


Unsupervised random forests for clustering

Master 2 Internship

Background

Random Forests¹ (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 drawbacks thanks to both i) random sampling (with replacement) of the observations in each tree, and ii) random sampling of variables (without replacement) at each tree node.

RF-related approaches have already been proposed in unsupervised contexts. Notably, some approaches generate synthetic data and apply RF to discriminate between those and real data, finally obtaining a proximity matrix^{2,3}. Besides, Yan *et al.* propose to apply the k-means on several resampled versions of the data to obtain a set of partitions, that are then all aggregated (through spectral clustering) — an algorithm that takes some *inspirations* from RF⁴. However, none of these previous works actually propose an extension of RF suitable to clustering problems.

Divisive clustering trees can be constructed for unsupervised classification (or clustering), by minimizing inertia while staying interpretable as decision trees⁵. This approach is implemented in the `divclust`  package. We propose to use those to propose unsupervised random forests of clustering trees.

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 approaches⁶).

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 EBOLA⁷, HIV and COVID-19⁸.

Required skills:

- Good knowledge in Biostatistics and/or Statistics
- Programming proficiency with 
- 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](#)

Centre Inria de l'Universit  de Bordeaux & Inserm U1219 *Bordeaux Population Health*

Location:

[Inserm U1219 Bordeaux Population Health research center – SISTM team](#)

Universit  de Bordeaux – ISPED

146, rue L o 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](mailto:boris.hejblum@u-bordeaux.fr) [boris.hejblum@u-bordeaux.fr] & [Robin Genuer](mailto:robin.genuer@u-bordeaux.fr) [robin.genuer@u-bordeaux.fr]

Bibliography

1. Breiman, L. [Random forests](#). *Machine learning* **45**, 5–32 (2001).
2. Shi, T. & Horvath, S. [Unsupervised learning with random forest predictors](#). *Journal of Computational and Graphical Statistics* **15**, 118–138 (2006).
3. Afanador, N. L., Smolinska, A., Tran, T. N. & Blanchet, L. [Unsupervised random forest: A tutorial with case studies](#). *journal of Chemometrics* **30**, 232–241 (2016).

4. Yan, D., Chen, A. & Jordan, M. I. [Cluster forests](#). *Computational Statistics & Data Analysis* **66**, 178–192 (2013).
5. Chavent, M., Lechevallier, Y. & Briant, O. [DIVCLUS-t: A monothetic divisive hierarchical clustering method](#). *Computational Statistics & Data Analysis* **52**, 687–701 (2007).
6. Hejblum, B. P., Alkassim, C., Gottardo, R., Caron, F. & Thi baut, 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).
7. Rechten, 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).
8. L vy, Y. *et al.* [CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death](#). *iScience* **24**, 102711 (2021).