Random forests variable importances

Towards a better understanding and large-scale feature selection

Antonio Sutera

Dept. of EECS, University of Liège, Belgium

COMPSTAT 2016,
Oviedo, Spain
August 24, 2016

Pierre Geurts, Louis Wehenkel (ULg),
Gilles Louppe (CERN & NYU)
Célia Châtel (Luminy)

Université de Liège

Ensemble of randomized trees: strengths and weaknesses

✓ Good classification method with useful properties:

- Universal approximation
- Robustness to outliers
- Robustness to irrelevant variables (to some extent)
- Invariance to scaling of inputs
- Good computational efficiency and scalability
- Very good accuracy

✗ Loss of interpretability w.r.t. standard trees

Ensemble of randomized trees: strengths and weaknesses

✓ Good classification method with useful properties:

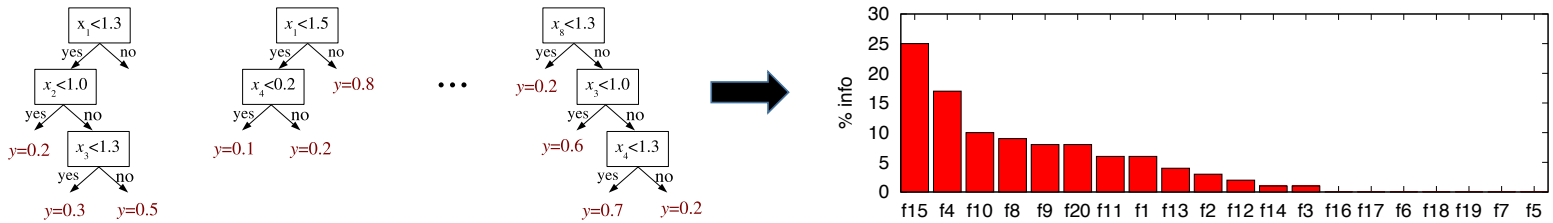
- Universal approximation
- Robustness to outliers
- Robustness to irrelevant variables (to some extent)
- Invariance to scaling of inputs
- Good computational efficiency and scalability
- Very good accuracy

✗ Loss of interpretability w.r.t. standard trees

$\Rightarrow$ but some interpretability can be retrieved through **variable importance scores**

Variable importance scores

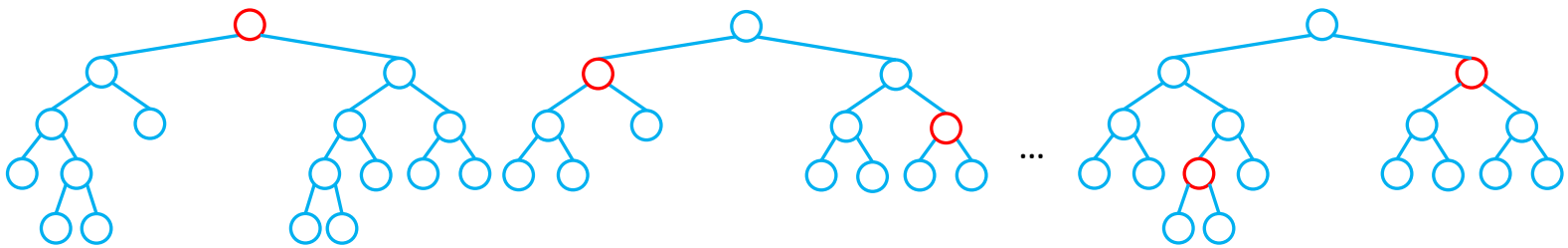
- Some interpretability can be retrieved through variable importance scores



- Two main importance measures:
  - **The mean decrease of impurity (MDI)**: summing total impurity reductions at all tree nodes where the variable appears (Breiman et al., 1984)
  - **The mean decrease of accuracy (MDA)**: measuring accuracy reduction on out-of-bag samples when the values of the variable are randomly permuted (Breiman, 2001)
- These measures have found many successful applications such as:
  - Biomarker discovery
  - Gene regulatory network inference

(Huynh-Thu et al, Plos ONE, 2010 and Marbach et al., Nature Methods, 2012)

Mean decrease of impurity (MDI): definition



Importance of variable X_m for an ensemble of N_T trees is given by:

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(t) = X_m} p(t) \Delta i(t)$$

where $p(t) = N_t / N$ and $\Delta i(t)$ is the impurity reduction at node t:

$$\Delta i(t) = i(t) - \frac{N_{t_L}}{N_t} i(t_L) - \frac{N_{t_r}}{N_t} i(t_R)$$

Motivation

Despite many successful applications in various domains, random forests variable importances are still poorly understood.

Our general objectives:

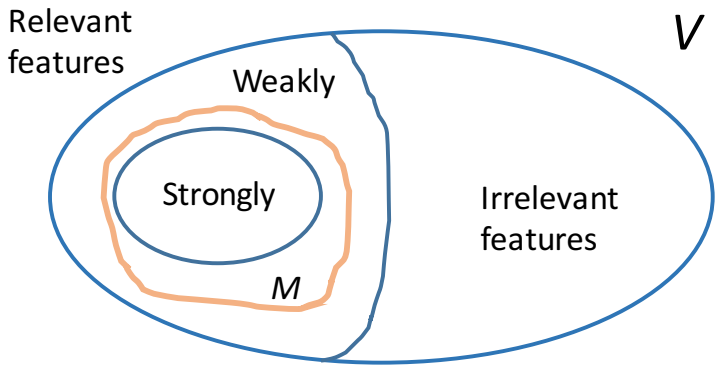
- Better understand the MDI importance measure, so as to provide advices on how to best interpret it and exploit it in practice
- Design more efficient feature selection procedures based on random forests.

Outline

1. Tree-based variable importance scores
2. Towards a better understanding of the MDI measure
3. Towards large-scale feature selection

Outline

Background: Feature relevance (Kohavi and John, 1997)

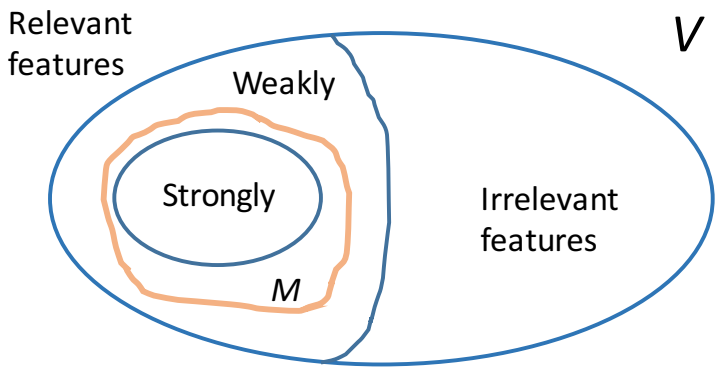


Given an output Y and a set of input variables V, $X \in V$ is

- **relevant** iff $\exists B \subseteq V$ such that $Y \not\perp\!\!\!\perp X|B$.
- **irrelevant** iff $\forall B \subseteq V$: $Y \perp\!\!\!\perp X|B$
- **strongly relevant** iff $Y \not\perp\!\!\!\perp X|V \setminus \{X\}$.
- **weakly relevant** iff X is relevant and not strongly relevant.

A **Markov boundary** is a minimal size subset $M \subseteq V$ such that $Y \perp\!\!\!\perp V \setminus M|M$.

Background: Feature selection (Nilsson et al., 2007)



Two different feature selection problems:

- **Minimal-optimal:** find a Markov boundary for the output Y.
- **All-relevant:** find all relevant features.

Notes:

- In general, both problems requires exhaustive subset search.
- When the input distribution is **strictly positive** ($f(x) > 0$), the markov boundary is unique and it contains all and only the strongly relevant features.

Assumptions

$$Imp(X_m) = \frac{1}{N_T} \sum_T \sum_{t \in T : v(t) = X_m} p(t) \Delta i(t)$$

Our working assumptions:

- All variables are discrete
- Multi-way splits à la C4.5, i.e. one branch per value of the variable
- Shannon entropy is used as the impurity measure:

$$i(t) = -\sum_c \frac{N_{t,c}}{N_t} \log \frac{N_{t,c}}{N_t}$$

- Asymptotic conditions: infinite sample size and number of trees

Two method parameters (with p the number of features):

- Number of features drawn at each node $K \in [1, p]$
- (Maximum tree depth $D \in [1, p]$)

Totally random unpruned trees

Thm. Variable importances provide **a three-level decomposition of the information jointly provided by all the input variables about the output**, accounting for all interaction terms in a **fair** and **exhaustive** way.

$$\underbrace{I(X_1, \dots, X_p; Y)}_{\text{Information jointly provided by all input variables about the output}} = \underbrace{\sum_{m=1}^{p} Imp(X_m)}_{\text{i) Decomposition in terms of the MDI importance of each input variable}}$$

$$Imp(X_m) = \underbrace{\sum_{k=0}^{p-1} \frac{1}{\binom{p}{k}(p-k)}}_{\text{ii) Decomposition along the degrees } k \text{ of interaction with the other variables}} \underbrace{\sum_{B \in \mathcal{P}_k(V^{-m})} I(X_m; Y|B)}_{\text{iii) Decomposition along all interaction terms } B \text{ of a given degree } k}$$

E.g.: $p = 3, \text{Imp}(X_1) = \frac{1}{3} I(X_1; Y) + \frac{1}{6}(I(X_1; Y|X_2) + I(X_1; Y|X_3)) + \frac{1}{3} I(X_1; Y|X_2, X_3)$

Link with common definitions of variable relevance

In asymptotic setting ($N = N_T = \infty$)

$K = 1$: Variable importances **depend only on the relevant variables**

- A variable X_m is relevant iff $Imp(X_m) > 0$
- The importance of a relevant variable is insensitive to the addition or the removal of irrelevant variables in V.

$\Rightarrow$ *Asymptotically, unpruned totally randomized trees thus solve the* **all-relevant** *feature selection problem*.

Link with common definitions of variable relevance

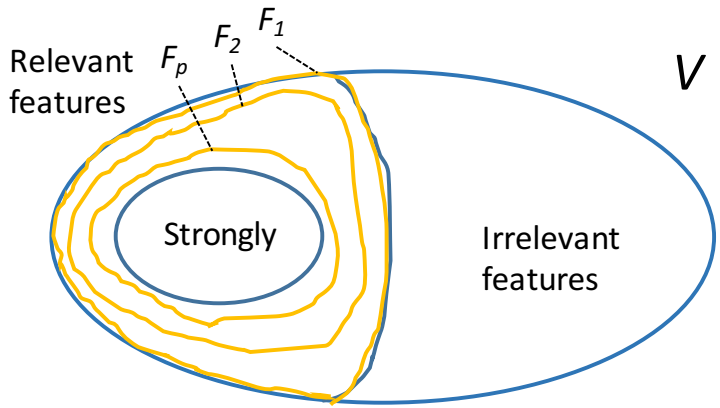
In asymptotic setting ($N = N_T = \infty$)

$K > 1$: Variable importances **can be influenced by the number of irrelevant variables** and **there can be relevant variables with zero importances** (due to masking effect)

But:

- X_m irrelevant $\Rightarrow Imp(X_m) = 0$
- X_m strongly relevant $\Rightarrow Imp(X_m) > 0$

Strongly relevant features can not be masked



$\Rightarrow$ *In the case of stricly positive distributions, non random trees always find a superset of the* **minimal-optimal** *solution which size decreases with K.*

Non asymptotic setting

- Finite number of trees

  In general, a **single tree** can not identify all relevant features, even the strongly relevant ones and an **ensemble of trees** is necessary.

  E.g.: $I(Y; X_1) = I(Y; X_2) = 0$ and $I(Y; X_1, X_2) > 0$

- Finite number of samples

  There is a positive bias in the estimation of mutual informations that depends on the cardinality of X and Y:

$$I(Y; X) = 0 \Rightarrow E\{\hat{I}(Y; X)\} = \frac{(|Y| - 1)(|X| - 1)}{2N_t \log 2}$$

Conclusions

Asymptotically, MDI is a sound statistic to detect weakly and strongly relevant features

As a quantitative score to rank relevant features, it should however be interpreted cautiously:

- Asymptotically, it is affected by the value of K, tree depth D, redundant and irrelevant variables (when $K > 1$).
- In finite settings, it is affected by biases in the estimation of impurity

To make the most of these scores, method parameters should be set appropriately and independently of predictive performance.

Future works:

- Finite sample analysis
- Numerical features
- Design alternative statistics with better or complementary properties.

Outline

- We want to address large-scale feature selection problems where one can not assume that all variables can be stored into memory
- Based on the previous analyses, we study and improve ensembles of trees grown from random subsets of features

(Work in progress)

Random subspace for feature selection

Simplistic memory constrained setting: We can not grow trees with more than q features

Straightforward ensemble solution: Random Subspace (RS)

Train each ensemble tree from a random subset of q features

1. Repeat T times:
   - 1.1 Let Q be a subset of q features randomly selected in V
   - 1.2 Grow a tree only using features in Q (with randomization K)
2. Compute importance $Imp_{q,T}(X)$ for all X

Proposed e.g. by (Ho, 1998) for accuracy improvement, by (Louppe and Geurts, 2012) for handling large datasets and by (Draminski et al., 2010, Konukoglu and Ganz, 2014) for feature selection

Let us study the population version of this algorithm.

RS for feature selection: study

Asymptotic guarantees:

- **Def.** $deg(X)$ with X relevant is the size of the smallest $B \subseteq V$ such that $Y \not\perp\!\!\!\perp X|B$
- $K = 1$: If $deg(X) < q$ for all relevant variables X: X is relevant iff $Imp_q(X) > 0$
- $K \geq 1$: If there are q or less relevant variables: X strongly relevant $\Rightarrow Imp_q(X) > 0$

Drawback: RS requires many trees to find high degree variables

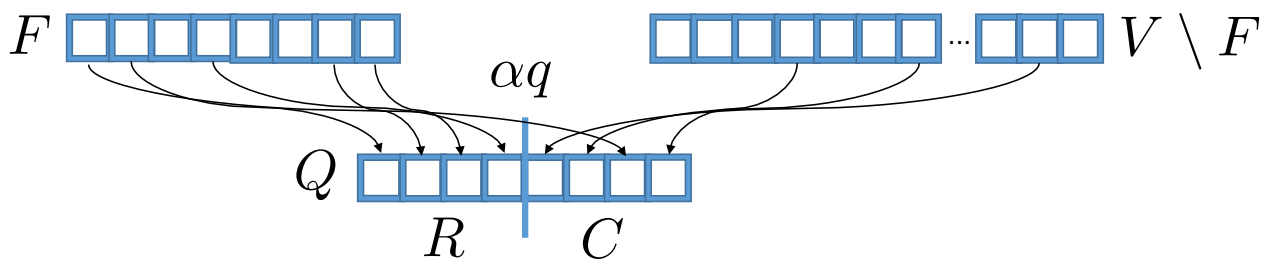
The probability to sample one feature X of degree $k < q$ together with its minimal conditioning is $\frac{\binom{p-k-\mathbf{1}}{q-k-\mathbf{1}}}{\binom{p}{q}}$

E.g.: $p = 10000, q = 50, k = 1 \Rightarrow \frac{\binom{p-k-\mathbf{1}}{q-k-\mathbf{1}}}{\binom{p}{q}} = 2.5 \cdot 10^{-5}$. In average, at least $T = 40812$ trees are required to find X.

Sequential Random Subspace (SRS)

Proposed algorithm:

1. Let $F = \emptyset$
2. Repeat T times:
   - 2.1 Let $Q = R \cup C$, where:
     - R is a subset of $\min\{\alpha q, |F|\}$ features randomly taken from F
     - C is a subset of $q - |R|$ features randomly selected in $V \setminus R$
   - 2.2 Grow a tree only using features in Q
   - 2.3 Add to F all features that get non-zero importance
3. Return F

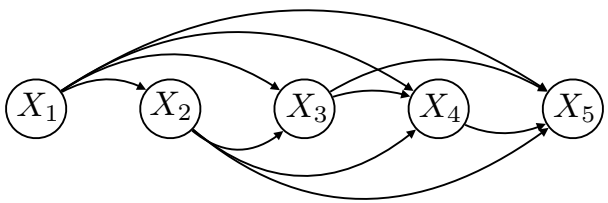


Compared to RS: *fill α% of the memory with previously found relevant variables and $(1 - \alpha)$% with randomly selected variables.*

SRS for feature selection: study

Asymptotic guarantees: similar as RS if all relevant variables can fit into memory.

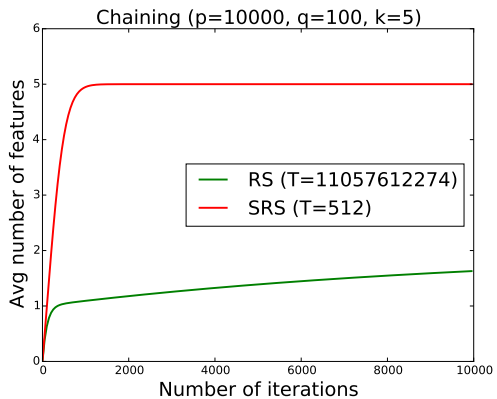
Convergence: SRS requires much less trees than RS in most cases.
For example,



Analytically

$$\overline{N}_{RS} \simeq (\frac{p}{q})^k$$
and
$$\overline{N}_{SRS} \simeq k\frac{p}{q}$$

Numerical simulation

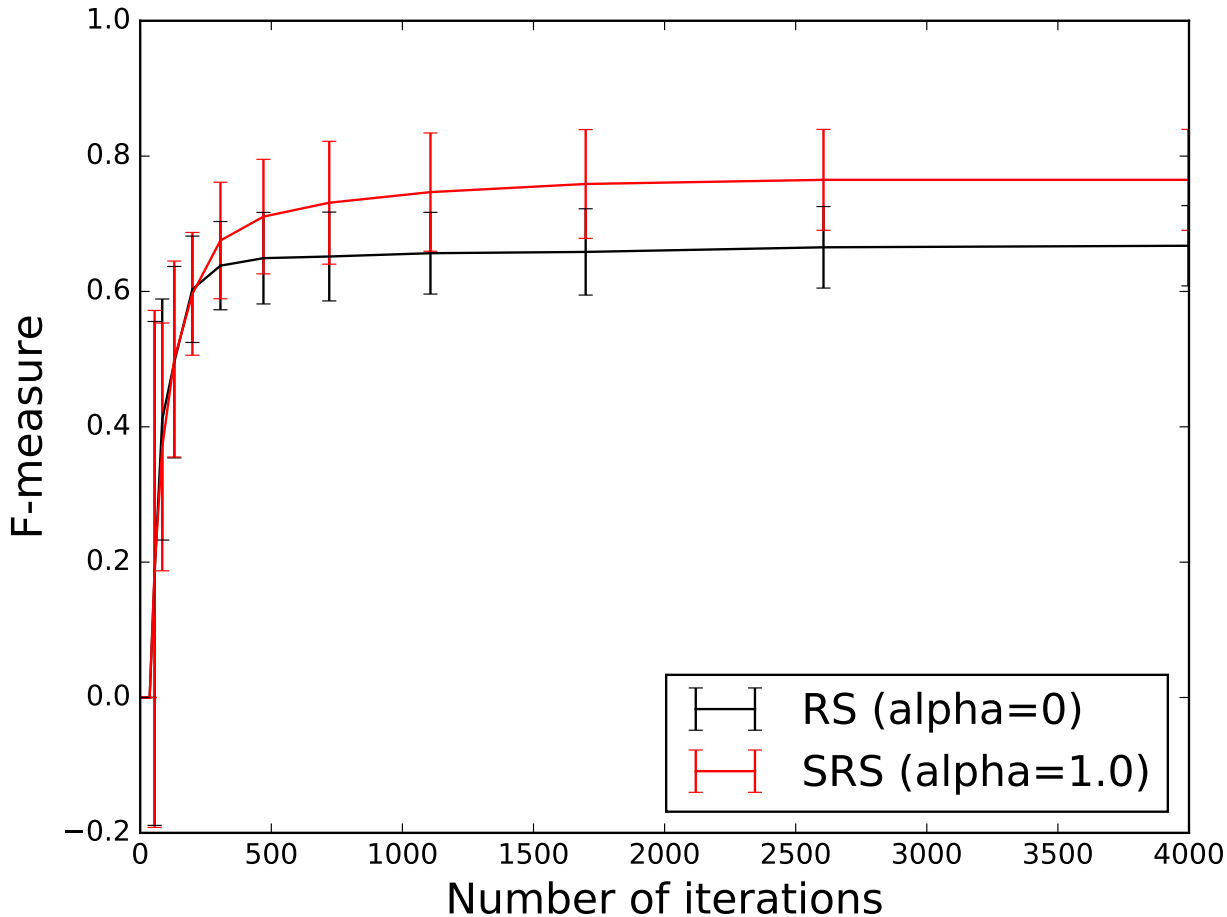


Note: $\alpha < 1$ ensures some permanent exploration of new features ($\alpha = 0 \Rightarrow$ RS).

Experiments: results in feature selection

Dataset: Madelon (Guyon et al., 2007)

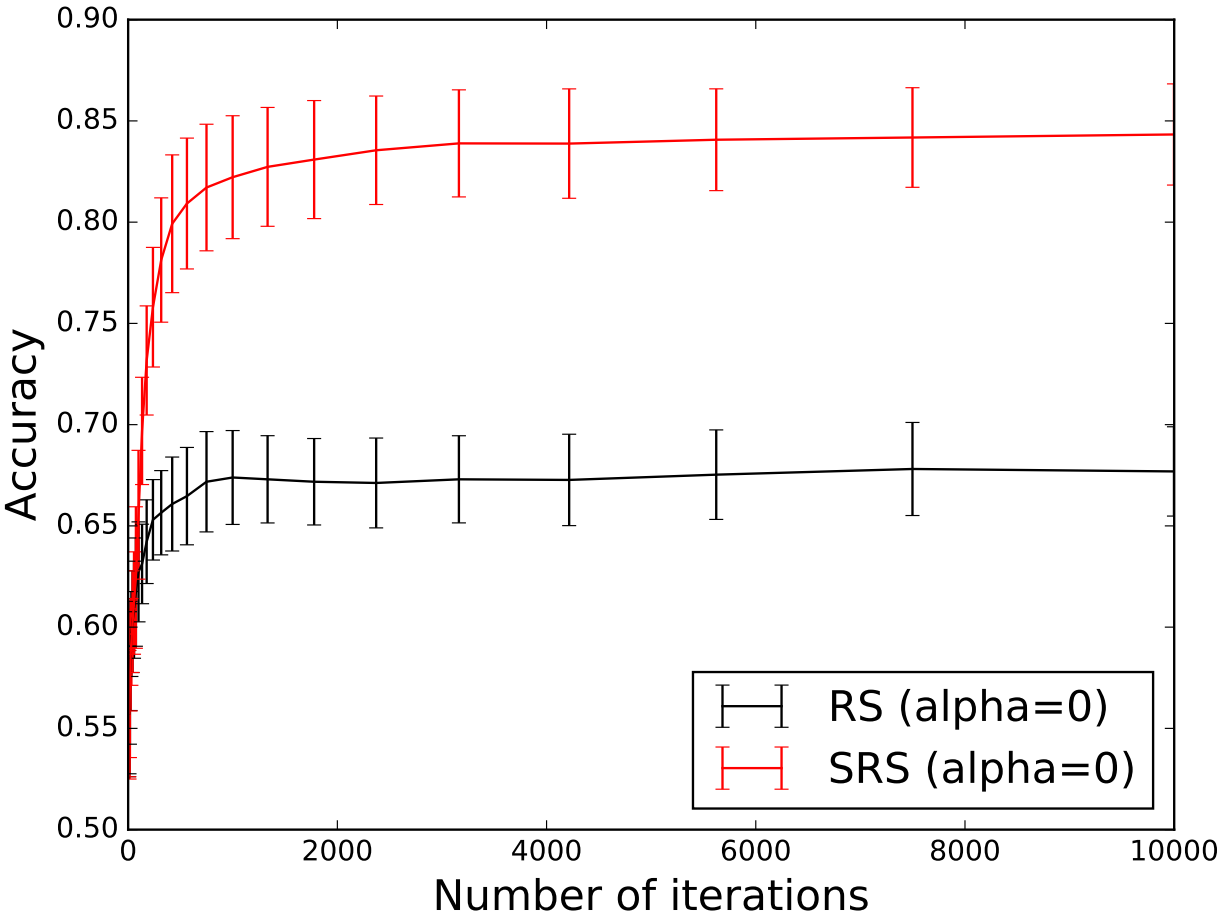
- 1500 samples (|LS|=1000, |TS|=500)
- 500 features whose 20 relevant features (5 features that define Y, 5 random linear combinations of the first 5, and 10 noisy copies of the first 10)



Parameter:

- q : 50

Experiments: results in prediction



Parameter:

- q : 50

After 10000 trees/iterations:

- RF ($K = max$): 0.81
- RF ($K = q$): 0.70
- RS : 0.68
- **SRS: 0.84**

Conclusions

Future works on SRS:

- Good performance of SRS are confirmed on other datasets but more experiments are needed.
- How to dynamically adapt K and α to improve correctness and convergence?
- Parallelization of each step or of the global procedure

General conclusion:

Interpreting random forests as a way to explore variable conditionings might shed new light on this algorithm and could suggest further improvements

References


Célia Châtel, *Sélection de variables à grande échelle à partir de forêtes aléatoires*, Master's thesis, École Centrale de Marseille/Université de Liège, 2015.


D. Marbach et al., *Wisdom of crowds for robust gene network inference*, Nature Methods **9** (2012), no. 8, 796–804.


V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, *Inferring regulatory networks from expression data using tree-based methods*, Plos ONE **5** (2010), no. 9, e12776.


V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts, *Statistical interpretation of machine learning-based feature importance scores for biomarker discovery*, Bioinformatics **28** (2012), no. 13, 1766–1774.


Gilles Louppe and Pierre Geurts, *Ensembles on random patches.*, ECML/PKDD (1) (Peter A. Flach, Tijl De Bie, and Nello Cristianini, eds.), Lecture Notes in Computer Science, vol. 7523, Springer, 2012, pp. 346–361.

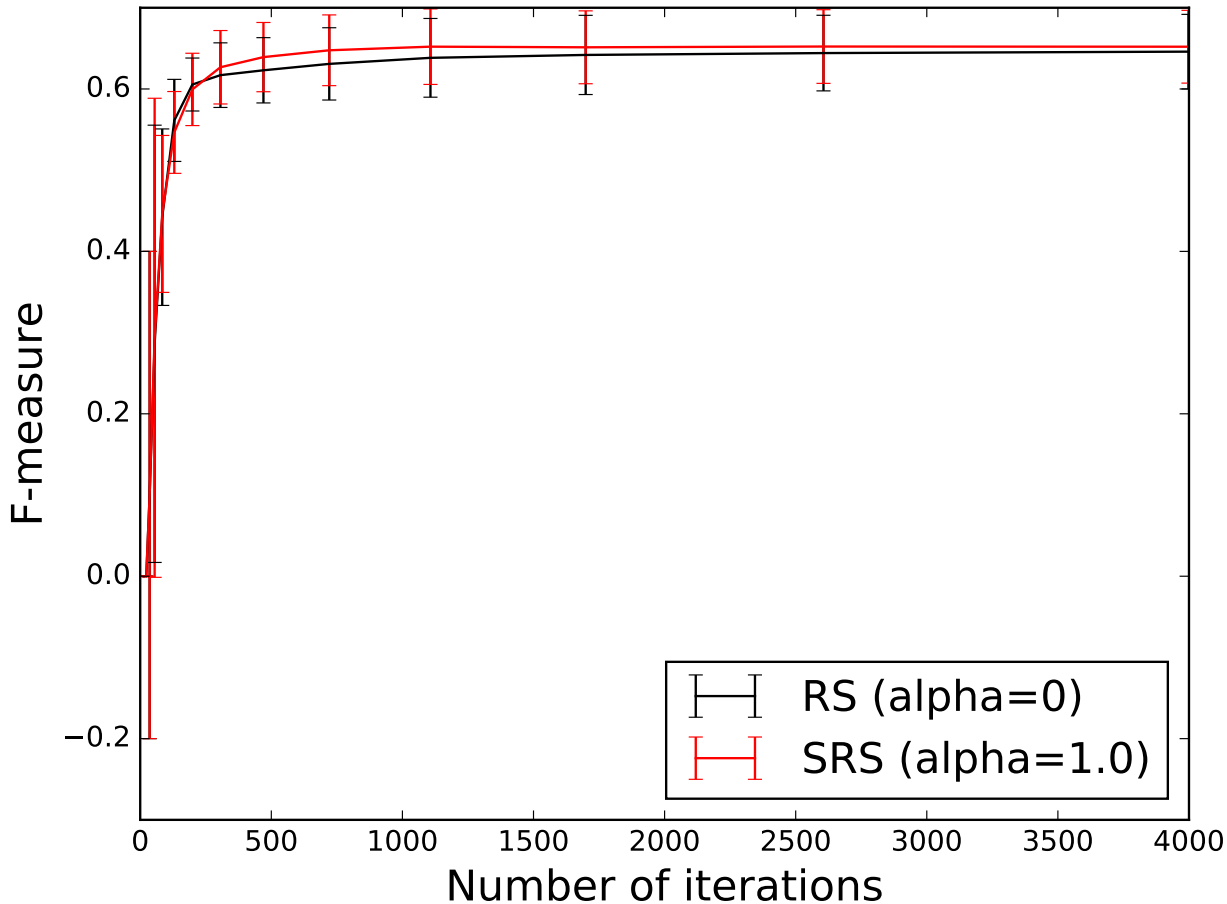

Gilles Louppe, *Understanding random forests: From theorey to practice*, Ph.D. thesis, University of Liège, 2014.


G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, *Understanding variable importances in forests of randomized trees*, Advances in neural information processing, 2013.

Experiments: results in feature selection

Dataset: TIS

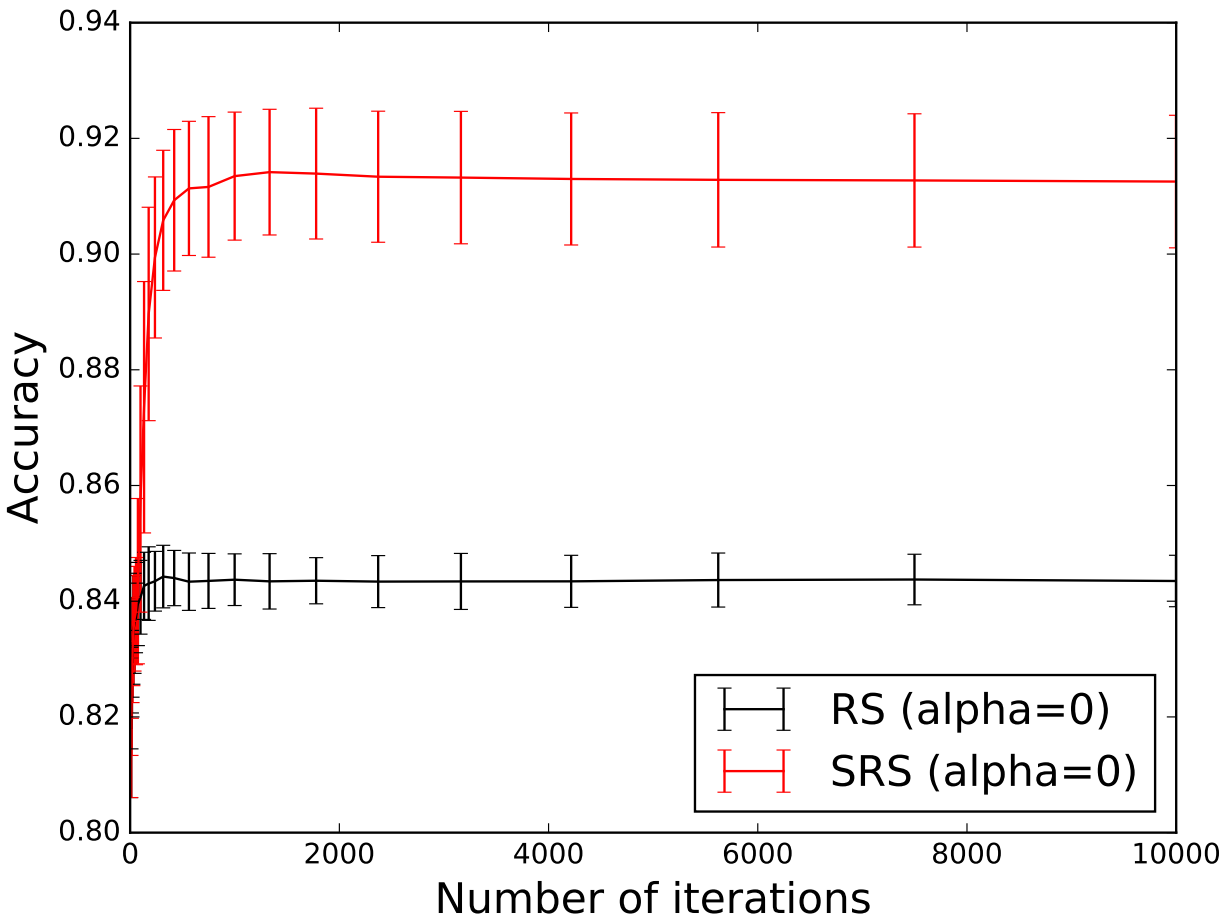
- 13375 samples
- 927 features



Parameter:

- q : 92

Experiments: results in prediction



Parameter:

- q : 92

After 10000 trees/iterations:

- RF ($K = max$): 0.91
- RF ($K = q$): 0.9
- RS : 0.84
- **SRS: 0.91**