Variable selection using random forests

Pattern Recognition Letters 31 (2010)

Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot

January 25, 2012

Outline

Motivations I

- Random forest for Variable selection.
- Methodology:
  - Provide some experimental insights about the behavior of the variable importance index
  - Propose a two-steps algorithm for two classical problems of variable selection.

Random Forests I

- **Principle:** 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.
- **Facts:**
  - at each node, a given number (*mtry*) of input variables are randomly chosen and the best split is calculated only within this subset.
  - no pruning step is performed, all the trees of the forest are maximal trees.

Random Forests I

- They focus on *randomForest* procedure of R package:
  - **2** parameters: ***mtry***, the number of input variables randomly chosen at each split and ***ntree***, the number of trees.
- They use the out-of-bag (oob) error estimation.

Random Forests I

- The algorithm:
  - Bootstrap sample of data.
  - Using 2/3 of the sample, fit a tree to its greatest depth determining the split at each node through minimizing the loss function considering a random sample of covariates (size is user specified)
  - For each tree. .
    - Predict classification of the leftover 1/3 using the tree, and calculate the misclassification rate = out of bag error rate.
    - For each variable in the tree, permute the variables values and compute the out-of-bag error, compare to the original oob error, the increase is a indication of the variable's importance
  - Aggregate oob error and importance measures from all trees to determine overall oob error rate and Variable Importance measure.

Random Forests II

- Oob Error Rate: Calculate the overall percentage of misclassification.
- Variable Importance: Average increase in oob error over all trees and assuming a normal distribution of the increase among the trees, determine an associated p-value.

Outline

Variable importance I

- **Variable importance index:** the increasing in mean of the error of a tree (mean square error (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.

Variable importance I

RF variable importance:

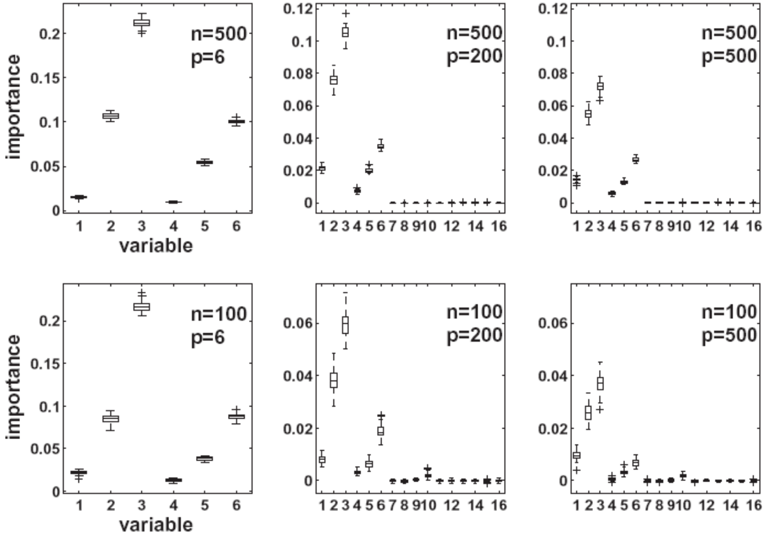
- For each tree t :
  - Consider the associated OOB_t sample.
  - Denote by $errOOB_t$ the error of a single tree t on this OOB_t sample.
  - Randomly permute the values of X^j in OOB_t to get a perturbed sample denoted by OOB_t^j and compute $errOOB_t^j$, the error of predictor t on the perturbed sample.

$$VI(X^j) = \frac{1}{ntree} \sum_t (err\widetilde{OOB}_t^j - errOOB_t),$$

Sensitivity to n and p I

- *ntree*=500 and *mtry*=$\sqrt{p}$
- Boxplots: 50 runs of RF algorithm. Plot only few variables. Graphs with n= 500 (top), n=100 (bottom)

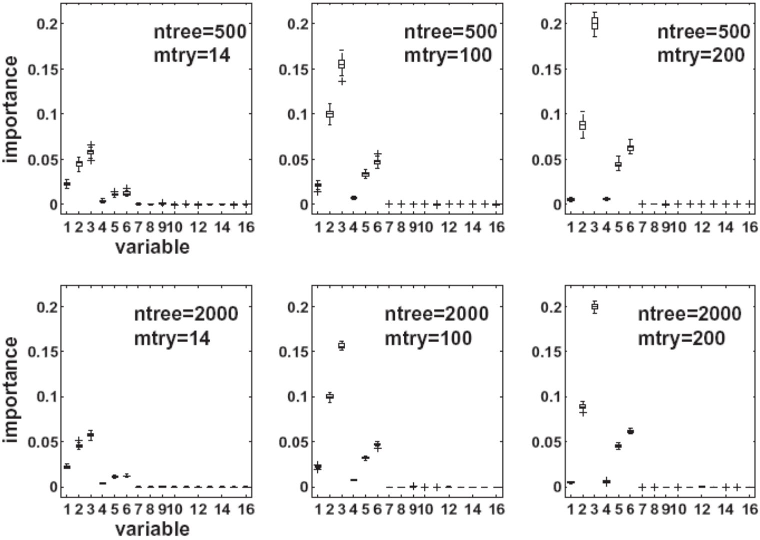
Sensitivity to n and p II



Sensitivity to mtry and ntree I

- We fix n=100 and p=200.

Sensitivity to mtry and ntree II



Outline

Variable selection I

We distinguish two variable selection objectives:

1. To find important variables highly related to the response variable for interpretation purpose;
2. To find a small number of variables sufficient to a good parsimonious prediction of the response variable.

Procedure I

Two-steps Procedure:

Step 1. Preliminary elimination and ranking:

- Sort the variables in decreasing order of RF scores of importance.
- Cancel the variables of small importance. Denote by m 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;

Procedure II

- 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.

Starting example I

- Simulated learning set n=100 and p=200.
- Run 50 forest with ntree=2000 an mtry=100
- *Variable ranking*. First we rank the variables by sorting the VI (averaged from the 50 runs) in descending order.
- *Variable elimination.* We set the threshold as the minimum prediction value given by a CART model fitting this curve.
- *Variable selection procedure for interpretation.* We compute OOB error rates of random forests.
- *Variable selection procedure for prediction.* We perform a sequential variable introduction with testing: a variable is added only if the error gain exceeds a threshold.

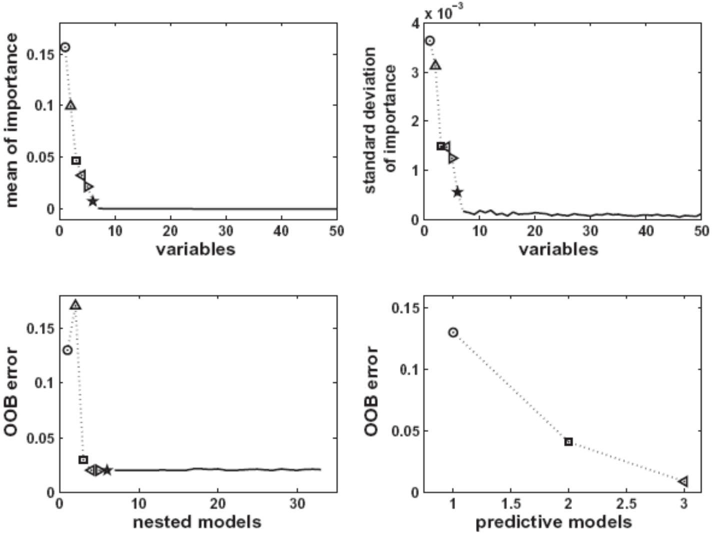
Starting example II

- The threshold is set to the mean of the absolute values of the first order differentiated OOB errors between the model with $p_{interp} = 4$ variables (the model we selected for interpretation, see the bottom left graph) and the one with all the $p_{elim} = 33$ variables:

$$\frac{1}{p_{elim} - p_{interp}} \sum_{j=p_{interp}}^{p_{elim}-1} |errOOB(j+1) - errOOB(j)|,$$

Starting example I

Starting example II

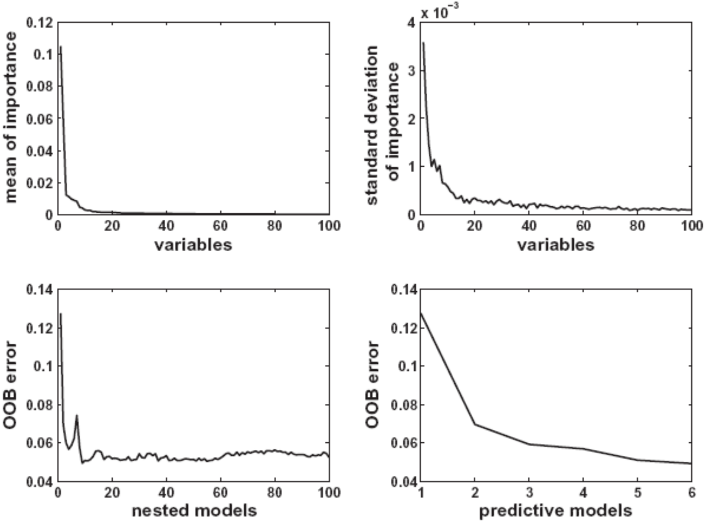


Outline

Prostate data I

- Prostate data: $n = 102$ and p =6033
- We use $ntree = 2000$; $mtry = p/3$

Prostate data II



Four high dimensional classification datasets I

Four well known high dimensional real datasets:

- Colon (n=62; p= 2000),
- Leukemia (n=38; p= 3051),
- Lymphoma (n = 62; p = 4026) and
- Prostate (n = 102; p = 6033).

To estimate the error rate we use a 5-fold cross-validation.

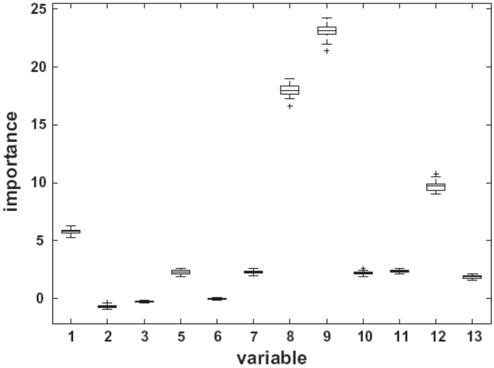
Table 2
Variable selection procedure for four high dimensional real datasets. CV-error rate and into brackets the average number of selected variables.

Dataset	Interpretation	Prediction	Original
Colon	0.16 (35)	0.20 (8)	0.14
Leukemia	0 (1)	0 (1)	0.02
Lymphoma	0.08 (77)	0.09 (12)	0.10
Prostate	0.085 (33)	0.075 (8)	0.07

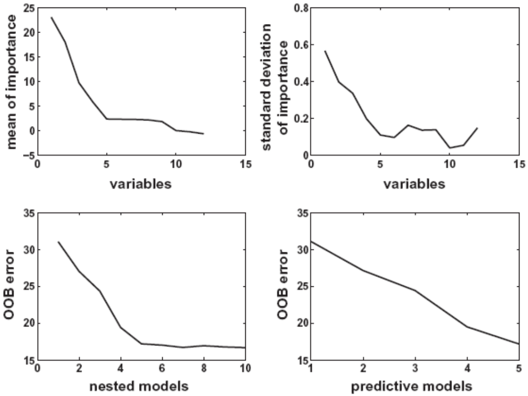
Ozone data I

- A standard regression dataset.
- We apply the entire procedure to the easy to interpret ozone dataset
- $n = 366$ observations of the daily maximum one-hour-average ozone together with $p = 12$ meteorologic explanatory variables.
- RF procedure: $mtry = p/3 = 4$ and $ntree = 2000$.
- *12 explanatory variables*: 1- Month, 2-Day of month, 3-Day of week, 5-Pressure height, 6-Wind speed, 7-Humidity, 8-Temperature (Sandburg), 9-Temperature (El Monte), 10-Inversion base height, 11-Pressure gradient, 12-Inversion base temperature, 13-Visibility.

Ozone data I



Ozone data I


mean of importance
variables
standard deviation of importance
variables
OOB error
nested models
OOB error
predictive models