBAE, which is an simple Artificial Neural Network which combines both clinical data and high dimensional cancer gene expression data for anlysing the survival of GBM cancer patients. And the ensemble model with max Linear predictors has achieved a very high c-index metric of 0.7314, which clearly illustrated that the proposed PBAE model has better performance and more effective than the state-of-art models.

1. Introduction

For human cancer diseases, high dimensional molecular profiles such as RNA-seq data and gene expression data can be effeciently obtained from platforms such as advanced molecular sequencing, etc [1]. The usage of high-dimensional genomic data has been increased for effective decision-making in clinical field, and for elucidating the biological mechanisms in the cancer gene data. For survival distribution estimation from the data, there are many methods available, which is collectively called as survival analysis which produces an output called the survival time which depends on the event of interest observed from cancer patients. Handling missing data is an predominant part of survival analysis. One such data is called right-censoring data of cancer patients. For analyzing clinical trials data such as time-to-event data, the survival Hazards regression model called Proportional hazard Cox model (Cox-PH) is the most widely used approach [2, 3]. With very few assumptions, Cox-ph is an effective semi-parametric model which interprets risk factor's effects in an effective way.

Eventhough, the Cox-PH model is high efficient, the conventional model have some limitations such as (1) Analyzing low-sample sized data which are high-dimensional; and (2) While handling the covariates which has some highly nonlinear relationship between them. When we train the datasets which are high-dimensional in case, most of the results were either over-fitted or it become infeasible [6]. As a consequence, clinical information datasets with large sample sized and which has a very low dimension, were used for the survival prediction of patients through the conventional Cox-PH. Additionally, for the selection algorithm to be guaranteed, an highly efficient approach for selecting features which includes almost all of the significant covariates [7]. For chronic human diseases, there is an highly non-linear impact on the patient survival by the genomic information of patients

[8], but the covariate's Linear contribution has been assumed by the basic cox-ph model. The censored survival phenotypes, have some linear effects on profiles of gene expressions like overall survival time and relapse time. To manage them, they implemented an cox-ph based on Kernel. [9]. It is still challenging to seek the optimal kernel function, Eventhough with hyper-parameters of optimal pair, it is difficult to get an kernel function which is optimal because, at the starting itself the models have to specify the kernel function incase of kernel-based models.

2. Related Works

For the survival prediction of cancer patients, many models related to deep learning techniques which have integrated the output layer as basic Cox-ph model have been proposed.

2.1 Cox-nnet Model

For regression problem, such as regularized Cox-PH regression, they proposed an simple Neural network called Cox-nnet to handle high-throughput RNA sequencing data [10]. The activation patterns in the hidden nodes of the Cox-nnet model were distinct and those features for reducing the dimension for sensitive reduction for survival. Moreover, the Cox-nnet model has one cox layer as output and 1-2 hidden layers only. Hence, the framework of Cox-nnet model is too simple.

2.2 DeepSurv Model

DeepSurv is a combination neural network with standard Cox-PH regression, and it is deep feed-forwarded for enhancement in prediciting the survival of cancer patients, and it should also serve for the personalized treatment as an recommendation system [11]. But the limitation of this deepsurv model is that, the clinical data with very low-dimension were anlaysed. And the number of variables was also less than 20.

2.3 CoxPASNet Model

In deep learning model it is a bit complex task to integrate Clinical dta or multi-omnics data or varieties of data. For improving the performance of survivcal prediction analysis, there has been many studies proposed to levarage the Clinical and Multi-omnics data [12, 13, 14]. With the help of an matrix form such as augmented matrix, we can retresent those heterogeneous datasets. In the CoxPASNet neural network model, the Pathway layers with the help of pathway masks obtained from pathway databases embed the patient's cancer survival informations and it also understands the complex bilogical mechanims associated with the survival of patients.

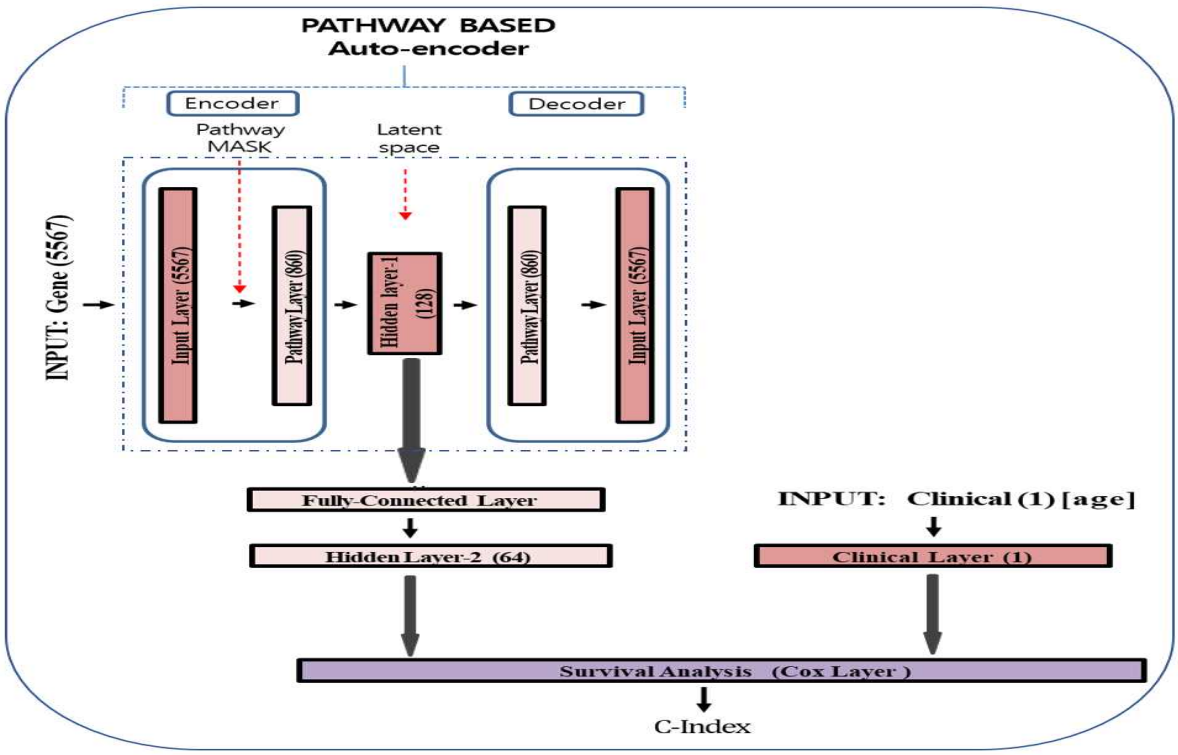
2.4 AECOX Model

As an effective dimensionality reduction technique, the Autoencoder (AE) framework can lead to efficient lower dimensional representations using unsupervised or supervised learning. An Autoencoder based approach (called AECOX) [15] for cancer prognosis prediction with simultaneous learning of lower dimensional representation of inputs. In AECOX, the code from AE will link to a Cox regression layer for the prognosis. Both losses from the AE networks and Cox regression layer will be counted to train the entire network weights through back-propagation.

3. PBAE Model:

The framework of PBAE contains two stages. First, training the autoencoder model completely, and with trained weights then the extracted input gene features are reduced into single-dimensional linear prediction in stage two. Then, resulting representations are fed to the Cox-PH survival Analysis. The whole PBAE framework is shown in Figure 1. PBAE has multiple layers in it. PBAE requires two types of ordered data, gene expression data and clinical data from the same patients. The pipeline layers of the two data types are merged in the last hidden layer and produces a Prognostic Index (PI), which is an input to Cox proportional hazards regression. In this study, we included only age as clinical data. Higher- dimensional clinical data are desired to be integrated with hidden layers in the clinical pipeline.

The gene layer is an input layer of PBAE, introducing zero-mean gene expression data (X) with n patient samples of p gene expressions, i.e., X={x1,...,xp} and ~N(0,1).The pathway layer incorporates prior biological knowledge, so that the neural network of PBAE can be biologically interpretable. To implement the sparse connections between the gene and pathway layers, we consider a binary bi-adjacency matrix called pathway masks. The hidden layers depict the nonlinear and hierarchical effects of pathways. The clinical layer introduces clinical data to the model separately from genomic data to capture clinical effects. The Cox layer is the output layer that has only one node. The node value produces a linear predictor, a.k.a. Risk Score from both the genomic and clinical data, which is introduced to a Cox-PH model.



(Figure 1) (Architecture of Pathway based Auto-encoder model)

4. Evaluation

From TCGA we obtained the GBM cancer dataset including both clinical and gene expression data cBioPortal (www.cbioportal .org/datasets). And the patients who doesn't have both death status event and survival time were excluded. For pathway based analysis, from the Molecular Signatures Database [45], we downloaded biological pathways from both Reactome and KEGG databases, because the pathways have the prior knowledge about the complex metabolism of cancer patients. We used C-index metrics, for measuring the performance of cancer survival prediction with censored data. The c-index measures concordant between observed survival time and the predicted score. It is also known as the rank- correlation method. For confirming the percentage of censoring is same on each test, validation and train data, we splited GBM datasets randomly like test data with 20% and the remaining dataset is divide into two parts as 80% training and 20% for validation for each experiments conducted.

Model	Loss Function	Linear prediction	C-index
Cox-PASNet	NLL	Normal	0.6716
PBAE	NLL+L2	Normal	0.6787
Ensemble Method	(NLL)+ (NLL +L2)	Min	0.6948
Ensemble Method	(NLL)+ (NLL +L2)	Average	0.7245
Ensemble Method	(NLL)+ (NLL +L2)	Max	0.7226
Ensemble Method	(NLL)+ (NLL +MSE)	Min	0.7265
Ensemble Method	(NLL)+ (NLL +MSE)	Average	0.7265
Ensemble Method	(NLL)+ (NLL +MSE)	Max	0.7289
Ensemble Method	(NLL)+ (NLL +MSE)	Min	0.7114
Ensemble Method	(NLL)+ (NLL +MSE)	Average	0.7285
Ensemble Method	(NLL)+ (NLL +MSE)	Max	0.7314

(Figure 2) (C-index performance comparison between CoxPASNet, PBAE and Ensemble Methods)

We have undergone an ablation study to enhance the performance of the Cox-Ph survival analysis. Hence, we adopted an ensemble method, in which the linear predictors for ensemble methods were obtained from both the models CoxPASNet and PBAE, and we made multiple experiments like taking only Minimum, or only Average or only Maximum of two set of linear predictors. With the help of Censored events (OS_Events) and ground truth (OS_Months) which has the time to event information, we computed the c-index. Out of several experiments, The Ensemble model with the Max of two set of Linear predictors achieved a very high c-index 0.7314 compared to other ensemble method experiments and with both individual models CoxPASNet and PBAE model. In the final experiment, There were totally 522 in the Linear predictors set. Out of 522 linear predictors, the PBAE model has highest number of Max Linear predictor (272) than the CoxPASNet model (250) respectively.

5. Conclusion

In this article, we proposed an auto-encoder model associated with cancer pathway mask called PBAE, for GBM cancer survival analysis. We assessed the TCGA's GBM dataset using the PBAE model. The results obtained from the experiments illustrated that the PBAE model has effectively overcomed the current state of art models in survival prediction, such as Cox-PASNet and we assesed the effectiveness of its prediction using statistical methods. For analysis purpose, we made an large matrix by integrating both clinical and genomic data. And during that integration, because of the unbalanced size between these two dataset's covariates, The cancer clinical data, might be dominated the genomic data due to its high-dimensionality effect. Hence, our PBAE considers two different layers for both genomic and clinical data inorder to intrpret these two datasets individually in an effective manner.

Acknowledgements

"This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF)& funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961)."

참 고 문 헌

[1] Lightbody G, et al.Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinformatics. 2018; 051. https://doi.org/10.1093/bib/bby051.

[2] Ahmed FE, Vos PW, Holbert D. Modeling survival in colon cancer: A methodological review. Mol Cancer. 2007; 6(1):15. https://doi.org/10.1186/1476-4598-6-15.

[3] Chen H-C, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC Med Res Methodol. 2012;12(1):102. https://doi.org/10.1186/1471-2288-12-102..

[4] Abadi A, et al. Cox Models Survival Analysis Based on Breast Cancer Treatments. Iran J Cancer Prev. 2014; 7(3):124 - 9.

[5] Atashgar K, Sheikhaliyan A, Tajvidi M, Molana SH, Jalaeiyan L. Survival analysis of breast cancer patients with different chronic diseases through parametric and semi-parametric approaches. Multidiscip Cancer Investig. 2018;https://doi.org/10.30699/acadpub.mci.2.1. 26.

[6] Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. Stat Methods Med Res. 2010;19(1):29 - 51. https://doi.org/10.1177/0962280209105024

[5] Fan J, Feng Y, Wu Y. High-dimensional variable selection for Cox's proportionalhazards model. Collections, vol. 6. Beachwood: Institute of Mathematical Statistics;2010, pp. 70 - 86. https://doi.org/10.1214/10-IMSCOLL606.

[6] Mallavarapu T, Hao J, Kim Y, Oh J, Kang M. Pathway-based deep clustering for molecular subtyping of cancer. Methods. 2019. https://doi.org/10.1016/j.ymeth.2019.06.017.

[7] Li H, Luan Y. Kernel Cox Regression Models for Linking Gene Expression Profiles to Censored Survival Data. In: Pac Symp Biocomput 8: 2003. p. 65 - 76. https://www.ncbi.nlm.nih.gov/pubmed/12603018. https://doi.org/10.1142/9789812776303_0007.

[8] Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method forprognosis prediction of high-throughput omics data. PLoS Comput Biol. 2018;14(4):1006076. https://doi.org/10.1371/journal.pcbi.1006076.

[9] Katzman JL, et al.DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018; 18(1):24. https://doi.org/10.1186/s12874-018-0482-1.28

[10] Yousefi S, et al.Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. 2017; 7(1):11707. https://doi.org/10.1038/s41598-017-11817-6.

[11] Wójcik PI, Kurdziel M. Training neural networks on high-dimensional data using random projection. Pattern Anal Appl. 2018:1 - 11. https://doi.org/10.1007/s10044-018-0697-0.

[12] Yousefi S, et al.Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. 2017; 7(1):11707. https://doi.org/10.1038/s41598-017-11817-6.

[13] Lu J, Cowperthwaite MC, Burnett MG, Shpak M. Molecular Predictors of Long-Term Survival in Glioblastoma Multiforme Patients. PLoS ONE. 2016; 29 11(4):0154313. https://doi.org/10.1371/journal.pone.0154313.

[14] Zhu B, et al.Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. Sci Rep. 2017; 7(1):16954. https://doi.org/10.1038/s41598-017-17031-8.

[15] Huang, Zhi, et al. "Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations." BMC medical genomics 13 (2020): 1-12.