Assessing the versions of Pathway based Autoencoder model for Cancer Survival Analysis

Velmurugan Arresh Balaji
Department of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
Email: arreshvnass@gmail.com

Chulwoong Choi
Department of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
Email: sentilemon02@gmail.com

Kyungbaek Kim
Department of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
Email: kyungbaekkim@jnu.ac.kr

ABSTRACT

Glioblastoma multiforme is a chronic tumor occur in brain, still its prognosis rate is poor in worldwide. Even with aggressive multimodality treatment, based on the molecular subgroups, the median survival is ~18-20 months only. By using clinical and genomic data we can improve the survival prediction of cancer patients through biological complex mechanisms, is more vital in these days. We proposed a biologically interpretable attention and pathway-based autoencoder model named PBAE for the GBM cancer survival analysis by using a basic neural network architecture which combines clinical dataset and cancer gene (HDLSS) dataset. In this paper, we made a study about the positioning of pathway mask used in the PBAE model to propose a highly efficient model for cancer survival analysis. The PBAE model with pathway mask initialized as layer weight has achieved a relatively better c-index metric of 0.6716. Moreover, our model has achieved relatively better results in survival probability-based performance metrics like Brier score and MAE as well.

KEYWORDS

Auto-Encoder, Deep Learning, Survival Analysis, Pathway Mask, Self-Attention

1 INTRODUCTION

The human disorders causing molecular profiles (for example, cancer) can be created using modern molecular high-throughput sequencing systems, which effectively make genomic expressions which have high dimensions (e.g., RNA-seq, cancer gene) [1]. The term "survival analysis" refers to a collection of techniques for predicting survival ranges from the data, with the conclusion being the time it takes for an observation to experience an event of interest. The typical Cox- PH model has two major challenges (1) assessing wide, limited (HDLSS) datasets (2) dealing with the nonlinear dynamic correlation analysis. Evaluating HDLSS data is important but difficult in bioinformatics although many datasets contain a small number of data (n) however a huge number of characteristics (p), given by the formula, p>>n. large data sets might lead to either learning unsustainability or test data underfitting [2]. Hence we proposed the PBAE, for analyzing the survival rate that combines clinical dataset with cancer gene expression datasets on a basic deep learning architecture. The proposed PBAE model may minimize irrelevant and misleading covariates while improving the performance of the Cox hazard analysis.

2 Pathway Mask Based Auto-Encoder Model

2.1 Architecture

The framework of PBAE contains two stages. Based on the models from CoxPASNet [3], Cox-nnet [4] and DeepSurv [5] respectively. 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. genomic dataset (X) with zero mean was introduced into the gene layer form of p gene expressions, of n patient samples i.e., X={x1,...,xp} and xi~N(0,1). The second layer is the pathway layer with each node indicating one pathway that are biologically distinct. Given pathway databases containing pairs of p genes and q pathways, the binary bi-adjacency matrix (A∈ $Bq\times p$) is constructed, where an element aij is one if gene j belongs to pathway i; otherwise, it is zero, i.e., A={aij|1≤i≤q,1≤j≤p} and aij={0,1}. Higher-level representations of biological pathways are expressed in the deeper hidden layer. To capture clinical impacts, this layer contributes survival clinical information (age) for the PBAE model apart from genomic data. The Cox layer is termed as the last output layer which contains only one node. From both genetic and clinical data, the value of node generates a risk score which is then used in a proportional hazard model.

2.2 Initializing Pathway mask as input

In the Figure 1, the pathway mask is used as input. Before training the model, the dataset is preprocessed, and we make the augmented matrix by combing the gene expression data and the bi-adjacency matrix called pathway mask and we generated the pathway-based gene expressions (Features) which is eventually fed into the PBAE model. In this case, the Layer 1 is considered as the pathway layer. And it has 860 nodes, out of which each nodes represents a pathway of GBM cancer patients. Only Features with high risk will be fed into the pathway layer. Eventually, the necessary features will be derived, and the rest unwanted features will be excluded.

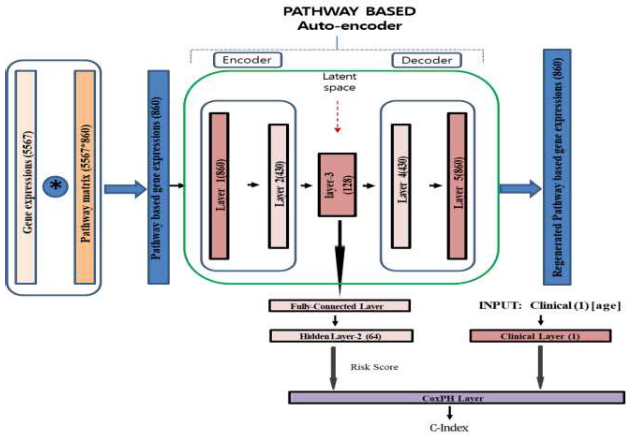

Figure 1: Initializing Pathway mask as input in PBAE model.

2.3 Initializing Pathway mask as Layer Weight

We made a little variation, by initializing the pathway mask as layer weight as shown in figure 2. In this model, the gene expression data will be fed into the input layer like the conventional way. But the difference is, the weight lies among the pathway and input layer will initialized by pathway mask at the time of training the model.

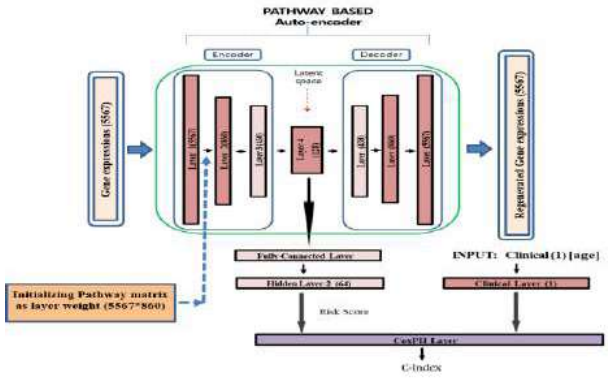

Figure 1: Initializing Pathway mask as Layer Weight in PBAE model.

Table 1: Performance metrics Comparison of PBAE model

Model	C-index	Brier Score	NBLL	MRL	Median Life
PBAE(Initializing Pathway Mask as Input)	0.6716	0.167786	2.1644062	19.238	20.122702
PBAE(Initializing Pathway Mask as Layer Weight)	0.63328	0.148715	1.9878166	17.841	18.251463

At the time of training, 5567 gene expression data will be fed into the 5567 input nodes of the input layer. Since our model is a deep learning method, the input genes are multiplied with the initialized weights (pathway mask) and fed into the pathway layer. Here, the dimensional reduced features (860 features) will be fed into the 860 nodes of pathway layer. With the help of pathway mask, only necessary features will be precisely obtained.

3 Experiments and Results

3.1 Dataset

For TCGA GBM cancer dataset, we gathered from the cBioPortal (www.cbioportal. org/datasets). We gathered biological pathways as previous information using Database called Molecular Signatures (MSigDB). Eventually, we obtained dataset from 522 GBM cancer sufferers with 5,567 genes, 860 pathways, and clinical (age).

3.2 Training and implementation detail

The Tanh function was chosen for the activation function since it generated the greatest Concordance-index score when contrasted to those other LeakyRELU and RELU activation functions. An empirical search yielded dropout rates of 0.7 and 0.5 in the very first hidden layer and the pathway layer correspondingly. Adaptive Moment Estimation (Adam) was used to improve the neural network. PBAE uses the negative log partial likelihood loss function, and the Mean Square Error (MSE) loss function obtained from the Autoencoder.

3.3 Evaluation Metrics

The c-index is mainly used for predicting the performance of model's discrimination. But the NBLL and brier score is used to predict the model's calibration and the discrimination as well. Median Life is basically the overall average of the subtraction of actual time values observed and time value predicted. The Mean Residual Life gives the expected remaining lifetime for individuals who have yet to experience the event of interest at time t.

4 Result Discussion

In this section, we described the performance metrics of the PBAE model for the GBM dataset The PBAE with pathway mask as input has achieved relatively higher C-index measure of 0.6716. The c-index turned down to be very low for the model with pathway mask as layer weight. But at the same time, the remaining performance metrics, the PBAE model has outperformed in Brier score, NBLL,

2

Mean Residual Life, and median Life, respectively. Hence, the model performed well except c-index. The results of PBAE model (Table 1) illustrates that, the positioning of pathway mask had a huge impact in the overall performance of the PBAE model.

5 Conclusion

In this paper, the PBAE model was proposed, for GBM cancer survival analysis to predict with gene expressions and clinical datasets of GBM patients. We made an assessment study about the positioning of the PM in our model which has achieved quite better results in performance metrics like, NBLL, Brier score and MAE except c-index, when the PM is initialized as layer weight. For the future work, we had planned to work with multi-modal data, for example expressions of mRNA, DNA methylation, etc.

ACKNOWLEDGMENTS

"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)".

REFERENCES

[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] Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. Stat Methods Med Res. 2010; 19(1):29–51. https://doi.org/10.117 7/ 096228020 9105024.

[3] Hao, J., Kim, Y., Mallavarapu, T. et al. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. BMC Med Genomics 12, 189 (2019). https://doi.org/10.1186/s12920-019-0624-2..

[4] 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.

[5] Bouveyron,C. and Brunet-Saumard,C. (2014) Model-based clustering of high-dimensional data: a review. Comput. Stat. Data Anal., 71, 52–78..