Recursive partitioning methods for identifying relevant variables in a sparse unbalanced data set

Orlane Briand[1], Ivanne Debec[1], Arnaud Montet[2], Lorraine de Malleray[2] & Evelyne Vigneau[1]

[1] *Sensometrics and Chemometrics Laboratory, ONIRIS, La Géraudière, Nantes*
E-Mail : *orlane.briand@oniris-nantes.fr, ivanne.debec@oniris-nantes.fr, evelyne.vigneau@oniris-nantes.fr*

[2] *IFF, Neuilly-sur-Seine & New York*
E-Mail : *arnaud.montet@iff.com, lorraine.de.malleray@iff.com*

Abstract

The aim of this article is to explore various recursive partitioning methods when dealing with a sparse and unbalanced data set. Random forests based on the Classification and Regression Trees (CART) are compared to alternative algorithms based on Conditional Inference trees (CI trees). The importance measures available in these two frameworks are discussed in order to determine relevant variables.

Keywords: Recursive partitioning methods, CART, CI trees, random forests, importance measures

Résumé

Le but de cet article est d'explorer différentes méthodes de partitionnement récursif lorsque le jeu de données est creux et non équilibré (contenant beaucoup de valeurs nulles). Les forêts aléatoires basées sur les arbres de classification et de régression (CART) sont comparées à des algorithmes alternatifs basés sur les arbres d'inférence conditionnelle (CI trees). Les mesures d'importance proposées par ces deux alternatives dans le but de déterminer les variables pertinentes sont discutées.

Mots-clés : Méthodes de partitionnement récursif, CART, CI trees, forêts aléatoires, mesures d'importance

1. Introduction

Recursive partitioning methods, among which Classification and Regression Trees (CART) (see Breiman *et al*., 1984), Conditional Inference trees (CI trees) (see Hothorn *et al*., 2006) and Random Forests (RF) (see Breiman, 2001), are emerging methods in statistical data analysis interesting when dealing with nonlinearities and interaction effects. Besides their predictive ability, these methods are easy to interpret, providing an efficient way to identify relevant variables.

They have been investigated in a case study in which a quantitative quality response, $\boldsymbol{y}$, was to be related to a large number of quantitative predictors (in the $\mathbf{X}$ matrix), most of them being sparse (with 0 values). Moreover, about one third of these predictors included only one non-null observation, giving rise to a very unbalanced cover of the observations' space.

The aim of this study was to explore several approaches based on recursive binary trees, combined with random forests, and more specifically CART and CI trees. A comparison of the subsets of variables highlighted, depending on the criteria used for assessing the variables' importance, was also investigated.

2. Methods

2.1. CART and Random Forests

The CART algorithm proposed by Breiman (1984) is a recursive binary splitting algorithm in which, at each node, the set of the observations reaching this node is partitioned into two subsets in order to maximize a measure related to the variation of the node impurity. An exhaustive search is performed among all the variables and all the possible values for splitting.
For regression models, the impurity measure *i(t)*, at the node *t*, is given by the variance of the response $\boldsymbol{y}$. The objective is to determine the threshold *s*, for a predictor $\boldsymbol{X_j}$, for which the partition of the N_t observations reaching the node *t* into child nodes t_L and t_R maximizes

$$\Delta \boldsymbol{i}(\boldsymbol{t}) = \boldsymbol{i}(\boldsymbol{t}) - \boldsymbol{p_L}\boldsymbol{i}(\boldsymbol{t_L}) - \boldsymbol{p_R}\boldsymbol{i}(\boldsymbol{t_R})$$

where $p_L=N_{tR}/N_t$ and $p_R= N_{tR}/N_t$.

Let us remark that a tree is intrinsically grown to fit a specific learning sample. So, small changes in the learning set can have an impact on the structure of the tree. As a solution, random forests have been proposed in order to average over an ensemble of trees. Random forests as introduced by Breiman (2001) is random in two ways: *(i)* each tree is based on a bootstrap sample of the initial observations set, *(ii)* only a random subset of the predictor variables is considered as candidates at each node in a tree.

However, random forests based on the CART algorithm have been shown to be biased for variables selection. In particular, CART favors splits in continuous variables and variables with numerous categories (e.g. Strobl *et al.*, 2007). In the null hypothesis (independence between **y** and **X**), variables with more possible cut-off points are more likely to produce the best Δi(t). In our case study, this is to be considered because, even if the variables $\boldsymbol{X}$ are all quantitative, the more they are sparse, the less there are choices in the cut-off points. In the most extreme case, when only one observation has a non-null value, there is only one possibility to split the variable.

2.2. CI trees and Random Forests

In order to overcome the bias selection problem, several unbiased tree algorithms have been suggested, among which the Conditional Inference trees algorithm proposed by Hothorn *et al.* (2006). Basically, at each node, the association of each predictor to the response is assessed by a permutation test framework. In a first step, each predictor is tested and if none of them reached the predefined significance level, the actual node is not further split. Otherwise, the predictor showing the strongest relationship to the response (lowest *p-value*) is chosen. In a second step, the selected predictor is used in order to split the set of observations into two subsets. Due to the fact that, at each node, variable selection and splitting procedure are separated, Hothorn *et al.* (2006) claim that the obtained tree structures do not suffer from a systematic tendency towards covariates with many possible splits or many missing values.

Likewise CART, CI trees can be involved in random forests following the same rationale as Breiman's original approach.

2.3. Variable importance measures

In random forests, based on either CART algorithm or CI trees, different criteria are proposed to assess the importance of the variables and afterwards their ranking.

In random forest implementations based on CART (using the "RandomForest" R package), two different criteria are used: the mean decrease in accuracy (denoted "MDA-CART" in the following) and the mean decrease in node impurity (denoted "MDI-CART"):

- "MDA-CART" is used to measure the impact of each variable on the accuracy of the model. It is determined by permuting the values of each variable and measuring how much the permutation decreases the accuracy of the model, as suggested by Breiman (2001). In this way, the permutation has little to no effect on model accuracy if variables are not important, while permuting important variables significantly decreases it.
- "MDI-CART" is a measure of how each variable contributes to the homogeneity of the nodes in the resulting random forest. It corresponds to the average, over all trees in the forest, of the (weighted) mean of the improvement in the splitting criterion, $\Delta \boldsymbol{i}$, due to a given variable.

In the conditional inference framework ("party" R package), only the MDA criterion could be evaluated. However, Hapfelmeier *et al.* (2014) proposed a modification of the usual permutation methodology (Breiman, 2001) especially applicable for data with missing values. Two slightly similar criteria are then available (denoted "MDA-CI-perm" and "MDA-CI-rdalloc" in the following):

- "MDA-CI-perm" criterion evaluated following the permutation principle of the MDA importance in 'RandomForest' but based on CI trees, instead of CART trees.
- "MDA-CI-rdalloc" criterion based on the random re-allocation of each observation to the child nodes in the node that uses the variable considered for splitting.

3. Results

Results show that the criteria of variable importance allow variable selection and ranking. Figures 1 to 4 illustrate, for each type of measure, the ranking of the first forty variables according to their importance. The standard deviation over all trees of a forest has been computed for each variable. The variability of the criterion values is represented by an error bar (± 2 sd) in the figures. Regarding the MDA criteria, when the lower bound of the error bar is below zero, the significance of the observed value is doubtful. Figures 1-4 are used as a basis for selecting a subset of variables.

The summary in Table 1 of the selected variables, according to the criterion considered, points out that the lists are almost the same. For instance, in any case, variables 302 and 181 are the first two ones. Then, variables 493, 290, 63, 68, 170 are selected with each type of measure importance with a high average ranking.

As previously mentioned, one third of the variables contains only one non-null value in the data set. It can be shown that MDA-CART and MDA-CI-perm criteria will always be equal to zero for these variables. The use of "MDA-CI-rdalloc" criterion allows such variables to have a significant measure of importance which can be interesting depending on the case study and type of variables. Variables 46* and 4*, listed in Table 1 for the MDA-CI-rdalloc criterion, are two of this type of variables.

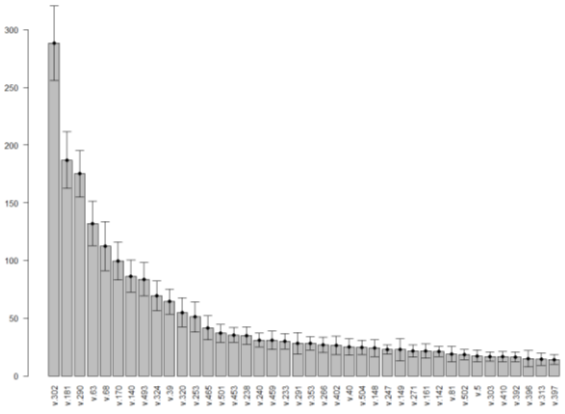

Figure 1: Order of importance / MDI

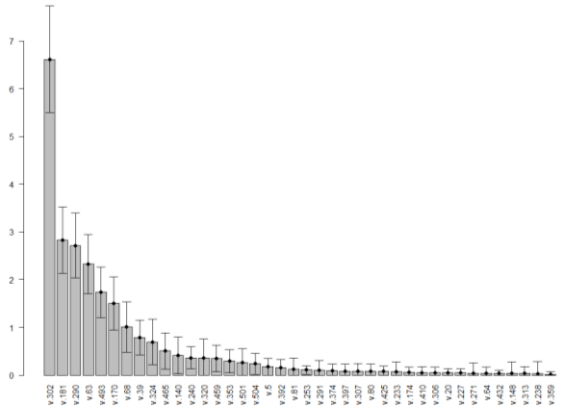

Figure 2: Order of importance / MDA-CART

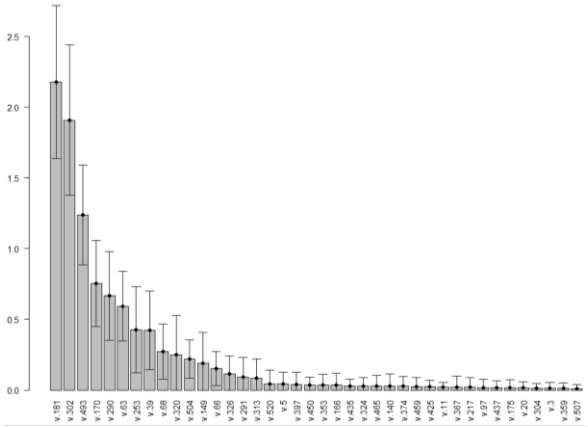

Figure 3: Order of importance / MDA-CI-perm

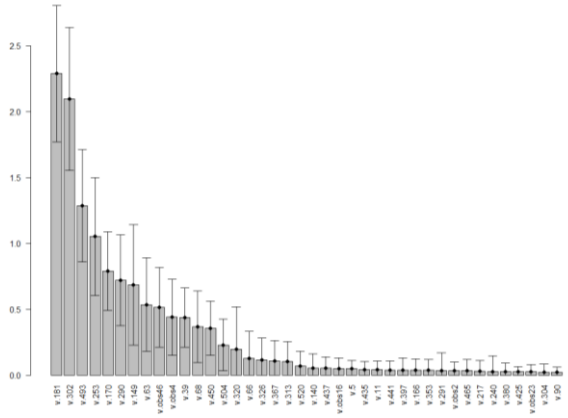

Figure 4: Order of importance / MDA-CI-rdalloc

MDI	MDA-CART	MDA-CI-perm	MDA-CI-rdalloc
302	302	181	181
181	181	302	302
290	290	493	493
63	63	170	253
68	493	290	170
170	170	63	290
140	68	253	149
493	39	39	63
324	324	68	46*
39	465	320	4*
320		504	39
253		149	68
465		66	450
			504

Table 1 : Summary table of importance measures and ranking

4. Discussion and conclusion

Regression trees and random forests are popular in many applications but they present some drawbacks such as bias in variable selection. In this paper, two frameworks of procedures were compared for variable selection purpose in the context of a sparse and unbalanced data set. Classification and Regression Trees (CART) and Conditional Inference trees (CI trees) algorithms allowed to rank and select important variables **X,** that have an effect on the quantitative response ***y***. According to the procedure, the criterion used for the construction of trees and for assessing the importance of a variable may differ. The Conditional Inference trees introduced by Hothorn *et al.* (2006) seem to be an appropriate solution to this kind of regression problems. In the Conditional Inference trees framework, the "MDA-CI-rdalloc" measure provides an unbiased variable selection and allow variables with only one non-null value to be pointed out as potentially relevant.

References

Breiman, L., Friedman, J. H., Olshen R.A. & Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

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

Hapfelmeier, A., Hothorn, T., Ulm, K. & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Stat Comput*, 24:21-34.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased Recursive Partitioning: A conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. (2007). Biais in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 1-21.