확률분포를 이용한 딥러닝 기반 비소세포폐암 환자 생존곡선 예측

두이 푸옹 다오*,양형정*, 김수형*,이귀상*,강세령**,오인재***
*전남대학교 인공지능융합학과
**화순전남대학교병원 핵의학과
***전남대학교 의과대학 내과학교실
e-mail: hjyang@jnu.ac.kr

Deep Learning-based Survival Curve Estimation using Probability Distribution in Patients with Non-small Cell Lung Cancer

Duy-Phuong Dao*, Hyung-Jeong Yang*, Soo-Hyung Kim*, Guee-Sang Lee*, Sae-Ryung Kang**, In-Jae Oh***
*Dept of Artificial Intelligence Convergence, Chonnam National University
**Dept of Nuclear Medicine, Chonnam National University Hwasun Hospital
***Dept of Internal Medicine, Chonnam National University Medical School and Hwasun Hospital

Abstract

Non-small-cell lung cancer (NSCLC) represents approximately 80-85% of lung cancer diagnoses and is the leading cause of cancer-related death worldwide. Prediction of time-to-death-event is necessary and helpful for deducing proper treatment at early stages. In this study, we utilize a Deep Neural Network to learn parameters of parametric distribution from well-known probability distribution. This model learns complex relationships between patient's covariates and produce individual's survival curve. We discover that most of time-to-event data encounters highly imbalance in terms of duration (time to event occurrence). Therefore, we propose a novel loss function to tackle this issue. The experimental results demonstrate that our proposal achieves competitive performance compared to conventional methods in terms of C-index metric.

1. Introduction

Survival analysis (also called time-to-event analysis) is a field of statistics that widely used in advertisement [1], medical [2, 3] and industry [4]. Recently, leverage advance in deep learning and machine learning, researchers use neural network to explore and understand the relationships between individual's covariates (e.g. staging, age, gender) to estimate probability of an event of interest until it occurs.

Especially in medical field, survival analysis is applied to estimate risk to death event at each patient based on clinical data (or medical imaging data). Knowing patients have low or high risk to help outline road map for treatment at early stages. Imbalance data is also a challenge in machine learning field. We ideate a novel loss function to solve this problem.

In this study, we propose a deep neural network learnt complex relationships between individual's covariates and distributional representation of whole survival time in dataset. The output of model is two parameters (scale and shape) of a parametric distribution. Through the scale and shape parameters, we can generate individual's survival probability curve which is 1.0 (or 100% alive) at time zero and monotonically decrease to 0 over time.

2. Related Work

In this section, we review the use of conventional approaches. Survival analysis was firstly constructed and utilized by medical

researchers and data analysts to measure the lifespans of a certain population. Nowadays, it has been found in various applications such as cancer patients [5], customer churn [6], credit scoring [7], and failure times of mechanical systems [8].

The Kaplan-Meier estimator method is firstly utilized in the statistical and medical field,

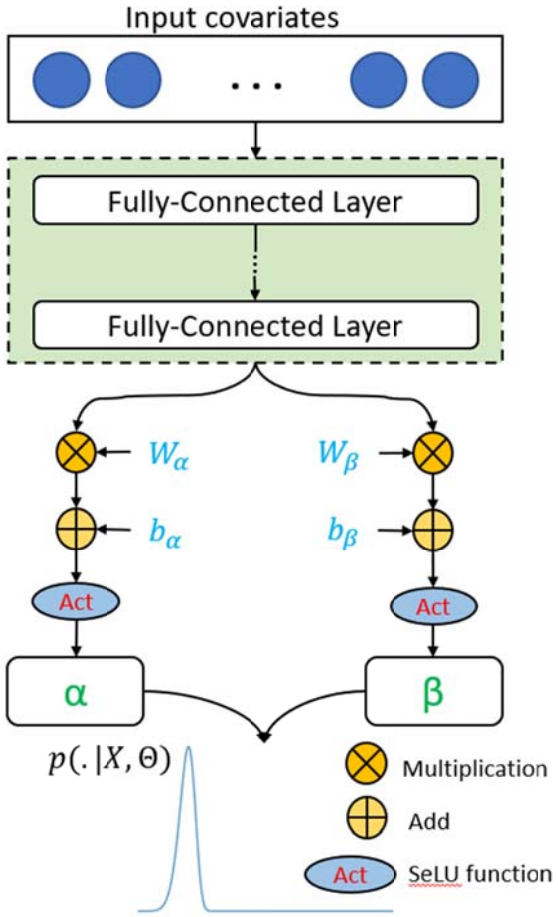

Figure 1. The overall architecture.

which are able to learn flexible and complex time distributions, but not incorporating patient's covariates. Therefore, it is hard to generate individual's survival curve. To tackle this issue, the Cox proportional hazard (CPH) model [9] was introduced to take patient's covariates. However, CPH has limitation in learning impact of covariates due to the use of constant factor.

Recently, DeepSurv [10] and DeepHit [11] were introduced and attracted much attention from researchers. However, Deephit exists some limitations with that its maximum of survival time is much larger than number of samples. In the meanwhile, DeepSurv model is not a fully parametric model.

3. Distribution Network

In this section, we describe our proposal, Parametric Distribution Estimation based Neural Network, depicted in Figure 1.

3.1 Survival Data

Survival data provides three type of information for each patient: i) observed covariates, ii) duration (time from start to event

ID patient	Age	Gender	...	T stage	N stage	Death event	Survival time (days)
LC04971	77.15068	Male	...	1	3	1	113
LC02966	68.31233	Female	...	2	1	0	1825
LC05230	73.64932	Male	...	3	2	1	362

Figure 2. Some samples of survival data.

occurrence), and iii) type of event (e.g., death or alive). Figure 2 shows some samples of survival data.

3.2 Distribution Network Architecture

The aim of survival analysis is to predict individual's survival curve, S(·|X) = P(T > t|X), which could be discrete, continuous or a mixture of both. In this study, we focus on the continuous probabilistic distribution. A very popular probabilistic distribution for survival is the Weibull distribution. The Weibull distribution is
usually known as a two-parametric probabilistic distribution since the location parameter is not used and set to zero. The Weibull distribution is presented as follows:

$$f(t,\alpha,\beta) = \beta\alpha t^{\alpha-1}e^{-\beta t^{\alpha}} \quad (1)$$

$$F(t,\alpha,\beta) = 1 - e^{-\beta t^{\alpha}} \quad (2)$$

$$S(t,\alpha,\beta) = 1 - F(t,\alpha,\beta) \quad (3)$$

Where $f(.)$ is the probability density function, $F(.)$ is the cumulative density function, $S(.)$ is the survival function, α is the shape parameter, β is the scale parameter and t are timepoints beyond the observed range.

The α and β parameters are learnt from input covariates and further generate probability density distribution and survival distribution of each patient beyond the observed range of time.

3.3 Objective function

As we aforementioned that survival dataset includes non-censored and censored data. Therefore, survival regression (or survival analysis) differs from conventional regression approaches. It takes both non-censored and censored data which usually is seen as missing

data. Herein, we apply maximum likelihood formulation to approximate the distribution of the survival duration distribution of the dataset.

- For non-censored data, we attempt to maximize the ELBO loss as follows:

$$\begin{aligned}\ln P(N|\boldsymbol{\Theta}) &= \ln\left(\prod_{i=1}^{|N|} P(T=t_i|X=x_i,\boldsymbol{\Theta})\right)\\ &\geq \sum_{i=1}^{|N|}\left(E[\ln P(T=t_i|X=x_i,\boldsymbol{\Theta}]\right)\\ &= \mathbf{ELBO}_N(\boldsymbol{\Theta})\end{aligned} \quad (4)$$

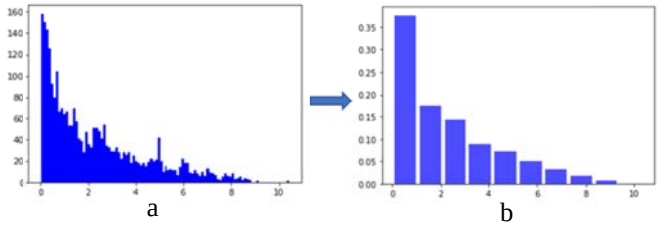

Figure 3. a) Distribution of survival time, b) Probability density of survival time.

- For censored data, we also apply ELBO loss with a change to fit with this type of data.

$$\begin{aligned}\ln P(C|\boldsymbol{\Theta}) &= \ln\left(\prod_{i=1}^{|C|} P(T>t_i|X=x_i,\boldsymbol{\Theta})\right)\\ &\geq \sum_{i=1}^{|N|}\left(E[\ln P(T>t_i|X=x_i,\boldsymbol{\Theta}]\right)\\ &= \mathbf{ELBO}_C(\boldsymbol{\Theta})\end{aligned} \quad (5)$$

Where $\boldsymbol{\Theta}$ is the weights of model, N is the set of non-censored data and C is the set of censored data.

To handle imbalance data, we propose a novel module to enhance effect of minor group. As we can see in Figure 3 (left), the distribution of survival time mostly distributed as early timepoint and decreased beyond follow-up time.

Therefore, it leads to less contribute to the loss function at latter timepoints and mostly bias on early timepoints. We come up with an idea that we first calculate probability density distribution of survival duration and then split into 10 bins as illustrated in Figure 3 (right). We classify which group the patients belong to base on their survival time. Finally, we integrate above loss functions to generate overall loss as follows:

$$Loss = \mathbf{ELBO}_N(\boldsymbol{\Theta}) * e^{1-p_n} + \Upsilon * \mathbf{ELBO}_C(\boldsymbol{\Theta}) * e^{1-p_c} \quad (6)$$

Where p_n and p_c is probability density of non-censored patients and censored patients, respectively. Υ is the scalar coefficient that presents the effect of loss of censored data to the combined loss.

4. Experimental result

We conducted comprehensive experiments on the dataset collected from Chonnam National University hospital (CNUH), South Korea. We also evaluate and compare our proposal with existing methods.

4.1 Dataset

This dataset was collected and verified by doctors at CNUH. The dataset consists of 2690 non-small cell lung cancer (NSCLC) patients with survival time, death event and 8 input covariates such as histology, age, gender, overall stage, M stage, N stage, T stage, smoking status and smoking amount. All patients were anonymized and replaced by ID number before analysis. We split data into two sets (training and testing set) with ratio of 80:20. In training set, we apply 5-fold cross-validation to evaluate model in training process.

Table 1. Comparison to prior methods (higher is better).

Model	C-index
CoxPH [5]	0.7555 ± 2.45e-4
DeepSurv [6]	0.7562 ± 9.13e-4
DeepHit [7]	0.7579 ± 1.73e-3
Our proposal	**0.7653 ± 1.64e-3**

4.2 Experimental settings

All experiments were conducted in Pytorch framework and trained on the Geforce RTX 2080ti GPU. In our model, we apply two fully-connected (FC) layers [100, 100] to learn the input covariates. After each FC layer, we use BatchNorm [12] and ReLU [13] to normalize and convert model to non-linear model. We utilize RMSprop [14] optimizer with a learning rate of 0.001, 100 epochs to optimize all of weights in our model. The batch size is set to the number of training samples to capture distribution of whole training set over every iteration. The Υ coefficient is chosen from {0.1, 0.3, 0.5, 0.7, 1}.

4.3 Experimental result

In this study, we implement existing methods which achieve remarkable performance in terms of C-index metric [15]. C-index is the most common metric to evaluate performance of survival models. It takes non-censored and censored data. C-index has ability to correctly provide a reliable ranking of the survival duration based on the individual's risk score. The C-index is measured as follows:

$$C-index = \frac{\sum_{i,j} 1_{T_i<T_j} * 1_{p_i<p_j} * e_i}{\sum_{i,j} 1_{T_i<T_j} * e_i} \quad (7)$$

Where T_i, T_j, e_i are survival time of patient i, j and event of interest of patient i, respectively. And p_i, p_j are survival probability of patient i and j, respectively.

The comparison results are shown in Table 1. Our proposed method achieves the best performance in terms of C-index with 0.7653. In addition, all pairwise comparisons were statistically significant ($p < 0.05$).

5. Conclusion

Survival analysis is an import and necessary technique in medical field to help doctors outline risk of death and propose proper strategies for patients. In this study, we leverage development of machine learning methods to estimate individual's survival curve. The study results demonstrated that our method can estimate survival curve more accurately than existing methods. Further, in order to obtain helpful features from CT and PET images, we will integrate all of them to boost performance.

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) and This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (NRF-2020R1A2B5B01002085).

References

[1] M. Richardson, E. Dominowska and R. Ragno. Predicting clicks: Estimating the Click-Through Rate for New Ads. In WWW, pp.521-529. ACM Press, 2007.

[2] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," Statistics in Medicine, vol. 32, no. 13, pp. 2173-2184, 2013.

[3] M. Luck, T. Sylvain, H. Cardinal, A. Lodi and Y. Bengio. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. arXiv:1705.10245v1 [cs.LG], 2017.

[4] S. Matsuno, T. Ito, Y. Uchida and T. Ito. Lifespan of information service firms in Japan: a survival analysis. International Journal of Information Systems and Project Management, Doi: 10.12821/ijispm060104, 2017.

[5] A. Vigan, M. Dorgan, J. Buckingham, E. Bruera and M. E. Suarez-Almazor. Survival prediction in terminal cancer patients: a systematic review of the medical literature. Palliative Medicine, 14(5):363-374, 2000.

[6] D. V. d. Poel and B. Lariviere. Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research, 157(1):196-217, 2004.

[7] L. Dirick, C. Claeskens and B. Baesens. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 68(6):652-665, 2017.

[8] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics, 11(3):812-820, 2015.

[9] D. R. Cox. Regression models and life-tables. Journal of Royal Statistical Society. Series B (methodological), 34(2):187-220, 1972.

[10] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology, vol. 18, no. 1, p. 24, 2018.

[11] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[12] S. Loffe. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32th International Conference on International Conference on Machine learning, vol. 37, 448-456, July 2015.

[13] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In International Conference on Machine Learning (ICML), 2010.

[14] T. Tieleman and G. Hinton. Lecture 6.5-RMSProp, COURSERA: Neural Networks for Machine Learning. Technical report, 2012.

[15] F. E. Harrell, R. M. Califf, D. B Pryor, K. L. Lee and R. A. Rosati. Evaluating the yield of medical tests. Journal of the American Medical Association 247(18):25432546, 1982.