28th IGES Annual Meeting, October 12-14, 2019, Houston TX, USA
Abstract for the poster session

TOWARDS AN ACCURATE CANCER DIAGNOSIS MODELIZATION: COMPARISON OF RANDOM FOREST STRATEGIES

Debit A[1,2,*], *Poulet C*[1], *Josse C*[1,4], *Azencott CA*[5], *Jerusalem G*[4], *Van Steen K*[2], *Bours V*[1,3]

[1]*University of Liege, GIGA-Research, Laboratory of Human Genetics, Liege, Belgium*
[2] *University of Liege, GIGA-Research, Medical Genomics, BIO3, Liege, Belgium*
[3]*University Hospital (CHU), Center of Human Genetics, Liege, Belgium*
[4]*University Hospital (CHU), Department of Medical Oncology, Liege, Belgium*
[5]*Centre for Computational Biology (CBIO) of Mines ParisTech, Institut Curie and INSERM., Paris, France*

[*] *adebit@uliege.be*

Abstract

Machine learning approaches are heavily used to produce models that will one day support clinical decisions. To be reliably used as a medical decision, such diagnosis and prognosis tools have to harbor a high-level of precision. Random Forests have been already used in cancer diagnosis, prognosis, and screening. Numerous Random Forests methods have been derived from the original random forest algorithm. Nevertheless, the precision of their generated models remains unknown when facing biological data. The precision of such models can be therefore too variable to produce models with the same accuracy of classification, making them useless in daily clinics. Here, we perform an empirical comparison of Random Forest based strategies, looking for their precision in model accuracy and overall computational time. An assessment of 15 methods is carried out for the classification of paired normal - tumor patients, from 3 TCGA RNA-Seq datasets: BRCA (Breast Invasive Carcinoma), LUSC (Lung Squamous Cell Carcinoma), and THCA (Thyroid Carcinoma). Results demonstrate noteworthy differences in the precisions of the model accuracy and the overall time processing, between the strategies for one dataset, as well as between datasets for one strategy. Therefore, we highly recommend to test several random forest strategies prior to modeling. This will certainly improve the precision in model accuracy while revealing the method of choice for the candidate data.