Group variable identification based on the random forests for grouped variables algorithm.

6 month-internship proposal

Institutions: Université Bretagne Sud, LMBA CNRS 6205, IRISA CNRS 6074.
Lieu: Campus Tohannic, Vannes, France.
Duration: 6 months.
Supervisors: Audrey Poterie & Charlotte Pelletier.
Contacts: audrey.poterie@univ-ubs.fr, charlotte.pelletier@univ-ubs.fr.

Keywords: Random forests for grouped variables, group variable identification, recursive feature selection, C++.

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 example, in the analysis of gene expression data, datasets contain the expression levels of thousands of genes in a much smaller number of observations. Thus it has become frequent to use in the analysis only a small number of genes, which are clustered into several groups representing putative biological processes [2, 3]. Another example is functional data, like spectrometry data.
(see Figure 1), where researchers are often more interested by identifying discriminatory parts of the curve rather than individual wavelengths. Finally, in some supervised methods such as the CART algorithm [4], categorical inputs are converted into a group of dummy variables. Thus, it seems quite natural to treat the dummy variables related to a categorical variable as a group. In all these situations, elaborating a prediction rule based on groups of variables rather than on the individual variables can improve both interpretation and prediction accuracy [5]. Several methods have already been proposed to deal with this problem. For instance, the logistic regression regularized by the Group Lasso penalty (GL) enables to elaborate classification rules based on groups of input variables [6]. Recently, two decision tree approaches and one random forests algorithm—called Random Forests for Grouped Variables—have been developed to deal with groups of inputs [7, 8]. These methods build prediction rules based on groups of inputs. 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 the algorithm named Random Forests for Grouped Variables [8] (RFGV). The project will consist of three objectives.

The first objective will be to propose an efficient implementation of the RFGV algorithm (for both the training phase and the prediction phase). An implementation in R has already been proposed for binary classification problems. The first aim is twofold: (1) to improve this implementation by proposing a multi-threading version in C++/Java/Cython, and (2) to extend the implementation so as to deal with multi-class classification tasks and regression problems.

In supervised problems with grouped inputs, the group structure is often unknown. Then, the second objective will consist in developing an original and data-driven method to perform group variable identification. The proposed strategy will be based on the RFGV algorithm and its grouped importance score [7]. A wrapper strategy, such as the feature elimination algorithm proposed by [9] or the selection approach introduced by [5], could be tested. The method will be thoroughly assessed through experimental studies on several synthetic data sets.

Finally, the third objective will be to apply the proposed method on real data.

**Candidate profile**

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

- Has strong programming skills in at least one of the following languages: C++, Java, Cython. Skills in R and/or Python will be appreciated.
- Is keen to learn about random forests, feature selection and more generally supervised learning.
- Has good communication skills in French or in English (oral/reading/writing).

Being familiar with random forests will be appreciated.

**Details**

The 6-month internship will take place at the Université Bretagne Sud in Vannes. The salary will be around 500€ per month.

The student will be supervised by:

- Audrey Poterie: audrey.poterie@univ-ubs.fr,
- Charlotte Pelletier: charlotte.pelletier@univ-ubs.fr.
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 2021.

References


