IMPROVING THE OPTIMAL CUT-POINT IDENTIFICATION OF CONTINUOUS COVARIATES FOR SURVIVAL OUTCOME PREDICTION WITH APPLICATION TO PUBLIC REAL DATA

Rancoita Paola M.V.[1]

[1]*University Centre for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffele, Milan, Italy*

Background: In the clinical practice, the categorization of clinical or biological continuous variables is commonly considered to develop classification rules for diagnosis or prognosis, which in turn may lead to different decisions on patient's treatment or management. Selecting the optimal cut-point is fundamental to guarantee the best prediction of the patient outcome. For survival outcome, few methods exist in the literature, such as the survival tree and the method of Contal and O'Quigley [1], which is based on the log-rank statistic. However, they may be prone to overfitting, since they rely on an exhaustive search of the optimal cut-point among the observed data.

Aim: To overcome the limitations of existing approaches, a new method is proposed for optimal cut-point identification in case of survival outcome prediction. Its performance has been evaluated and compared to that of more standard approaches.

Methods: Assuming a proportional hazards model, the proposed method estimates the cut-point as a parameter of the Cox's regression model. The estimation requires the numerical maximization of the partial likelihood, which is not straightforward, since it may present many local maxima. Thus, alternative maximization procedures were explored by specifying a grid of initial values of the parameters and different criteria for selecting the maximum. In a simulation study, the version of the proposed method, using a criterion considering also the log-rank statistics, resulted the best one.
An extensive simulation study was conducted to compare the performance of the proposed method with the one of the survival tree and the Contal and O'Quigley's method. The simulated data were generated mimicking challenging issues typical of real-world data. Different scenarios were realized by varying: the sample size, the percentage of censored data, the distribution of the continuous covariate, the balancing of the two groups defined by the true cut-point, and the distance between the corresponding survival curves.
All the methodologies were also compared by using real training-test public data on breast cancer ([2] for the training, [3-6] for the test) and diffuse large B-cell lymphoma ([7] for the training, [8] for the test).

Results: In the simulation setting, the method of Contal and O'Quigley showed good performances only when the groups defined by the true cut-point were almost balanced, otherwise it got the worst estimation of the cut-off. Conversely the proposed method outperformed the survival tree in more challenging scenarios, involving asymmetrically distributed covariates or unbalanced groups and a high percentage of censored data.
When applied on real data, the proposed methodology provided a more robust estimation of the cut-point compared to the survival tree. In fact, for some variables, the latter method found a cut-point in the training set which showed less consistent results in predicting the survival outcome in the corresponding test set. For example, in the breast cancer context, the proposed methodology found a cut-off for the variable age of 36.41 corresponding to an hazard ratio, HR [95% confidence interval, CI] of 0.51 [0.34-0.58] in the training set and of 0.6 [0.38-0.94] in the test set. Instead, the survival tree selected as cut-off 31.5 corresponding to an HR [95% CI] of 0.18 [0.09-0.34] in the training set and of 0.48 [0.21-1.08] in the test set. Thus, the relative difference of the HR in the test set with respect to the training set was only 18% for the proposed method, while 167% for the survival tree. The percentage of 95% CI in the training set which was overlapping with the 95% CI in the test set was 91% for the proposed method, while only 52% for the survival tree.

Conclusions: Overall, the proposed methodology showed a better estimation of the cut-point in the survival setting compared to other methods in the literature on both simulated and public real data.

References

[1] Contal C., O'Quigley J., An application of changepoint methods in studying the effect of age on survival in breast cancer. Computational Statistics & Data Analysis, 1999; 30(3):253-70.
[2] Royston P., Ambler G., Sauerbrei W., The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol., 1999; 28(5):964-74.
[3] van 't Veer L.J., Dai H., van de Vijver M.J. et al., Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002; 415(6871):530-6.
[4] Miller K.D., Chap L.I., Holmes F.A. et al., Randomized phase III trial of capecitabine compared with bevacizumab plus capecitabine in patients with previously treated metastatic breast cancer. J Clin Oncol., 2005; 23(4):792-9.
[5] Sotiriou C., Wirapati P., Loi S. et al., Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst., 2006; 98(4):262-72.
[6] Desmedt C., Piette F., Loi S., et al., Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res., 2007; 13(11):3207–14.
[7] Lenz G., Wright G., Dave S.S. et al., Stromal gene signatures in large-B-cell lymphomas. N Engl J Med., 2008; 359(22):2313-23.
[8] Schmitz R., Wright G.W., Huang D.W. et al., Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. N Engl J Med., 2018; 378(15):1396-1407.