Analyzing High-Dimensional Survival Data Using Random Forests

Motarjem, K.[1], Mohammadzadeh, M.[1] and Abyar, A.[1]

[1] *Department of Statistics, Tarbiat Modares University, Tehran, Iran.*
E-mail: k.motarjem@modares.ac.ir

Abstract:

Recently developments in genetics create cases that contain numerous covariates and few observations. These kinds of data are called High-dimensional data. In these cases, use of usual statistical methods is not appropriate. Especially in survival analysis, modeling of high-dimensional survival data is of utmost importance. One of the usual methods for analyzing survival data is the use of logistic regression. In this paper will be shown that the performance of logistic regression is not acceptable for analyzing high-dimensional survival data [1]. Generalized Breimanâ€■s Random Forests method for analyzing high-dimensional survival data, namely Random Survival Forests (RSF), has a strong efficiency for modeling survival data [2-3]. The main advantage of this method is that there is no distributional assumption on the regression model. Then the efficiency of RSF and usual methods for analyzing survival data such as logistic regression model are compared for modeling a real high-dimensional survival dataset and results are evaluated.

Finally using random forests and logistic regression models, a real dataset has been modeled and the accuracy of results are compared by Akaike Information Criterion [4]. It is shown that the random forests method and logistic regression model may presented different significant covariates. So, based on the data conditions the researchers should precisely choose the models.

Keywords: Random Survival Forests, Logistic regression, High-dimensional data, Survival analysis.

2010 Mathematics Subject Classification: 62F01, 62F02, 62F99

References

[1] Mittal, S. and Madigan, D., "High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis", *Biostatistics*, Vol. 15, pp. 1â€“15, 2014.
[2] Braiman, L., "Random forests, Machine Learn", Vol. 45, pp. 5-32, 2001.
[3] Genuer, R., "Random Forest: some methodological insights", 2008.
[4] Akaike, H., "Information theory and the extension of the maximum likelihood principle", *In: Petrov, V., Csaki, F., Editors, Proceedings of the Second International Symposium on Information Theory*, Budapest: Akailseoniaikiudo, pp. 267-281, 1973.