Forêts aléatoires : remarques méthodologiques

Robin Genuer, Jean-Michel Poggi, Christine Tuleau

Forêts aléatoires : remarques méthodologiques

Robin Genuer & Jean-Michel Poggi & Christine Tuleau

Université Paris-Sud, Mathématique, Bât. 425, 91405 Orsay, France

Résumé : On s'intéresse à la méthode des forêts aléatoires d'un point de vue méthodologique. Introduite par Leo Breiman en 2001, elle est désormais largement utilisée tant en classification qu'en régression avec un succès spectaculaire. On vise tout d'abord à confirmer les résultats expérimentaux, connus mais épars, quant au choix des paramètres de la méthode, tant pour les problèmes dits "standards" que pour ceux dits de "grande dimension" (pour lesquels le nombre de variables est très grand vis à vis du nombre d'observations). Mais la contribution principale de cet article est d'étudier le comportement du score d'importance des variables basé sur les forêts aléatoires et d'examiner deux problèmes classiques de sélection de variables. Le premier est de dégager les variables importantes à des fins d'interprétation tandis que le second, plus restrictif, vise à se restreindre à un sous-ensemble suffisant pour la prédiction. La stratégie générale procède en deux étapes : le classement des variables basé sur les scores d'importance suivi d'une procédure d'introduction ascendante séquentielle des variables.

Môts clés : Forêts aléatoires, Régression, Classification, Importance des Variables, Sélection des Variables.

Abstract: This paper examines from an experimental perspective random forests, the increasingly used statistical method for classification and regression problems introduced by Leo Breiman in 2001. It first aims at confirming, known but sparse, advice for using random forests and at proposing some complementary remarks for both standard problems as well as high dimensional ones for which the number of variables hugely exceeds the sample size. But the main contribution of this paper is twofold: to provide some insights about the behavior of the variable importance index based on random forests and in addition, to propose to investigate two classical issues of variable selection. The first one is to find important variables for interpretation and the second one is more restrictive and try to design a good prediction model. The strategy involves a ranking of explanatory variables using the random forests score of importance and a stepwise ascending variable introduction strategy.

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

Random forests (RF henceforth) is a popular and very efficient algorithm, based on model aggregation ideas, for both classification and regression problems, introduced by Breiman (2001). It belongs to the family of ensemble methods, appearing in machine learning at the end of nineties (see for example Dietterich (1999) and (2000)). Let us

briefly recall the statistical framework by considering a learning set $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ made of n i.i.d. observations of a random vector (X, Y). Vector $X = (X^1, ..., X^p)$ contains predictors or explanatory variables, say $X \in \mathbb{R}^p$, and $Y \in \mathcal{Y}$ where $\mathcal{Y}$ is either a class label or a numerical response. For classification problems, a classifier t is a mapping $t : \mathbb{R}^p \to \mathcal{Y}$ while for regression problems, we suppose that $Y = s(X) + \varepsilon$ and s is the so-called regression function. For more background on statistical learning, see Hastie *et al.* (2001). Random forests is a model building strategy providing estimators of either the Bayes classifier or the regression function.

The principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the learning sample L and choosing randomly at each node a subset of explanatory variables X. More precisely, with respect to the well-known CART model building strategy (see Breiman *et al.* (1984)) performing a growing step followed by a pruning one, two differences can be noted. First, at each node, a given number (denoted by $mtry$) of input variables are randomly chosen and the best split is calculated only within this subset. Second, no pruning step is performed so all the trees are maximal trees.

In addition to CART, another well-known related tree-based method must be mentioned: bagging (see Breiman (1996)). Indeed random forests with $mtry = p$ reduce simply to unpruned bagging. The associated R packages are respectively `randomForest` (intensively used in the sequel of the paper), `rpart` and `ipred` for CART and bagging respectively (cited here for the sake of completeness).

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) and Bühlmann, Yu (2002) for bagging). Nevertheless, Breiman (2001) sketches an explanation of the good performance of random forests related to the good quality of each tree (at least from the bias point of view) together with the small correlation among the trees of the forest, where the correlation between trees is defined as the ordinary correlation of predictions on so-called out-of-bag (OOB henceforth) samples. The OOB sample which is the set of observations which are not used for building the current tree, is used to estimate the prediction error and then to evaluate variable importance.

Tuning method parameters

For details about tuning method parameters, see Genuer *et al.* (2008).

RF variable importance

The quantification of the variable importance (VI henceforth) is an important issue in many applied problems complementing variable selection by interpretation issues. In the linear regression framework it is examined for example by Grömping (2007), making a distinction between various variance decomposition based indicators: "dispersion impor-

tance", "level importance" or "theoretical importance" quantifying explained variance or changes in the response for a given change of each regressor. Various ways to define and compute using R such indicators are available (see Grömping (2006)).

In the random forests 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. Often, such random forests VI is called permutation importance indices in opposition to total decrease of node impurity measures already introduced in the seminal book about CART by Breiman *et al.* (1984).

Even if only little investigation is available about RF variable importance, some interesting facts are collected for classification problems. This index can be based on the average loss of another criterion, like the Gini entropy used for growing classification trees. Let us cite two remarks. The first one is that the RF Gini importance is not fair in favor of predictor variables with many categories while the RF permutation importance is a more reliable indicator (see Strobl *et al.* (2007)). So we restrict our attention to this last one. The second one is that it seems that permutation importance overestimates the variable importance of highly correlated variables and they propose a conditional variant (see Strobl *et al.* (2008)). Let us mention that, in this paper, we do not notice such phenomenon. For classification problems, Ben Ishak, Ghattas (2008) and Díaz-Uriarte, Alvarez de Andrés (2006) for example, use RF variable importance and note that it is stable for correlated predictors, scale invariant and stable with respect to small perturbations of the learning sample. But these preliminary remarks need to be extended and the recent paper by Archer *et al.* (2008), focusing more specifically on the VI topic, do not answer some crucial questions about the variable importance behavior: like the importance of a group of variables or its behavior in presence of highly correlated variables. This one is the second goal of this paper.

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. Three types of variable selection methods are distinguished (see Kohavi *et al.* (1997) and Guyon *et al.* (2003)): "filter" for which the score of variable importance does not depend on a given model design method; "wrapper" which include the prediction performance in the score calculation; and finally "embedded" which intricate more closely variable selection and model estimation.

For non-parametric models, only a small number of methods are available, especially for the classification case. Let us briefly mention some of them, which are potentially competing tools. Of course we must firstly mention the wrapper methods based on VI coming from CART, see Breiman *et al.* (1984) and of course, random forests, see Breiman (2001). Then some examples of embedded methods: Poggi, Tuleau (2006) propose a method based on CART scores and using stepwise ascending procedure with elimination

step; Guyon *et al.* (2002) (and Rakotomamonjy (2003)), propose SVM-RFE, a method based on SVM scores and using descending elimination. More recently, Ben Ishak *et al.* (2008) propose a stepwise variant while Park *et al.* (2007) propose a "LARS" type strategy (see Efron *et al.* (2004) for classification problems.

Let us recall that two distinct objectives about variable selection can be identified: (1) to find important variables highly related to the response variable for interpretation purpose; (2) to find a small number of variables sufficient for a good prediction of the response variable. The key tool for task 1 is thresholding variable importance while the crucial point for task 2 is to combine variable ranking and stepwise introduction of variables on a prediction model building. It could be ascending in order to avoid to select redundant variables or, for the case $n << p$, descending first to reach a classical situation $n \sim p$, and then ascending using the first strategy, see Fan, Lv (2008). We propose in this paper, a two-steps procedure, the first one is common while the second one depends on the objective interpretation or prediction.

Bibliographie

[1] Archer K.J. and Kimes R.V. (2008) *Empirical characterization of random forest variable importance measures.* Computational Statistics & Data Analysis 52:2249-2260

[2] Ben Ishak A. and Ghattas B. (2008) *Sélection de variables en classification binaire : comparaisons et application aux données de biopuces.* To appear, Revue SFDS-RSA

[3] Biau G., Devroye L., and Lugosi G. (2008) *Consistency of random forests and other averaging classifiers.* Journal of Machine Learning Research, 9:2039-2057

[4] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) *Classification And Regression Trees.* Chapman & Hall

[5] Breiman, L. (1996) *Bagging predictors.* Machine Learning, 26(2):123-140

[6] Breiman L. (2001) *Random Forests.* Machine Learning, 45:5-32

[7] Breiman L. (2004) *Consistency for a simple model of Random Forests.* Technical Report 670, Berkeley

[8] Breiman L. and Cutler, A. (2005) *Random Forests.* Berkeley,
http://www.stat.berkeley.edu/users/breiman/RandomForests/

[9] Bühlmann, P. and Yu, B. (2002) *Analyzing Bagging.* The Annals of Statistics, 30(4):927-961

[10] Cutler A. and Zhao G. (2001) *Pert - Perfect random tree ensembles.* Computing Science and Statistics, 33:490-497

[11] Díaz-Uriarte R. and Alvarez de Andrés S. (2006) *Gene Selection and classification of microarray data using random forest.* BMC Bioinformatics, 7:3, 1-13

[12] Dietterich, T. (1999) *An experimental comparison of three methods for constructing ensembles of decision trees : Bagging, Boosting and randomization.* Machine Learning, 1-22

[13] Dietterich, T. (2000) *Ensemble Methods in Machine Learning.* Lecture Notes in Computer Science, 1857:1-15

[14] Efron B., Hastie T., Johnstone I., and Tibshirani R. (2004) *Least angle regression*. Annals of Statistics, 32(2):407-499
[15] Fan J. and Lv J. (2008) *Sure independence screening for ultra-high dimensional feature space*. J. Roy. Statist. Soc. Ser. B, 70:849-911
[16] Genuer R., Poggi J.-M., Tuleau C. (2008) *Random Forests: some methodological insights*. RR-6729 - http://hal.inria.fr/inria-00340725/fr/
[17] Guyon I., Weston J., Barnhill S., and Vapnik V.N. (2002) *Gene selection for cancer classification using support vector machines*. Machine Learning, 46(1-3):389-422
[18] Guyon I. and Elisseff A. (2003) *An introduction to variable and feature selection*. Journal of Machine Learning Research, 3:1157-1182
[19] Grömping U. (2007) *Estimators of Relative Importance in Linear Regression Based on Variance Decomposition*. The American Statistician 61:139-147
[20] Grömping U. (2006) *Relative Importance for Linear Regression in R: The Package relaimpo*. Journal of Statistical Software 17, Issue 1
[21] Hastie T., Tibshirani R., Friedman J. (2001) *The Elements of Statistical Learning*. Springer
[22] Ho, T.K. (1998) *The random subspace method for constructing decision forests*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(8):832-844
[23] Kohavi R. and John G.H. (1997) *Wrappers for Feature Subset Selection*. Artificial Intelligence, 97(1-2):273-324
[24] Liaw A. and Wiener M. (2002). *Classification and Regression by randomForest*. R News, 2(3):18-22
[25] Park M.Y. and Hastie T. (2007) *An L1 regularization-path algorithm for generalized linear models*. J. Roy. Statist. Soc. Ser. B, 69:659-677
[26] Poggi J.M. and Tuleau C. (2006) *Classification supervisée en grande dimension. Application à l'agrément de conduite automobile*. Revue de Statistique Appliquée, LIV(4):39-58
[27] Rakotomamonjy A. (2003) *Variable selection using SVM-based criteria*. Journal of Machine Learning Research, 3:1357-1370
[28] Strobl C., Boulesteix A.-L., Zeileis A. and Hothorn T. (2007) *Bias in random forest variable importance measures: illustrations, sources and a solution*. BMC Bioinformatics, 8:25
[29] Strobl C., Boulesteix A.-L., Kneib T., Augustin T. and Zeileis A. (2008) *Conditional variable importance for Random Forests*. BMC Bioinformatics, 9:307