Variable importance and selection using random forests

Abderezzak Amimer , Madi Kassymbekov

17/04/2021

1. Presentation of the subject

This project is interested in prpoposing a tutorial for variable importance and variable selection in random forests using different R ressources. The methods used in the following sections are purely designed for prediction setting and not an inference one. Before diving into the tutorial, each element discussed are briefly reviewed.

1.3 Random Forests Random forests are in the family of ensemble methods. The principle is fairly simple. It combines many regression trees or classification trees using several bootstrap samples from the training data and randomly selecting a subset of the explanatory variables at each node in order to split the data. Amongst the random subset, one is selected using CART or Inference method to split. Additional details will be provided on this matter and these very powerful algorithms.

1.1 Variable importance Variable importance can be defined as the contribution of each predictor to the model. it is usually represented as a ranking of the variables based on the effect they have on the generated model. Knowing which variables significantly impacts the quality of the predictions helps analysts weed out those that are not necessary and help , in certain cases, improve the quality of the predictions.

1.2 Variable selection Variable selection consists of removing certain predictors from a model in order to improve the quality of a prediction. Following Kohavi and John (1997) and Guyon and Elisseff (2003), three groups of methods are distinguishable. First, "wrapper", which look closely at the prediction performance of a model to make variable selection. Second, "filter", the score of variable importance is not based on a model design and finally, "embedded" which combine both model estimation and variable selection.

Outline This project is organized as follows. A description of important literature and the most recent algorithms used in R regarding variable importance and selection using trees. Next, an overview of the methods to be used in the tutorials. To continue, a section will be deticated to review R references that will be used on a simulated dataset.This last part will be a demonstration of how to perform variable selection and variable importance using the references presented in the appropriare sections.

2. Litterature review

2.1 Random forest Random forests are based on decision trees with bootstrap aggregation called "bagging". It is a method used to generate multiple versions of a predicor to get an average predictor. Randomness is introduced in each predictor by making bootstrap replicates. As low bias is the feature of tree based models, high variance is one of its drawbacks dependent on the number of features that leads to larger trees and potential vulnerability to overfitting. However, compared to other tree models like CART and conditional inference tree, random forest models use bootstrap aggregation technique with random features to do prediction and afterwards uses weighted averaging which lowers the variance of predictions compared to classical decision tree models. The algorithm could be summarized by the following pseudo code Breiman (2001) :

1) For a training set of length N, sample N instances at random with replacement.
2) Grow a tree on the bootstrap training set using a specified number of random features
3) Repeat step 1 and 2 for the set number of estimators.
4) Average predictions or take majority vote (depending on type of task).

2.2 Variable importance

Sensitivity to n and p All learners are sensitive to the nature of the data and it's quality. Two important components to consider when fiting a model is the number of observations n and the number of explanatory variables p. Genuer et al. (2010) ran an experiment using simulated data to understand how the variable importance index would vary when $n << p$. It was shown that as n decreases and becomes smaller compared to p, the variable importance index of significant variables would decrease and get closer to 0. The variablity of the index also increases greatly and that is due to the intrsinc algorithm of random forests.

Sensitivity to mtry and ntree Mtry hyperparameter in the random forest algorithm is defined as the subset of covariates in the data that is used to perform the best split at different nodes. According to Genuer et al. (2010), the increase of the number of variables in the subset leads to an increase in the magnitude of variable importance index of useful predictors. On the other hand, ntree, an other hyperparameter of the random forest algorithm, is defined as the number of trees. It's increase leads to a decrease of the standard deviantion of the variable importance index. It is important to note, that although the standard deviation decreases, using a large number of trees can prompt overfitting on the training data.

Sensitiity to correlated covariates Many studies were performed to evaluate the impact of having correlated covariates on the variable importance index. Archer and Kimes (2008) paper indicates that the increase in the number of correlated covariates makes it difficult to pick significant variables due to the dilution of the variable importance score which is confirmed by the experiments of Auret and Aldrich (2011).

2.3 Variable selection In this project, the focus will be put on the "wrapper" methods which base the variable selection on a score that includes the prediction performance. Usually, the performance measure used is the mean squared error for regression tasks and the misclassification rate for classification tasks.

3. Brief review of methods

In the following section, a brief description of the methods used to calculate variable importance and to perform relevant variable selection will be presented.

3.1 Variable importance calculation methods First we start with variable importance calculation as it is the main index used to perform variable selection using random forests.

The idea behind this measure is that a covariate X is important if the prediction error increases whent the relationship between X and Y is modified. To illustrate how to calculate the variable importance once can simply follow these steps for the classical approach which is presented by Breiman (2001)

1) Fit a random forest on the training data.
2) Select an out of bag sample of observations that the model was not trained on (OOB) and calculate the mean squared error on each tree.
3) Select one covariate in the covariate space and permute the values of each instances in order to break the link between the response variable and the covariate selected.
4) Make new predictions using these permuted values and obtain a new mean squared error for OOB sample.
5) Perform the steps 3 and 4 for each covariate in the covariate space.
6) To calculate the variable importance, the sum of differences of the MSE between the permuted covariate and the non permuted covariate is taken and divided by the total number of OOB.

The following formula displays the results of this algoritm :

$$VIMP(X_j) = \frac{1}{B}\sum_{b=1}^{B}[MSE(\hat{f}_b, OOB_b^j) - MSE(\hat{f}_b, OOB_b)]$$

where j is the covariate for which the value are permuted. b is the bth OOB sample. B is the number of out of bag samples which is equal to the number of trees. $\hat{f}_b$ is the prediction model for the bth sample.

Isharwan H. (2007) proposed a different way of breaking the link between Y and X. Instead of permuting the values of X_j they proposed to assign the node at random instead of chosing the best variable to split on. The second option, consists of systematically choosing the opposite split each time a split is performed with that variable. This adds noise to the model and the error is then used to calculate the difference in error with the original model.

3.2 Variable selection methods Variable selection in random forest is mainly based on the variable importance index. The following section explores and briefly describes the different ways to perform variable selection in order to get a better prediction performance.

Recursive feature elimination (RFE) and non recursive feature elimintation (NRFE) RFE's objective is to find the smallest number of variables that help get the best predictive model. It follows the following steps :

1. Train a random forests
2. Compute the permutation importance measure
3. Eliminate the less relevant variable(s).
4. Repeat steps 1 to 3 until no further variables remain.

This method is an improved version of the NRFE, which was proven to be less efficient on correlated data. The update of the ranking based on variable importance has proven to be more effective in presence of correlation withing the covariates. For that reason, NRFE will not be used in the following tutorials. Gregorutti et al (2016)

Boruta This method was developped to find all relevant variables within a classification framework. Kursa and Rudnicki (2010) describe the Boruta algorithm with the following steps:

The Boruta algorithm consists of following steps: 1. Extend the training data by adding copies of all variables (the data is always extended by at least 5 duplicate attributes, even if the number of attributes in the original set is lower than 5).

2. Shuffle the added attributes to remove their correlations with the response.The goal here is to break the link between Y and X
3. Run a random forest classifier on the extended dataset and gather the Z scores computed. The Z score is calculted based on the loss of accuracy in classification task.
4. Find the maximum Z score among duplicate attributes, and then assign a hit to every attribute that scored better than the maximum Z score.
5. For each attribute with undetermined importance perform a two-sided test of equality with the maxium z score
6. Deem the attributes which have importance significantly lower than the maximum z score as 'unimportant' and permanently remove them from the dataset.
7. Deem the attributes which have importance significantly higher than the maximum z score as 'important'.
8. Remove all duplicate attributes.
9. Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

Vsurf The Vsurf method is both usefull in regression and classification tasks. It is also used for interpretation and prediction. The algorithm in question is performed in two different steps and it is described as follows. Genuer et al. (2015)

Step 1 : Preliminary ranking and elimination.

a) This step consists of ranking the variables using the classical variable importance index. Genuer et al.(2015) suggests using a typical 50 trees to estimate the variable importance index.

b) Based on a certain threshold, eliminate all variables for which the variable importance index is below the threshold. The threshhold in question is estimated by calculating the standard deviation of variable importance. Other strategies could be used to find the best threshold.

Step 2 : variable selection.

a) For interpretation. The algorithm uses a very simple method to select the most performing model. For $k = 1$ to m, a Random Forest model is built using m important variables following the ranking previously estimated where 1 is the most important variable. Using a different number of runs, typically 25, select the variables that lead to the smallest OOB error.

b) For prediction. In order to improve the prediction performance, start with the ordered sequence of variables selected in the interpretation step. From there, build different random forests by following the order of variable importance from the previous step and use a stepwise algorithm to eliminate variables that increase the OOB prediction error by a previously selected threshhold.

Jefferie S. Evans et al. method for model selection (Parsimony) The purpose of the authors of the following method was to simplify as much a s possible the complexity of the models. They used the permuted variable importance measure to do so. The method starts by fiting a random forest with all variables and estimates variable importance. The values are then ranked and standardized to a certain ratio. Then, it iteratively subsets variables within a given ratio and fitting a model for each subset. The resulting model is compared to the original model which is kept constant. The performance is calculated on the OOB error which can include a penalty for the number of parameters in order to reduce the complexity of the model.

Altmann Altmann's method is based on an approach that keeps the values and the correlation structure of the covariates intact while computing variable importance under a null hypothesis of no association between predictor and the response variable. Instead of permuting the values of X_j, it permutes the values of the outcome and then multiple random forests are trained and new variable importance are calculated. If the variable importance in the orginal model without permutation is significantly different than 0, then the variable is kept. Otherwise, it is deemed as non important and can be disgcarded from the model. This initial approach is called the permutation approach and is a basic method. Altmann's modifications are implemented with the distribution of the variable importance under the null hypothesis. In other words, in order to allow less trainings of random forests and to get a conclusive test, it is important to make the right assumptions on the variable importance distribution and the author of the described method allows a parametric approach to estimate the P-values by fitting a specific probability distributin such as a normal, a lognormal etc.

Recurrent relative variable importance This approach is used in case the number of unimportant variables in a dataset is large. The selection method is simply based on a ratio of different random forest fits. In other words, first, several random forests are generated based on the data set and parameter values differing only in the seed of the random number generating process. Each random forest is used to compute the variable importance using the classical method and the values obtained are divided by the absolute minimal importance observed in each run. This outputs relative values and each variable having an importance greater or equal to a predetermined factor is kept.

Vita approach