Variable Selection in Gompertz Parametric Survival Regression Model

Taha dhaher Abd

Department of Statistics and Informatics, College of Computer science & Mathematics, University of Mosul, Mosul, Iraq

E-mail: tahadabd@gmail.com

Mohammed Khalid Mohammed Nory

Department of Statistics and Informatics, College of Computer science & Mathematics, University of Mosul, Mosul, Iraq

E-mail: mm98989844@gmail.com

Zakariya Yahya Algamal

Department of Statistics and Informatics, College of Computer science & Mathematics, University of Mosul, Mosul, Iraq

E-mail: zakariya.algamal@uomosul.edu.iq

Article Info

Page Number: 1426 - 1431

Publication Issue:

Vol 71 No. 3 (2022)

Abstract

The common issues of high dimensional gene expression data for survival analysis are that many of genes may not be relevant to their diseases. Gene selection has been proved to be an effective way to improve the result of many methods. The Gompertz parametric survival regression model is the most popular model in regression analysis for censored survival data. In this paper, an invasive weed optimization (IWO) as an evolutionary algorithm is employed in Gompertz proportional hazards regression model is proposed, with the aim of identification relevant genes and provides high classification accuracy. Experimental results show that the IWO significantly outperforms two competitor methods, AIC and BIC in terms of the area under the curve and the number of the selected genes.

Article History

Article Received*: 12 January 2022*

Revised*: 25 February 2022*

Accepted*: 20 April 2022*

Publication*: 09 June 2022*

Keywords: Gompertz distribution; parametric survival regression model; invasive weed optimization; gene selection.

1. Introduction

The problem of analyzing time to event data arises in a number of applied fields, such as medicine, biology, public health, and epidemiology [1, 2]. Nowadays, high dimensional gene expression data are increasingly used for modeling various clinical outcomes to facilitate disease diagnosis, disease prognosis, and prediction of treatment outcome [3].

Regression modeling is a standard practice to study jointly the effects of multiple predictors on a response. The parametric proportional hazards model is ubiquitous in the analysis of time-to-event data. When the number of predictors is large, building a parametric proportional hazards model including all of them is undesirable because it has low prediction accuracy and is hard to interpret [4, 5]. For these reasons, variable selection has become an important focus in parametric proportional hazards modeling.

2. Gompertz parametric survival regression model

It is considered one of the most popular distribution when analyzing survival data We know that the form of Gompertz distribution is:

$$f\left(t;\lambda,\theta\right)=\lambda e^{\theta t}\exp\left(\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)\right)\text{for t} \geq 0, \tag{1}$$

Where $\lambda,\theta>0$ is the scale and shape parameters. The CDF is

$$F\left(t;\lambda,\theta\right)=1-\exp\left(\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)\right)$$

(2)

The survival function and hazard function is defined as

$$S\left(t\right)=\exp\left\{\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)\right\} \tag{3}$$

$$h\left(t;\lambda,\theta\right)=\lambda e^{\theta t} \tag{4}$$

The parametric proportional hazards model of Gompertz distribution is

$$h_i\left(t\right)=\exp(\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi})\lambda e^{\theta t} \tag{5}$$

Then

$$S\left(t\right)=\exp\left\{\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)\right\}.\exp\left(x_j^T\beta\right) \tag{6}$$

The log-likelihood of Eq. (5) is

$$\ln L\left(\beta',\lambda,\theta\right)=\sum_{i=1}^{n}\left[\ln\left(\lambda e^{\theta t}.\exp\left(x_j^T\beta\right)\right)\right]+\ln\left[\exp\left\{\frac{\lambda}{\theta}\left(1-e^{\theta t}\right)\right\}.\exp\left(x_j^T\beta\right)\right]$$

(7)

Then the estimated proportional hazards and survival function are

$$\hat{h}\left(t\mid x_j\right)=\hat{\lambda}e^{\hat{\theta}t}\exp\left(\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p\right) \qquad (8)$$

$$\hat{S}\left(t\right)=\exp\left\{\frac{\hat{\lambda}}{\hat{\theta}}\left(1-e^{\hat{\theta}t}\right)\right\}.\exp\left(\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p\right)(9)$$

3. Invasive Weed Optimization Algorithm(IWO)

Invasive weed optimization (IWO) is an evolutionaryalgorithm inspired from the biological behavior of weeds.The characteristics of IWO algorithm are its robustness, adaptation andrandomness which make it more effective for global search. The following equations are used for IWO algorithm:

$$seed_i=floor\left(\frac{f_i-f_{\min}}{f_{\max}-f_{\min}}\left(S_{\max}-S_{\min}\right)\right)+S_{\min} \qquad (10)$$

$$\sigma_{iter}=\frac{\left(iter_{\max}-iter\right)^n}{\left(iter_{\max}\right)^n}\left(\sigma_{initial}-\sigma_{final}\right)+\sigma_{final} \qquad (11)$$

$$\chi_{son}=\chi_{parent}+sd=\chi_{parent}+random*\sigma_{iter} \qquad (12)$$

The proposed variable selection is as following:

(1) The number of weeds, is set to 30 and the number of iterations is $t_{\max}=250$.

(2) The positions of each weed are randomly determined. The position of a weed represents the variables. The initial positions of the weeds are generated from a uniform distribution within the range [0,1].

(3) The fitness function is defined as

$$\text{fitness}=\min\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \qquad (12)$$

(4) The positions are updated using Eq. (12).

(5) Steps 3 and 4 are repeated until a $t_{\max}$ is reached.

4. Real Application

To evaluate the performance of the proposed method, three real gene datasets were used. A brief introduction and summary of the used datasets are given in Table 1. The first dataset is the Diffuse large B-cell lymphoma dataset (DLBCL) [6]. There are 240 lymphoma patients' samples. Each patient's data consists 7399 gene expression measurements, and its survival time, including censored or not.

The second dataset is the Dutch breast cancer dataset (DBC) [7]. In this dataset, there was 295 breast cancer patients' information collected in this dataset. Each patient's data consist 4919 gene expression measurements.

The third dataset is the Lung cancer dataset (LC) [8]. This dataset contains 86 lung cancer patients' information including 7129 gene expression measurements, survival time and whether the survival time is censored.

Table 1: The detail of the used three real microarray datasets

Dataset	Sample	Gene	Censored
DLBCL	240	7399	102
DBC	295	4919	207
LC	86	7129	62

To demonstrate the usefulness of the proposed method, comparative experiments with the AIC and BIC are conducted. To do so, each gene expression dataset is randomly partitioned into the training dataset and the test dataset, where 70% of the sample are selected for training dataset and the rest 30% are selected for testing datasetTo assess how well the model predicts the outcome, the idea of time-dependent receiver-operator characteristics (ROC) curves for censored data and area under the curve (AUC) as our criteria. The real application results are summarized in Tables 2 – 4.

Table 2 shows the average results of different used methods applied to the three real datasets. It is obviously that the numbers of genes selected by AIC are much more than those of the BIC and the IWO. Among the other two methods, the IWO selected the least subset of genes.

Table 2: The selected genes results

	AIC	BIC	IWO
DLBCL	100	81	33
DBC	77	61	42

LC	81	73	22

In order to test the prediction accuracy of the different used methods, their average values of AUC for both the training and testing dataset were given in Tables 3 and 4, respectively. In the observation of Table 3, in terms of AUC, the IWO achieved a maximum accuracy of 97.6%, 98.2% and 99.5% for DLBCL, DBC, and LC datasets, respectively. Furthermore, it is clear from the results that the IWO outperformed the AIC and BIC for all datasets. Moreover, the IWO improved the classification accuracy compared to AIC. The improvements were 10.5%, 11.2%, and 9.7% for the DLBCL, DBC, and LC datasets, respectively.

Table 3: The AUC results for the training dataset

	AIC	BIC	IWO
DLBCL	0.892	0.933	0.976
DBC	0.905	0.945	0.982
LC	0.922	0.958	0.995

It can also be seen from Table 4 that the proposed method has the best results in terms of the AUC for the testing dataset. The IWO has the largest AUC of 95.5%, 96.8%, and 97.9% for the DLBCL, DBC, and LC datasets, respectively. This indicated that the IWO significantly succeeded in identifying the patients who are in fact having the cancer with a probability of greater than 0.95.

Table 4: The AUC results for the testing dataset

	AIC	BIC	IWO
DLBCL	0.875	0.926	0.955
DBC	0.835	0.932	0.968
LC	0.903	0.944	0.979

5. Conclusions

This paper presents Gompertz proportional hazards regression model by employing the IWO algorithm to identify the relevant genes in gene expression data. Our proposed method was experimentally tested and compared with other existing methods. The superior prediction performance of the proposed method was shown through the AUC. Meeting this criterion nominates the proposed method as a promising gene selection method.

REFERENCES

1. Cockeran, M., S.G. Meintanis, and J.S. Allison, *Goodness-of-fit tests in the Cox proportional hazards model.* Communications in Statistics - Simulation and Computation, 2019: p. 1-12.
2. Emura, T., Y.H. Chen, and H.Y. Chen, *Survival prediction based on compound covariate under Cox proportional hazard models.* PLoS One, 2012. **7**(10): p. e47627.
3. Huang, J., et al., *Group selection in the Cox model with a diverging number of covariates.* statistica Sinica, 2014.
4. Karabey, U. and N.A. Tutkun, *Model selection criterion in survival analysis.* 2017. **1863**: p. 120003.
5. Leng, C. and H. Helen Zhang, *Model selection in nonparametric hazard regression.* Journal of Nonparametric Statistics, 2006. **18**(7-8): p. 417-429.
6. Rosenwald, A., et al., *The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.* New England Journal of Medicine, 2002. **346**(25): p. 1937-1947.
7. van Houwelingen, H.C., et al., *Cross-validated Cox regression on microarray gene expression data.* Statistics in medicine, 2006. **25**(18): p. 3201-3216.
8. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nature medicine, 2002. **8**(8): p. 816.