Bootstrap approach for multivariate survival analysis of cancer patients.

felipe rodolfo camargo dos santos, Gabriela Guardia, Pedro A F Galante

Instituto de Ensino e Pesquisa - Hospital Sírio Libanês

Abstract

Gene expression is an important factor correlated with survival of cancer patients. It has been shown that tumors harbour many synergistic and antagonistic interactions that impact disease progression. Given this complexity, the use of regression methods for survival analysis is a valuable resource for understanding the molecular profile of tumors and its impact on tumor progression, especially multivariate models, which take into account the important contribution of the context in which each variable is regarded. Despite their importance, the existing algorithms present some limitations, in particular for the analysis of high dimensional datasets, such as handling attribute/observation proportion issues (dimensionality problem) and control of attribute-attribute correlations. In order to achieve analysis convergence for high dimensional datasets we propose a bootstrap approach, making use of current regression methods that, in turn, use lasso and/or ridge parameters for the obtention of reliable regression coefficients in a bias-variance trade-off optimization. Our tool was applied in a real scenario for the investigation of survival differences between men and women patients with glioblastoma and other cancer types. These analyses revealed protein coding genes and long non-coding RNAs (lncRNAs) that may contribute to prognosis differences based on the gender of patients. Interestingly, these enriched lncRNAs are still poorly characterized and potentially subject to further investigations. Our results suggest that multivariate analysis combined with bootstrap algorithm improves prognostic prediction in comparison with commonly used methods that rely on univariate regression filters for lowering data dimensionality. In summary, we believe our method provides a more reliable and adjusted list of genes than current strategies because it takes into account gene-gene interactions.