VSURF : un package R pour la sélection de variables à l'aide de forêts aléatoires

R. Genuer[a], **J.-M. Poggi**[b] and **C.Tuleau-Malot**[c]

[a]Université de Bordeaux
ISPED, INSERM U-897, INRIA équipe SISTM
146, rue Léo Saignat, 33000 Bordeaux
Robin.Genuer@isped.u-bordeaux2.fr

[b]Université d'Orsay
Laboratoire de Mathématiques
Bâtiment 425, 91405 Orsay
Jean-Michel.Poggi@math.u-psud.fr

[c]Université de Nice Sophia Antipolis
CNRS, LJAD, UMR 7351
06100 Nice
Malot@unice.fr

Mots clefs : Forêts aléatoires, Sélection de variables, Package R

Variable selection is a crucial issue in many applied classification and regression problems. It is of interest for statistical analysis as well as for modelization or prediction purposes to remove irrelevant variables, to select all important ones or to determine a sufficient subset for prediction. These main different objectives on a statistical learning perspective involve variable selection to simplify statistical problems, to help diagnosis and interpretation, and to speed up data processing.
The authors have proposed a variable selection method based on random forests [4], and the aim of this presentation is to describe the (recently available on CRAN) associated R [5] package called **VSURF** and to illustrate its use on real datasets.
Introduced by Breiman [1], random forests (abbreviated RF in the sequel) is an attractive non-parametric statistical method to deal with such problems, since it requires only mild conditions on the model supposed to have generated the observed data. Indeed, since it is based on decision trees and it uses aggregation ideas, RF allow to consider in an elegant and versatile framework different models and problems, namely regressions, two-class or multiclass classifications.

The strategy

In Genuer et.al. [4] we have distinguished two variable selection objectives: interpretation and prediction. The first is to find important variables highly related to the response variable in order to select all the important variables, even with high redundancy. The second is to find a small number of variables sufficient to a good parsimonious prediction of the response variable. We have proposed the following two-step procedure, the first one is the same for the two situations while the second one depends on the objective:

- Step 1. Preliminary elimination and ranking:
  - Compute the RF scores of importance, order the variables in decreasing order of importance;
  - Cancel the variables of small importance (let m denotes the number of remaining variables).

- Step 2. Variable selection:
  - For *interpretation*: construct the nested collection of RF models involving the k first variables, for $k = 1$ to m and select the variables involved in the model leading to the smallest OOB error (estimate of the prediction error);
  - For *prediction*: starting from the ordered variables retained for interpretation, construct an ascending sequence of RF models, by invoking and testing the variables stepwise. The variables of the last model are selected.

VSURF in action

The dataset, called `srbct` and available in the R package mixOmics [2], allows us to illustrate our procedure in a high-dimensional multiclass classification framework.
This real dataset considered is relative to small round blue cell tumors of childhood. It is composed of a data frame, called `gene`, of size 63×2308 which contains the 2308 gene expression; and a response factor of length 63, called `class`, indicating the class of each sample (there are 4 classes).

```
library(VSURF)
library(mixOmics)
data(srbct)
```

```
vSRBCT <- VSURF(x = srbct$gene, y = srbct$class)
```

```
summary(vSRBCT)

##
##  VSURF computation time: 2.7 hours
##
##  VSURF selected:
##  651 variables at thresholding step (in 7.2 mins)
##  25 variables at interpretation step (in 2.6 hours)
##  13 variables at prediction step (in 14.6 secs)
```

On this dataset, the procedure leads to 25 and 13 selected variables after the interpretation and prediction step respectively, and the number of selected variables as the selected variables themselves are stable.
We can compare these numbers of selected variables with those obtained in Díaz-Uriarte et.al. [3] where the authors select 22 genes on the original dataset and their number of selected variables is also quite stable.

Références

[1] Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5-32
[2] Dejean, S., Gonzalez, I., Le Cao, KA with contributions from Monget, P., Coquery, J., Yao, F., Liquet, B. and Rohart, F. (2013). ***mixOmics:*** *Omics Data Integration Project*. R package version 5.0-1
[3] Díaz-Uriarte, R. and Alvarez De Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(1), 3
[4] Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, **31**(14), 2225-2236
[5] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/