Group input identification based on the random forests for grouped inputs algorithm: application on aircraft engine data.

6 month-intership proposal possibly following by a PhD thesis (CIFRE thesis)

Institutions:	Université Bretagne Sud (UBS), LMBA CNRS 6205 and Safran Aircraft Engines.
Lieu:	Campus Tohannic, Vannes, France.
Duration:	6 months.
Supervisors :	Audrey Poterie (UBS) & Erwan Scornet (École Polytechnique) & Jérôme Lacaille (SAFRAN) François Septier & Emmanuel Frenod (UBS).
Contacts:	`audrey.poterie@univ-ubs.fr`

Keywords : Random forests for grouped inputs, group importance score, group identification, model interpretability.

Subject

Supervised learning consists in explaining and/or predicting an output variable by using some inputs. Here, we consider the context in which the inputs have a known and/or obvious group structure. In many supervised problems, inputs can have a group structure or groups of inputs can be defined to capture the underlying input associations. In these cases, the study of groups of variables can make more sense than the study of inputs taken individually. For instance, when we are given a large number of inputs that are correlated, using groups of inputs instead of individual inputs can allow to "control" high correlations and can make model understanding easier. Indeed, in the analysis of gene expression data it has become frequent to use in the analysis group of genes representing putative biological processes instead of individual genes [1,2]. Furthermore, when the inputs are some observations of a time series, it could be interesting to cluster the inputs into groups that represent several time periods. Several methods have already been proposed to deal with this problem. For instance, the regression regularized by the Group Lasso penalty (GL) enables to elaborate prediction rules based on groups of input variables [3,4]. Recently, two decision tree approaches and one random forests algorithm *–called Random Forests for Grouped Inputs (RFGI)–* have been developped to deal with groups of inputs [5,6]. These methods build tree-based prediction rules based directly on groups of inputs and do not need any input-transformation process. Furthermore in addition to the prediction purpose, these three new methods can also be used to perform group variable selection thanks to the introduction for each of them of some grouped variable importance scores.

Objective

The internship will focus on Random Forests for Grouped Inputs (RFGI) [6]. More precisely, the project will consist of the following three objectives.

The first objective will be to study the grouped importance score [5]. This score is the natural extension of the mean decrease in accuracy (MDA) introduced by [7]. Some recent papers highlight some MDA's weaknesses and suggest using other importance scores. Then, the first goal will aim at assessing and improving performance of the group importance score.

In supervised problems in which inputs have a group structure, groups are often poorly known or sometimes even completely unknown. Then, this internship's main objective will be to develop an original and data-driven method to perform group identification. The proposed strategy will use the RFGI algorithm. Some approaches based on the grouped importance score [5] and a wrapper strategy such as for instance the methods introduced in [8–10] could be studied. Other methods computationally more efficient could also be proposed.

Finally, as with any ensemble method, RFGI models are not directly interpretable. Moreover, the group importance score only enables the identification of relevant predictor groups and does not provide any information about model interpretability [11–13]. So, the third objective of this internship will be to develop a tool to provide insights into how groups of inputs and outputs are related and overcome the lack of interpretability of RFGI models.

Each objective will start by establishing a state-of-the-art. Moreover, all methods developed will be thoroughly assessed through experimental studies on several synthetic data sets and on high-dimensional and highly correlated data about aircraft engines.

This project will be continued in a PhD thesis (CIFRE thesis).

Candidate profile

We are looking for a motivated and talented student in Master 2 (or equivalent) who:

- Has strong skills in statistics and known statistical learning. Being familliar random forests and clustering will be very appreciate.
- Has strong programming skills in R and Python. Skills in Cython and C++ will be appreciate.
- Interested in doing a PhD thesis.
- Is keen to develop new statistical methods and to implement them in Python (or Cython).
- Has good communication skills in French or in English (oral/reading/writing).

Details

This intership will be a collaboration between SAFRAN and the Laboratoire de Mathématiques Bretagne Athlantique (LMBA). The 6-month internship will take place at the Université Bretagne Sud in Vannes.

To apply for this position, the candidate is requested to firstly send a CV, a motivation letter and academic records to `audrey.poterie@univ-ubs.fr`. **The application deadline is: 31th January 2023**.

References

[1] P. Tamayo, D. Scanfeld, B. L. Ebert, M. A. Gillette, C. W. Roberts, and J. P. Mesirov, "Metagene projection for cross-platform, cross-species characterization of global transcriptional states," *Proceedings of the National Academy of Sciences*, vol. 104, no. 14, pp. 5959–5964, 2007.

[2] S.-I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, no. 11, p. R76, 2003.

[3] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[4] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[5] A. Poterie, J.-F. Dupuy, V. Monbet, and L. Rouvière, "Classification tree algorithm for grouped variables," *Computational Statistics*, vol. 34, no. 4, pp. 1613–1648, 2019.

[6] A. Poterie, *Arbres de décision et forêts aléatoires pour variables groupées*. PhD thesis, INSA de Rennes, 2018.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[9] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[10] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis," *Computational Statistics & Data Analysis*, vol. 90, pp. 15–35, 2015.

[11] S. Rüping *et al.*, "Learning interpretable models," 2006.

[12] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, "Interpretable predictions of tree-based ensembles via actionable feature tweaking," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 465–474, 2017.

[13] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.