Unsupervised random forest for clustering

Master 2 Internship

Background

Random Forests[1] (RF) are a powerful machine learning technique. They are most often used for prediction, either in a regression context (with a continuous outcome) or in a classification context (with a categorical outcome). At their core, RF build upon classification and regression trees (CART), overcoming their shortcomings thanks to both random sampling (with replacement) of the observations at each tree and random sampling of variables (without replacement) at each node.

Besides, divisive clustering trees can be constructed for unsupervised classification (or clustering), by minimizing inertia while staying interpretable as decision trees[2]. This approach is implemented in the `divclust` R package.

Subject

This internship aims at leveraging both divisive clustering trees as well as the RF framework to propose a new clustering algorithm. A key contribution of this internship will be the definition of an aggregation strategy of clusterings across trees, which can rely on a consensus similarity matrix and its postprocessing (not unlike solutions adopted in non parametric Bayesian clustering approaches[3]).

Objectives

1. Implement *divclust random forests*, featuring the ensemble clustering (through partition aggregation across trees).
2. Investigate the impact of the RF tuning parameters on the results in numerical studies, in particular the number of trees, the number of randomly selected variables per nodes, and the trees depth.
3. Apply this new clustering approach to high-dimensional transcriptomics data in the context of vaccine development against EBOLA[4], HIV and COVID-19[5].

Required skills:

- Good knowledge in Biostatistics and/or Statistics
- Programming proficiency with R

- An interest for biomedical research, and in particular in vaccine research
- English proficiency (both written and spoken)
- Scientific curiosity
- Master 1/Bachelor/Engineering school with a major in Biostatistics and/or Statistics

Hosting laboratory:
SISTM team
Inria Bordeaux Sud-Ouest & Inserm U1219 *Bordeaux Population Health*

Location:
Inserm U1219 *Bordeaux Population Health* research center – SISTM team
Université de Bordeaux – ISPED
146, rue Léo Saignat
33076 Bordeaux Cedex

Duration:
Internship of 4 to 6 month available starting from January 2024.

Compensation:
Intern gratification according to the official recommendations (15% of social security ceiling, i.e. around 625€/month).

Contact:
Send a detailed CV and a motivation letter to both **Boris Hejblum** [boris.hejblum@u-bordeaux.fr] & **Robin Genuer** [robin.genuer@u-bordeaux.fr]

Bibliography

1. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

2. Chavent, M., Lechevallier, Y. & Briant, O. DIVCLUS-t: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis* **52**, 687–701 (2007).

3. Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F. & Thiébaut, R. Sequential dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data. *Annals of Applied Statistics* **13**, 638–660 (2019).

4. Rechtien, A. *et al.* Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. *Cell Reports* **20**, 2251–2261 (2017).

5. Lévy, Y. *et al.* CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. *iScience* **24**, 102711 (2021).