Multi-Task with Variational Autoencoder for Lung Cancer Prognosis on Clinical Data

Thanh-Hung Vo
Chonnam National University
Gwangju, South Korea

Guee-Sang Lee
Chonnam National University
Gwangju, South Korea

Hyung-Jeong Yang
Chonnam National University
Gwangju, South Korea

Sae-Ryung Kang
Chonnam National University
Gwangju, South Korea

In-Jae Oh
Chonnam National University
Gwangju, South Korea

Soo-Hyung Kim
Chonnam National University
Gwangju, South Korea
shkim@jnu.ac.kr

ABSTRACT

Due to the increase of lung cancer in Korea, survival analysis for this kind of cancer gets emerging in recent years. Statistical and traditional machine learning methods usually used by medical doctors for this task. Which the success of deep learning in many tasks of computer vision, natural language processing, some studies starting to use DL for this task. Differ than many fields, data in medicine is difficult to collect and process, then the number of samples usually small and a little bit difficult to apply deep learning approach. In this study, we apply variational autoencoder together with the normal task of survival analysis and analysis the effect of it's it on the target task. The results show that when combine the VAE with the target task, the network architecture less sensitive with the training size, and then could be trained with small number of sample. The limit of this study is using the internal dataset, then it is difficult to compare to the others.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Variational autoencoder, clinical, tabular, lung cancer

1 INTRODUCTION

Lung cancer is the leading cause of death among many types of cancer worldwide [8]. Similar, in Korea, lung cancer is the most kind of cancer cause death [1]. There are two types of lung cancer, including non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). Survival analysis (SA) for lung cancer is a task that analyzes the expected duration of time from the time of cancer detection until the death of the patient. The input for SA maybe includes the image, clinical data, microRNA, etc. While image, microRNA, and other kinds of data type are widely used for analysis, the clinical data got less consider. In this study, we do the SA task based on clinical data information. Clinical data is in the form of tabular, which includes many rows, each row for one patient, and many columns where each column is one filed, such as age, sex, smoking status. Some fields in the tabular data are in numerical, some in categorical kind and the other in text or some other forms.

Naive Bayes, Decision Tree, Random Forest, SVM, and traditional machine learning methods were used for the SA task, such as in [2, 6, 7]. Most of them only support numerical data; only several algorithms could work with categorical data. In recent years, Deep learning (DL) got the promising results in many tasks such as computer vision, natural language processing, there are some studies that build the neutron network to solve the SA [3]. In this study, we focus on clinical data, while other types are left for future work.

The DL method is a hungry of data, that mean to applying it to special task, a large of data sample is needed. But, it is a little bit difficult to get the large sample in the field of medicine due to some high requirement of collect and process data of patient. With the small amount of sample, DL easy to get overfitting. Variational autoencoder (VAE) is one of method that could be help in this situation [5, 9]. It is a probabilistic models, that can be learn by stochastic process from dataset based on DL algorithm.

The rest of the paper is as follows. The section 2 shows the approach. The next section 3 gives information about the dataset that to be used in this study. The experimental and results are presented in section 4. Finally, section 5 is the conclusion and discussion about future work.

2 PROPOSED METHOD

2.1 Multi-task network architecture with variational autoencoder and classifier

Fig. 1 shows the proposed network architecture, named MVAEC. The network includes three separated blocks including one encoder and two decoder. In this network, we combine variational autoencoder and classifier together. For this task, full connected neuron network, DNN, is used for all three parts. The left part is an encoder, which received a vector of real number as input and generate two vectors, μ and σ. Two vectors are mean and variance of the encode of the input in the latent space.

There are two decoders in the right of fig. 1, survival classifier in the top and reconstruction in the bottom. Classifier module is the original task which return the score of each class. The reconstruction module aims to rebuild the original input from latent vector.

SMA 2020, September 17-19, 2020, Jeju, Republic of Korea

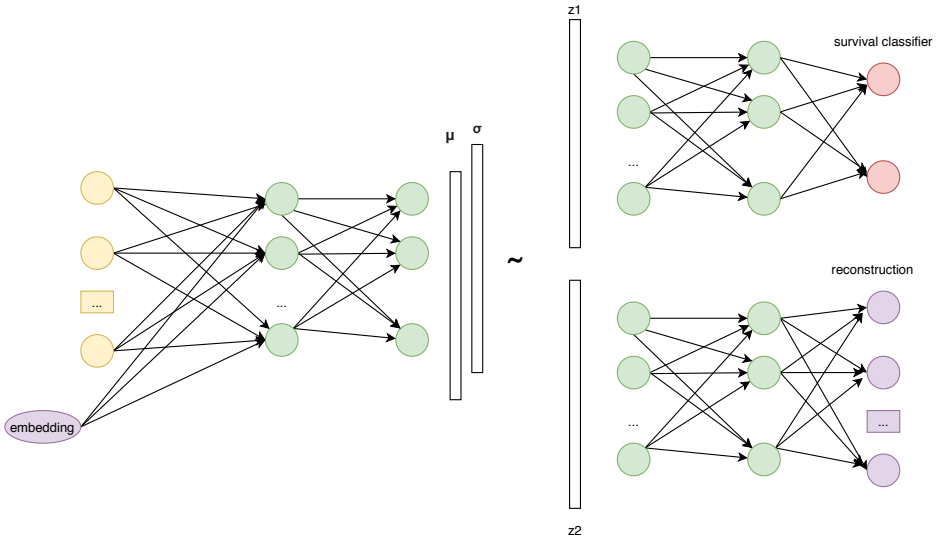

Figure 1: VAE multi-task network architecture

The input of each decoder modules are separated sampling from μ and σ with Gaussian distribution by reparameterization trick [4].

2.2 Loss function

Our network was designed for multi-task including variational autoencoder part and survival classifier part, the loss function should be designed to the network learn both tasks together.

The loss function is the combination of $\mathcal{L}_{VAE}$ and $\mathcal{L}_{Classifier}$ which given as eq. 1.

$$\mathcal{L} = \mathcal{L}_{VAE} + \mathcal{L}_{Classifier} \tag{1}$$

VAE loss function, $\mathcal{L}_{VAE}$, is the combination of two partitions, reconstruction, and KL divergence. Reconstruction loss is binary cross-entropy, calculate the difference between the output and the original input.

For the loss of the classifier part, we use cross-entropy as usual.

3 DATASET

The Chonnam National University Hospital Clinical dataset (CNUHC) was used to analysis in our study. The dataset is collected in Chonnam National University Hospital from 2007 to 2020 by medical doctors. The data is the information of 2298 patients. There are four categories and 13 continue fields in clinical and the survival time that need to be predicted.

The research flow is shows in Fig. 2. The total samples in the CNUHC are $n = 2298$. We excluded $n = 102$ samples, which missing information. The remaining $n = 2196$ then filter to remove unfollowing up where patients discontinue with the hospital, may be dead or more to another. The rest of the dataset $n = 2107$ is consider to analysis.

Table 1 shows some statistical information on selected dataset. The mean of age is 68.59 (95% CI 68.2 to 69). There are 509 females (24.15%) and 1598 males (75.84%). With 95% confidence the population survival time mean is between 1.94 and 2.1, based on 2107 samples. In this study, we focus on the classification problem, where the survival time less or equal 1.5 year (low) and mark as high if

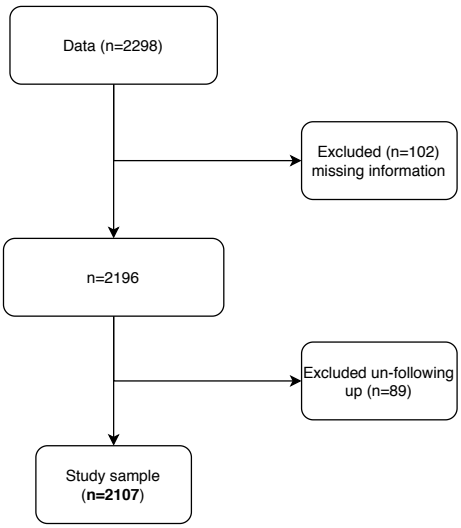

Figure 2: The pipeline of project.

Table 1: Patient characteristics (n=2107)

Characteristic	Value
Age (years), mean (95% CI)	68.59 (68.20-69.00)
Gender (female/male), n(%)	509 (24.15%)/1598 (75.84%)
Survival time (years), mean (95% CI)	2.02 (1.94-2.10)
Survival classes (Low/High), n (%)	1136 (53.92%)/971 (46.08%)

this value more than 1.5 years. Low survival time rate is 53.92%, and high rate is 46.08%.

4 EXPERIMENTAL AND RESULTS

4.1 Pre-processing

The pre-processing step includes some smaller steps. Firstly, the data need to clean. All missing information rows were removed. There are also $n = 89$ patients who did not follow up with the hospital, maybe they were dead or moved to treat in another hospital. Those rows were also removed before analysis. The next step is data transform. Some binary categorical fields can easily convert to a binary number, that 0 or 1.

4.2 Experimental setup

Because the number of samples in our dataset is small, we use the k-fold cross-validation $k = 5$. There are four fields in the form of categorical that need embedding step to transform to numerical data type including *mcode*, *description*, *histology*, and *overall stage*. The embedding size are set as 5, 5, 3, and 3, respective. There are nine fields in the form of numerical value. So the input size pass to the network is

$$input\_size = n_{numeric-field} + n_{embedding-size} = 9 + 16 = 25$$

Latent space dimension was set to 32, that means $\mu, \sigma \in R^{32}$.

In the reconstruction part, the number of output nodes was set equal to the input of the network after embedding, 25. The number of nodes in the classifier was set to 2, low and high.

4.3 Results

We analyze the effect of the sample size on the accuracy of the classifier. While our dataset is small, the total number of patients is 2107, we randomly division to 5 separated folds. For each round, we keep test fold as original while keeping the training and validation set from the four other folds with the *rate* from 0.1 to 1.0 with step 0.1. Fig. 3 shows the results. With full dataset for training and validation ($rate = 1.0$), our network architecture archive accuracy at 70.89%. When the keep rate decreasing, due to the number of samples to training lower, the accuracy starting to drop. While accuracy drop is a normal trend, in this case, the speed is slow. At the lowest keep rate, $rate = 0.1$, meaning we have only about 120 samples, very small to DL algorithm, the accuracy is 65.09%. The results show that event the number of samples is small, the network could learn for target task. This is because the VAE module in our architecture, it is a probabilistic model, that can learn from the small data points.

5 CONCLUSION

In this study, we analyze the affect of the VAE in the lung cancer prognosis. We proposed the network architecture to deal with the small number of sample in the clinical data problem. We add VAE to the target network, classification in this case. The results show that VAE helps the target network less sensitive with the number of samples in the training set. Besides that, the network architecture has some weaknesses. Firstly, the network architecture is not deep enough, it depends on the number of features in the tabular dataset is small and the number of samples too. Second, currently, we only analyze the sensitivity with the number of samples for training without comparison to other methods, that need future work.

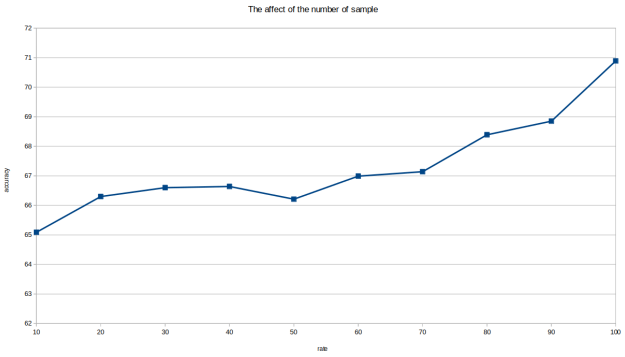

Figure 3: The effect of the number of sample to the performance

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). and by a grant (HCRI 19 001-1*H2019-0295) Chonnam National University Hwasun Hospital Institute for Biomedical Science. The corresponding author is Soo-Hyung Kim.

REFERENCES

[1] Jung Chi Young Cho Deog Gon Jeon Jae Hyun Lee Jeong Eun Ahn Jin Seok Kim Seung Joon Kim Yeongdae Choi Yoo-Duk Suh Yang-Gun Kim Jung-Eun Lee Boram Won Young-Joo Kim Young-Chul Choi Chang-Min, Kim Ho Cheol. 2019. Report of the Korean Association of Lung Cancer Registry (KALC-R), 2014. *Cancer Res Treat* 51, 4 (2019), 1400–1410. https://doi.org/10.4143/crt.2018.704

[2] Yuming Jiang, Jingjing Xie, Zhen Han, Wei Liu, Sujuan Xi, Lei Huang, Weicai Huang, Tian Lin, Liying Zhao, Yanfeng Hu, Jiang Yu, Qi Zhang, Tuanjie Li, Shirong Cai, and Guoxin Li. 2018. Immunomarker Support Vector Machine Classifier for Prediction of Gastric Cancer Survival and Adjuvant Chemotherapeutic Benefit. *Clinical Cancer Research* 24, 22 (2018), 5574–5584. https://doi.org/10.1158/1078-0432.CCR-18-0848

[3] Tejaswini Mallavarapu Jung Hun Oh Mingon Kang Jie Hao, Youngsoon Kim. 2019. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Med Genomics* 189 (2019). https://doi.org/10.1186/s12920-019-0624-2

[4] Diederik P. Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*.

[5] Carnegie Mellon and U C Berkeley. 2016. Tutorial on Variational Autoencoders. In *arXiv*. 1–23.

[6] Yip Cheng Har Pietro Lio Sarinder Kaur Dhillon Mogana Darshini Ganggayah, Nur Aishah Taib. 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak* 48 (2019). https://doi.org/10.1186/s12911-019-0801-4

[7] Wei C Price D Zhang F Courneya KS Kakadiaris IA. Paxton RJ, Zhang L. 2019. An exploratory decision tree analysis to predict physical activity compliance rates in breast cancer survivors. *Ethn Health* 24, 7 (2019), 754–766. https://doi.org/10.1080/13557858.2017.1378805

[8] Lindsey A. Torre, Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. 2015. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians* 65, 2 (2015), 87–108. https://doi.org/10.3322/caac.21262 arXiv:https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21262

[9] Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*. https://doi.org/10.1016/j.neuroimage.2016.11.058