P-DEEPSURV: A CANCER SURVIVAL PREDICTION SYSTEM BASED ON DEEPSURV (COMBINED WITH A FILTER LAYER)

QIUXIA ZHANG[1], JIALIN SUI[2], NINI ZHENG[3], XIAOHUI LI[4]

Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
E-MAIL: A10715006@mail.ntust.edu.tw

Abstract:

The cancer survival prediction system is a system that uses survival analysis to predict the survival time of cancer patients based on their survival information，such as genetic and clinical information. Deepsurv, a Cox proportional hazards deep neural network, uses a nonlinear deep neural network for modeling, and achieves the effect of a linear Cox proportional hazards model while increasing efficiency. But the Deepsurv does not perform well in high-dimensional real data sets. Therefore, P-Deepsurv, a Cox proportional hazards deep neural network combined with a filter layer based on Deepsurv is proposed by us. It uses statistical methods to reduce the dimensionality of high-dimensional data in the new filter layer. Through the controlled experiments conducted on the public data set, P-Deepsurv was proved to perform better than Deepsurv.

Keywords:

cancer survival prediction system; Cox proportional hazards deep neural network; survival analysis

1. Introduction

Over the years, the diagnosis rate and death rate of cancer have been increasing, so it is necessary to develop a cancer prediction system to effectively help us analyze the survival of cancer patients.

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems.

The traditional method used for survival analysis is linear Cox proportional hazards model (CPH)[1], but it requires extensive feature engineering or prior medical knowledge for feature filtering. Non-linear survival methods, such as neural network, can automatically model these features without manual feature engineering. But the performance of the traditional neural network models (i.e., classification methods[2][3], time-encoded methods[4][5], risk-predicting methods[6]) is not as good as CPH.

Deepsurv[7] is a Cox proportional hazards deep neural network proposed in 2018, which combines the advantages of deep neural networks and CPH, has achieved better results than CPH.

In the actual survival analysis process, it is not enough to simply consider the patient's clinical information, only by adding genetic information can make the survival analysis more accurate. However, Deepsurv didn’t perform well enough in real medical data sets, where clinical data and genetic information were added together.

The main reason is the survival data, including clinical and genetic data, have massive features and high redundancy. These massive features come from tens of thousands of people’s genetic information. Too many features will cause the dimensionality of the input layer become extremely high, which will increase the analysis time and reduce the accuracy rate.

In order to solve this problem, we built a new model based on Deepsurv. We added a filter layer before the input layer in Deepsurv to perform feature engineering. In the filtering layer, we use statistical methods to filter out invalid data of clinical data, and use edgeR[8] and DESeq[9] to screen out differential genes of genetic data.

We uploaded the project to GitHub, you can click on the link: https://github.com/ChoushaChang/P-Deepsurv to know more details.

2. Related works

Here are some related works about our system:

2.1. Gene differential expression analysis

Differential gene refers to genes that are different from normal genes due to structural changes such as gene mutations or methylation. Gene differential expression analysis is the process of analyzing the expression amount of genes.

2.2. DESeq

Deseq is an R language analysis suite that is used to analyze RNA-seq data and gene expression differences. In the random sampling and sequencing of simulated genes, the distribution of the data is estimated according to negative binomial distribution, and the FDR (false discovery rate) is used to calculate whether there are significant differences between the genes. DESeq can estimate local variation. Different variation parameters for different gene expression can reduce the bias caused by high expression and obtain more accurate results.

2.3. edgeR

edgeR implements novel statistical methods based on the negative binomial distribution as a model for count variability, including empirical Bayes methods, exact tests, and generalized linear models. The package is especially suitable for analysing designed experiments with multiple experimental factors but possibly small numbers of replicates.

It has unique abilities to model transcript specific variation even in small samples, a capability essential for prioritizing genes or transcripts that have consistent effects across replicates.

2.4. Deepsurv

Deepsurv proposed the use of COX based on deep learning for survival analysis and using it into clinic systems. Deepsurv uses a deeper network and more advanced training techniques, such as ReLU and Dropout. The output is a risk value which used C-index as evaluation indicator. The loss function of Deepsurv is as follows. h (x) is a risk function and is estimated through network θ’s weight. Finally, Deepsurv sets the loss function (1) to be the negative log partial likelihood of the equation:

$$l(\theta) := -\sum_{i:E_i=1}\left(\hat{h}_\theta(x_i) - \log\sum_{j\in\Re(T_i)} e^{\hat{h}_\theta(x_j)}\right). \tag{1}$$

At the same time, different treatment options will also affect the survival prediction of the patient. Based on the assumption that each individual has the same baseline hazard function λ0(t). Deepsurv take the log of the hazards ratio(2) to calculate the personal risk:

$$\begin{aligned}\mathrm{rec}_{ij}(x) &= \log\left(\frac{\lambda(t;x|\tau=i)}{\lambda(t;x|\tau=j)}\right) = \log\left(\frac{\lambda_0(t)\cdot e^{h_i(x)}}{\lambda_0(t)\cdot e^{h_j(x)}}\right)\\ &= h_i(x) - h_j(x).\end{aligned} \tag{2}$$

Note: τ are treatment groups.

Deepsurv's simulation experiment includes two types: linear and nonlinear. In linear risk model Deepsurv has the same performance as COX but it’s far better than COX in non-linear models.

3. P-Deepsurv

The architecture of P-Deepsurv is shown in Fig. 1: The filter layer receives the patient’s clinical information and genetic information as input. It maintains a table T, which containing differentially expressed genes’ name, non-differentially expressed genes are filtered out according to the table. The filtered data will be input to the Deepsurv neural network as the nodes of the input layer, a survival time will be obtained as the output.

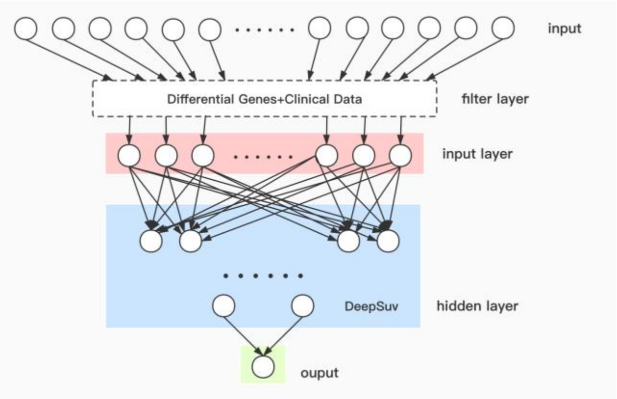

Fig. 1. The architecture of P-Deepsurv

This table is obtained by using DESeq and edgeR to perform differential expression analysis on the tumor tissue and normal tissue genes of cancer patients. After the analysis, we will get the table D(id, gene, baseMeanA, baseMeanB, foldchange, logFC, pval, padj) and table E (id, gene, logFC, Pval, FDR), you can get the specific principle, operation steps and description of attributes by referring [9] and [10].

Fold change refers to the multiple of the gene difference, where the values of D.logFC and E.logFC are the logarithmic values of fold change taking the base 2 value. The greater the absolute value of logFC, the more significant the gene difference. pval, means P-value, refers to the probability that the observation result of the sample or the more extreme result will appear when the hypothesis is ‘there is no difference between the gene in normal tissue samples and tumor samples’. D.padj is the adjusted value of D.pval, and E.FDR, the false discovery rate, is the value after E.pval adjusting.

The filter layer selects entries with logFC greater than

1 and padj less than 0.05 from Table D to obtain D’ (3), and selects entries with logFC greater than 1 and FDR less than 0.05 from Table E to obtain E’ (4). Then it respectively project the columns corresponding to the gene names from D’ and E’, perform the intersection operation on them, and finally obtain the table T (5). This process can be represented by the following Relational Algebra:

$$D' = \sigma_{\text{logFC>1 and padj<0.5}}(D) \tag{3}$$

$$E' = \sigma_{\text{logFC>1 and FDR<0.5}}(E) \tag{4}$$

$$T = (\Pi_{\text{gene}}(D')) \cap (\Pi_{\text{gene}}(E')) \tag{5}$$

Fig. 2 and Fig. 3 respectively show us the differential gene volcano plot[11] after differential expression analysis by DESeq and edgeR, among which the red and green dots are the differential genes the filter layer screened.

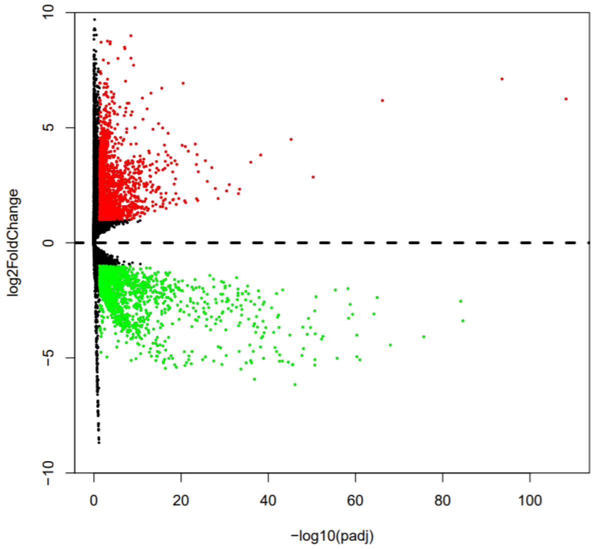

Fig. 2. Differential gene volcano plot after differential analysis using DESeq

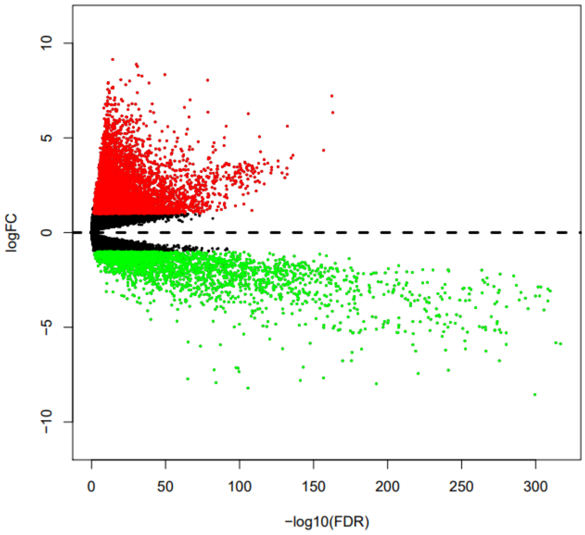

Fig. 3. Differential gene volcano plot after differential analysis using edgeR

In the Deepsurv network, in order to obtain the optimal network hyperparameters, we performed a random hyperparameter optimization search[12].

4. Experiment

In order to verify the feasibility of P-Deepsurv, we set up a control experiment to test the performance of P-Deepsurv.

4.1. Data set

In our experiment, we used TCGA-BRCA as our data set. It contains information of 1109 patients, which including 113 normal tissue samples and 1109 tumor tissue samples, and contains 60484 genes’ information. And we divided the data set into a training set and a testing set at a ratio of 4:1.

4.2. Validation metrics

To evaluate our model’s performance, we used two metrics in our experiment: C-index[13] and running time.

- **C-index**

C-index(concordance index) is used to estimate the predictive ability of the model. Its specific method is to randomly pair all the research objects into pairs and compute the proportion of all pairs whose predicted results are consistent with actual results. In Survival analysis, two patients are randomly selected. If the predicted survival time of one patient who has the longer real survival time is also longer than that of the other patient, it means the predicted result is consistent with actual result. The C-index is between 0.5 to 1. 0.5-0.7 is considered low accuracy, and 0.9 or more is considered high accuracy.

- **Running time**

To test the running time of the model, we calculate the running time by the CPU time's difference before and after modelling.

4.3. Control group

In order to check the distribution of difference genes, we first used DESeq and edgeR in filter layer to make differentially expressed gene analysis. When setting fold Change=1, P-value adjusted=0.5, 3618 difference genes were screened by DESeq and 9065 by edgeR. We combined their results and selected the top 10, 100, 1000, 2000, and 3000 genes, and combined them with clinical information as the control group used to measure P-Deepsurv.

At the same time we randomly selected 10, 100, 1000, 2000, 3000 genes to simulate Deepsurv's input without filter layer.

4.4. Result

We respectively trained Deepsurv and P-Deepsurv in the same hidden layer, the result of the C-index is as the Fig.4 to Fig.6 show.

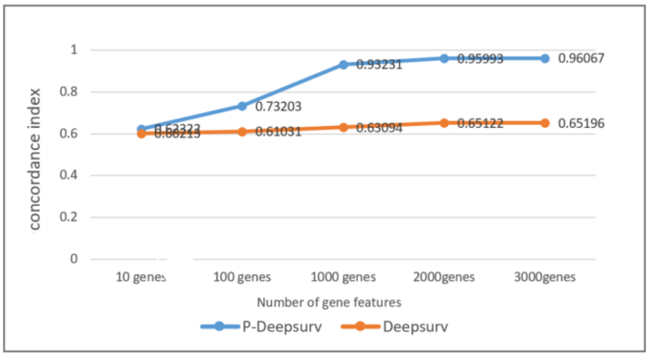

Fig. 4. C-index on the training set

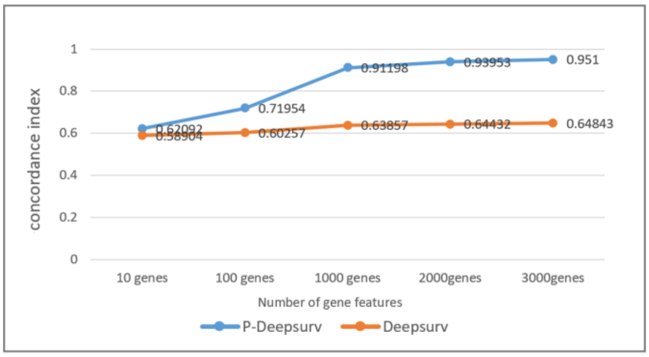

Fig. 5. C-index on the testing set

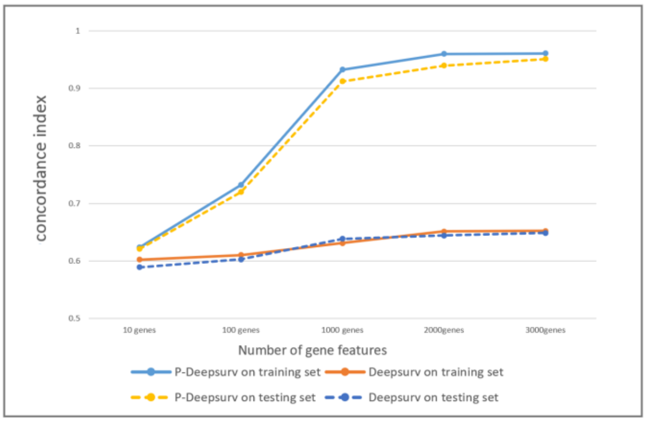

Fig. 6. Comparison of C-index using Deepsurv model and P-Deepsurv model

We can see that whether it's in training set or the testing set, P-Deepsurv has better survival prediction accuracy than Deepsurv, it means that our model is feasible. In addition, as the number of input features increases, the accuracy of P-Deepsurv will increase accordingly, while the increase of Deepsurv is not obvious. This shows that using prior knowledge to screen genes can effectively improve the accuracy of survival analysis.

The result of running time is shown in Fig. 7. We can find that the speed of P-Deepsurv is a little slower than Deepsurv due to the addition of filter layer.

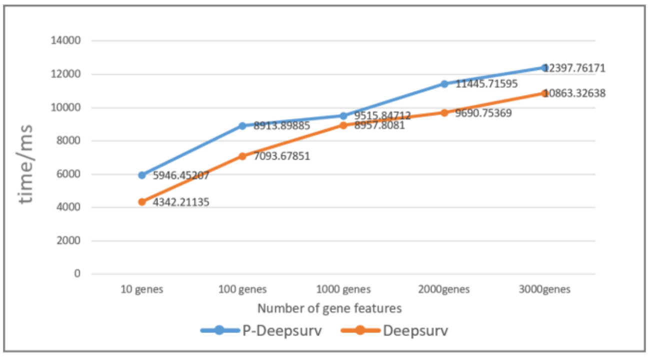

Fig. 7. Comparison of running time using Deepsurv model and P-Deepsurv model

During the experiment we found that when features were increased from 100 to 1000, the growth rate of P-Deepsurv running speed decreased. Therefore, considering both the prediction accuracy and the running speed, we choose 1000 as the number of gene screening in the filter layer.

5. Discussion

The results of running time in Fig. 7 may be a little confusing: We have reduced the dimensionality of the input layer by adding a new filter layer, which should reduce the training time of the Deepsurv network to a certain extent, but on the basis of Deepsurv, the P-Deepsurv's overall speed has decreased. We think that this is because the execution time of the filter layer is greater than the time saved by the dimensionality reduction operation, so the overall speed is reduced. It is difficult to balance accuracy and running speed at the same time. Although P-Deepsurv sacrifices performance in terms of running time to a certain extent, it still achieves relatively good results in terms of accuracy.

6. Conclusion

In this paper, we proposed a COX proportional hazard neural network model based on Deepsurv combined with a filter layer, and verified that our model perform better than Deepsurv. We use DESeq and edgeR in the new filter layer to effectively screen out the differential genes through statistical methods and reduce the dimensionality of the input layer. Although the addition of a filter layer makes the running speed of the whole model slower, it can effectively improve the accuracy of survival analysis.

References

[1] D. R. Cox, “Regression Models and Life-Tables”, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2 (1972), pp. 187-220, January 1972.

[2] Knut Liestbl, Per Kragh Andersen, and Ulrich Andersen, “Survival analysis and neural nets”, *Statistics in medicine*, vol.13, Issue12, pp.1189-1200, June 199a4.

[3] W Nick Street, “A neural network model for prognostic prediction”, in *ICML*, pp. 540–546, 1998.

[4] Leonardo Franco, Jose M Jerez, and Emilio Alba, “Artificial neural networks and prognosis in medicine. survival analysis in breast cancer patients”, in *ESANN*, pp.91–102. i6doc, January 2005.

[5] Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini, “Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach”, *Statistics in medicine*, vol.17, Issue10, pp.1169-1186, May 1998.

[6] David Faraggi and Richard Simon, “A neural network model for survival data”, *Statistics in medicine*, vol.14, Issue1, pp.73-82, January 1995.

[7] Jared L. Katzman, Uri Shaham, Alexander Cloninger, and et al., “DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network”, *BMC Medical Research Methodology*, Feb 2018.

[8] Mark D. Robinson, Davis J. McCarthy and Gordon K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”, *Bioinformatics*, vol. 26, Issue 1, pp.139–140, 1 January 2010.

[9] Simon Anders, Wolfgang Huber, “Differential expression analysis for sequence count data”, *nature precedings*, April 2010.

[10] Yunshun Chen, Aaron T. L. Lun and Gordon K. Smyth, “Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR”, in *Statistical Analysis of Next Generation Sequencing Data*, ch.3, pp.51-74, June 2014.

[11] Wentian Li, “Volcano Plots In Analyzing Differential Expressions With mRNA Microarrays”, *Journal of Bioinformatics and Computational Biology*, vol.10, no.6, October 2012.

[12] James Bergstra, Yoshua Bengio, “Random Search for Hyper-Parameter Optimization”, *The Journal of Machine Learning Research*, February 2012.

[13] FRANK E. HARRELL Jr., KERRY L. LEE and DANIEL B. MARK, “Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and

reducing errors", *Statistics in Medicine,* vol.15, Issue4, pp.361-387, February 1996.