

Variable Selection using Random Forests

Jean-Michel Poggi¹

Université Paris-Sud, e-mail: Jean-Michel.Poggi@math.u-psud.fr

Robin Genuer

Université Paris-Sud, e-mail: Robin.Genuer@math.u-psud.fr

Christine Tuleau

Université Nice Sophia Antipolis, e-mail: Christine.Malot@unice.fr

Abstract: This paper deals with variable selection using random forests. The main contribution is methodological and is twofold. First, we provide some insights about the behavior of the variable importance index based on random forests. Second, we propose to investigate two classical issues of variable selection: to find important variables for interpretation and the more restrictive one, to design a good prediction model. The proposed strategy involves a ranking of explanatory variables using the random forests score of importance and a stepwise ascending variable introduction. In addition, this procedure uses automatic data-driven threshold parameters.

Keywords: Random Forests, Regression, Classification, Variable Importance, Variable Selection

1. Random Forests

The principle of random forests (RF henceforth), introduced by Breiman in 2001, is to combine many binary decision trees (for classification or regression problems) built using several bootstrap samples coming from the learning sample L . Instead of CART model building strategy, each tree of the forest is obtained by recursive partitioning, choosing randomly at each node a subset (of given size) of explanatory variables X , and without pruning step so each tree is a maximal one.

RF algorithm becomes more and more popular and appears to be very powerful in a lot of different applications (see for example Díaz-Uriarte and Alvarez de Andrés (2006) for gene expression data analysis) even if it is not clearly elucidated from a mathematical point of view (see the recent paper by Biau *et al.* (2008)).

2. Variable importance

In the RF framework, the most widely used score of importance of a given variable is the increasing in mean of the error of a tree (MSE for regression and misclassification rate for classification) in the forest when the observed values of this variable are randomly permuted in the OOB samples. The OOB sample is the set of observations which are not used for building the current tree, and is used to estimate the prediction error and

¹Address of correspondence: Université Paris-Sud, Laboratoire de Mathématiques, Bât. 425, 91405 Orsay, France & Université Paris Descartes

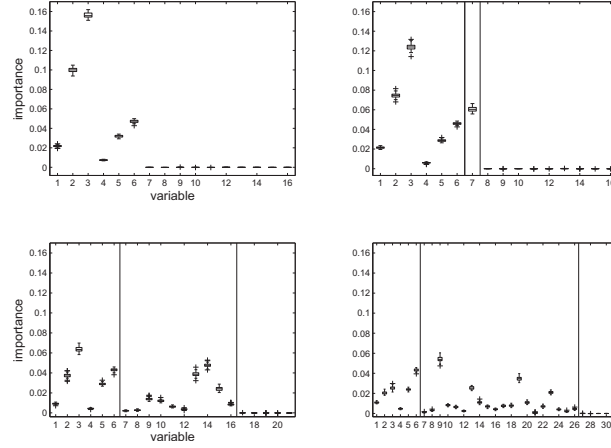
also to evaluate variable importance. Often, such random forests VI is called permutation importance indices in opposition to total decrease of node impurity measures already introduced in CART procedure.

We investigate in this paper some crucial questions about the VI behavior like the importance of a group of variables or its behavior in presence of highly correlated variables.

Let us examine a simulated dataset for the case $n \ll p$, called “toys data” in the sequel. It is an equiprobable two-class problem, $Y \in \{-1, 1\}$, with 6 true variables, the others being some noise. This example is interesting since it constructs two near independent groups of 3 significant variables (highly, moderately and weakly correlated with response Y) and an additional group of noise variables, uncorrelated with Y .

For example, let us illustrate the behavior in presence of several highly correlated variables. Starting from the previous framework with $n = 100$, $p = 200$, $ntree = 2000$ and $mtry = 100$. Then we add highly correlated replications of some of the 6 true variables.

Figure 1: Variable importance of a group of correlated variables



The first graph of Figure 1 is the reference one and for the three other cases, we simulate 1, 10 and 20 variables with a correlation of 0.9 with variable 3 (the most important one). These replications are plotted between the two vertical lines. As it can be seen, the magnitude of importance of the group 1, 2, 3 is steadily decreasing when adding more replications of variable 3. On the other hand, the importance of the group 4, 5, 6 is unchanged. Notice that the importance is not divided by the number of replications. Indeed in our example, even with 20 replications the maximum importance of the group containing variable 3 (that is variable 1, 2, 3 and all replications of variable 3) is only three times lower than the initial importance of variable 3. Finally, let us note that even if some variables in this group have low importance, they cannot be confused with noise.

3. Variable selection

Many variable selection procedures are based on the cooperation of variable importance for ranking and model estimation to evaluate and compare a family of models (for example Díaz-Uriarte, Alvarez de Andrés (2006) and Ben Ishak, Ghattas (2008)). We distinguish two variable selection objectives: to find important variables highly related to the

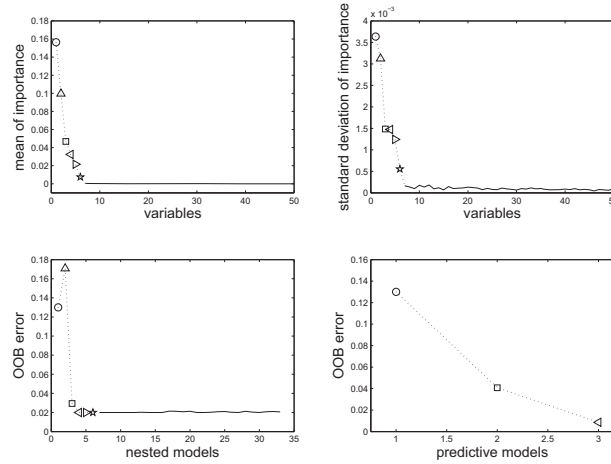
response variable for interpretation purpose; to find a small number of variables sufficient to a good prediction of the response variable. The first is to magnify all the important variables, even with high redundancy, for interpretation purpose and the second is to find a sufficient parsimonious set of important variables for prediction.

We propose the following two-steps procedure, the first one is common while the second one depends on the objective:

1. Preliminary elimination and ranking: cancel the variables of small RF importance; and order the m remaining variables.
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;
 - 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.

To both illustrate and give more details about this procedure, we apply it on a simulated learning set of size $n = 100$ from the classification toys data model (see) with $p = 200$. The results are summarized in Figure 2. The true variables (1 to 6) are respectively represented by ($\triangleright$, $\triangle$, $\circ$, $\star$, $\triangleleft$, $\square$). We compute, thanks to the learning set, 50 forests with $ntree = 2000$ and $mtry = 100$, which are values of the main parameters considered as well adapted for VI calculations.

Figure 2: *Variable selection for interpretation and prediction*



Let us detail the main stages of the procedure together with (in *italics*) the results obtained on toys data:

- First we rank the variables by sorting the VI in descending order.
The result is drawn on the top left graph for the 50 most important variables. Note that true variables are significantly more important than the noisy ones.
- We keep this order in mind and plot the corresponding standard deviations of VI.
We use this graph to estimate some threshold for importance, and we keep only the

variables of importance exceeding this level. More precisely, we select the threshold as the minimum prediction value given by a CART model fitting this curve. This rule is, in general conservative and leads to retain more variables than necessary in order to make a careful choice later.

The standard deviations of VI can be found in the top right graph. We can see that true variables standard deviation is large compared to the noisy variables one, which is close to zero. The threshold leads to retain 33 variables.

- Then, we compute OOB error rates of random forests (using default parameters) of the nested models starting from the one with only the most important variable, and ending with the one involving all important variables kept previously. The variables of the model leading to the smallest OOB error are selected.

Note that in the bottom left graph the error decreases quickly and reaches its minimum when the first 4 true variables are included in the model. Then it remains constant. We select the model containing 4 of the 6 true variables. More precisely, we select the variables involved in the model almost leading to the smallest OOB error, i.e. the first model almost leading to the minimum. The actual minimum is reached with 24 variables.

The expected behavior is non-decreasing as soon as all the "true" variables have been selected. It is then difficult to treat in a unified way nearly constant or slightly increasing: we propose to use an heuristic rule similar to the 1 SE rule of Breiman used for CART cost-complexity pruning procedure.

- We perform a sequential variable introduction with testing: a variable is added only if the error gain exceeds a threshold. The idea is that the error decrease must be significantly greater than the average variation obtained by adding noisy variables. *The bottom right graph shows the result of this step, the final model for prediction purpose involves only variables 3, 6 and 5. The threshold is set to the mean of the absolute values of the first order differentiated errors between the model with 5 variables (the first model after the one we selected for interpretation, see the bottom left graph) and the last one.*

We then illustrate the procedure on numerous artificial as well real classification or regression datasets, being standard or high dimensional (depending on n the number of observations, and p the number of variables: $n \gg p$ or $n \ll p$ resp.).

References

- Ben Ishak A. and Ghattas B. (2008) Sélection de variables en classification binaire : comparaisons et application aux données de biopuces, *Revue SFdS-RSA*, 149.
- Biau G., Devroye L. and Lugosi G. (2008) Consistency of random forests and other averaging classifiers, *Journal of Machine Learning Research*, 9, 2039–2057.
- Breiman L. (2001) Random forests, *Machine Learning*, 45, 5–32.
- Breiman L., Friedman J., Olshen R. and Stone C. (1984) *Classification And Regression Trees*, Chapman & Hall.
- Díaz-Uriarte R. and Alvarez de Andrés S. (2006) Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, 7.
- Genuer R., Poggi J.M. and Tuleau C. (2008) Random forests: some methodological insights, <http://hal.inria.fr/inria-00340725/fr/>, *Technical Report INRIA RR-6729*, 1–42.