Random forests variable importances

Towards a better understanding and large-scale feature selection

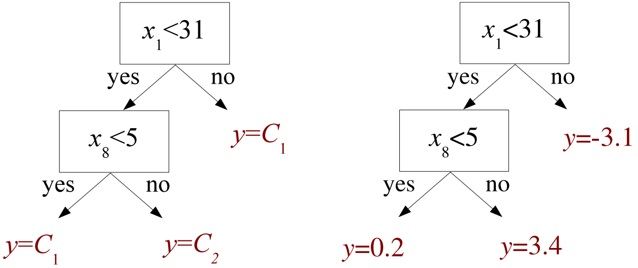
Pierre Geurts

Dept. of EECS, University of Liège, Belgium

CIRM,
Marseille, France
Feb 2, 2016

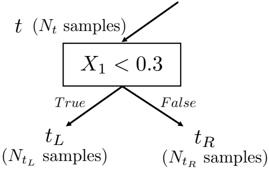
Antonio Sutera, Gilles Louppe, Vân Anh Huynh-Thu, Louis Wehenkel (ULg),
Célia Châtel (Luminy)

Classification and regression trees (Breiman et al., 1984)



- A very popular supervised learning algorithm that uses tree structured input/output models
- The learning procedure:
  - Recursively split the learning sample with tests based on the inputs trying to reduce as much as possible the **impurity** of the output (entropy, variance...)
  - Stop when the output is constant in the leaf or some stopping criterion is met (e.g., depth of the node is above some threshold D)

Impurity reduction



- The best split is the one that maximises impurity reduction:

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

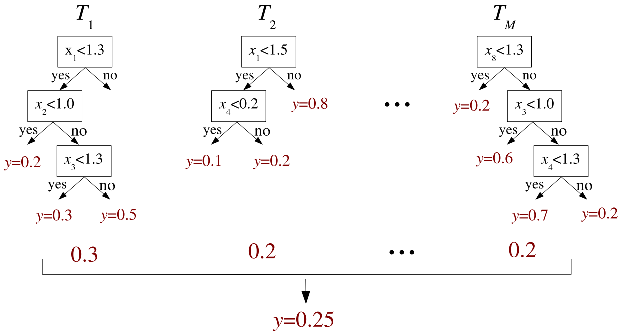
where N_t is the number of samples reaching node t.

- Standard impurity measures:

$$\begin{aligned} i_{regr}(t) &= \frac{1}{N_t} \sum_{i \in S(t)} (y_i - \frac{1}{N_t} \sum_{i \in S(t)} y_i)^2 \text{ (variance)} \\ i_{clas}(t) &= -\sum_c \frac{N_{t,c}}{N_t} \log \frac{N_{t,c}}{N_t} \text{ (Shannon entropy)} \end{aligned}$$

where $N_{t,c}$ is the number of samples of class c in node t.

Ensemble of randomized trees



- Improve trees by reducing their variance
- Many examples: Bagging (Breiman, 1996), Random Forests (Breiman, 2001), Extremely randomized trees (Geurts et al., 2006)
- Breiman (2001)'s Random Forests:
  - Each tree is built from a bootstrap sample
  - The best split at each node is chosen among K inputs selected (*locally*) at random

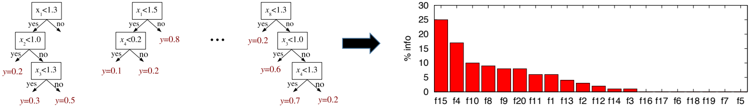
Ensemble of randomized trees: strengths and weaknesses

- 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

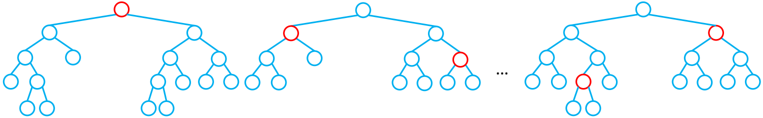
Variable importance scores

- 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)
- We focus here on the MDI measure
  - It is faster to compute (no permutations needed)
  - It does not require to use bootstrap sampling
  - Empirically, it correlates well with the MDA measure (except in specific conditions)

Mean decrease of impurity (MDI): definition



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

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

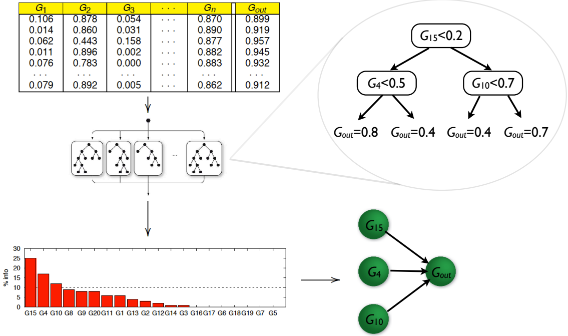
where $p(t) = N_t / N$ and $\Delta i(t)$ is the impurity reduction at node t:

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

One successful application: Gene network inference

GENIE3

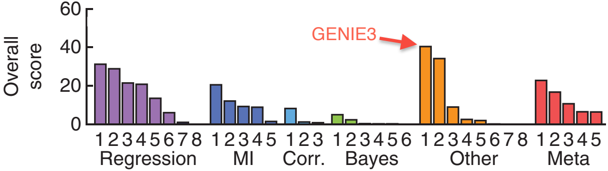
(Huynh-Thu et al, Plos ONE, 2010)



G_1	G_2	G_3	. . .	G_n	G_{out}
0.106	0.878	0.054	. . .	0.870	0.899
0.014	0.860	0.031	. . .	0.890	0.919
0.062	0.443	0.158	. . .	0.877	0.957
0.011	0.896	0.002	. . .	0.882	0.945
0.076	0.783	0.000	. . .	0.883	0.932
. . .	. . .	. . .	. . .	. . .	. . .
0.079	0.892	0.005	. . .	0.862	0.912

DREAM5 competition

(Marbach et al., Nature Methods, 2012)



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

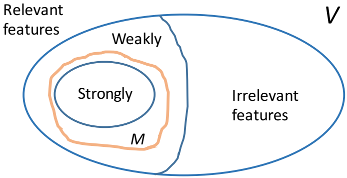
2. Towards a better understanding of the MDI measure

3. Towards large-scale feature selection

Outline

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

Background: Feature relevance (Kohavi and John, 1997)



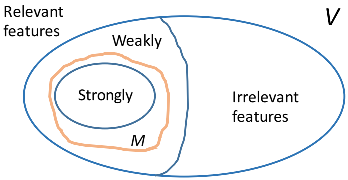
Given an output Y and a set of input variables V, $X \in V$ is

- **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.

The **degree**, $deg(X)$, of a relevant variable X is the smallest size of a subset $B \subseteq V$ such that $Y \not\perp\!\!\!\perp X|B$.

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 $(K = 1, D = p)$

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

Impact of rrelevant variables $(K = 1, D = p)$

Variable importances **depend only on the relevant variables**

Prop. A variable X_m is irrelevant if and only if $Imp(X_m) = 0$.

Prop. 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.

Non-totally randomized trees (K>1, D=p)

Most properties are lost as soon as $K > 1$

$\Rightarrow$ There can be relevant variables with zero importances (due to masking effect).

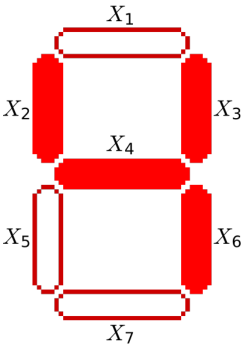
Example:
$I(X_1;Y) = H(Y)$, $I(X_1;Y) \approx I(X_2;Y)$, $I(X_1;Y|X_2) = \epsilon$ and $I(X_2;Y|X_1) = 0$

- $K = 1 \rightarrow Imp_{K=1}(X_1) \approx \frac{1}{2} I(X_1;Y) + \epsilon$ and $Imp_{K=1}(X_1) \approx \frac{1}{2} I(X_2;Y)$
- $K = 2 \rightarrow Imp_{K=2}(X_1) = I(X_1;Y)$ and $Imp_{K=2}(X_2) = 0$.

$\Rightarrow$ The importance of relevant variables can be influenced by the number of irrelevant variables

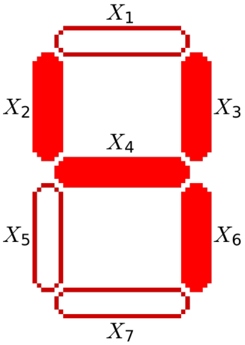
- $K = 2$ and we add a new irrelevant variable $X_3 \rightarrow Imp_{K=2}(X_2) > 0$

Illustration: 7-segment display (Breiman et al., 1984)



y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
0	1	1	1	0	1	1	1
1	0	0	1	0	0	1	0
2	1	0	1	1	1	0	1
3	1	0	1	1	0	1	1
4	0	1	1	1	0	1	0
5	1	1	0	1	0	1	1
6	1	1	0	1	1	1	1
7	1	0	1	0	0	1	0
8	1	1	1	1	1	1	1
9	1	1	1	1	0	1	1

Illustration: 7-segment display (Breiman et al., 1984)



$$Imp(X) = \sum_{k=0}^{p-1} \frac{1}{\binom{p}{k}(p-k)} \sum_{B \in \mathcal{P}_k(V^{-m})} I(X_m; Y|B)$$

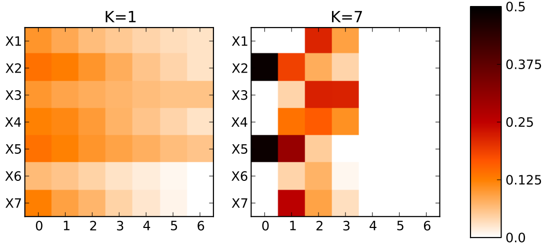
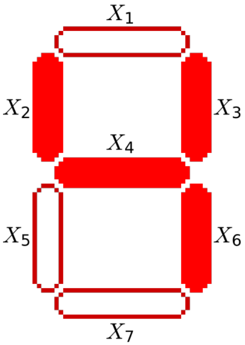
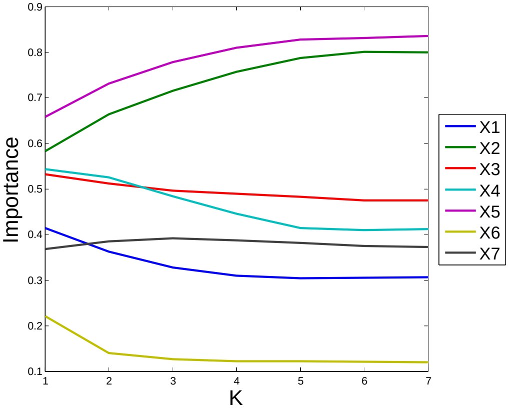


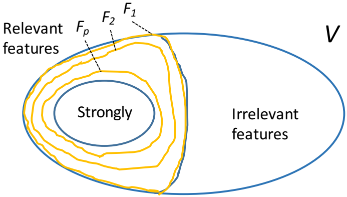
Illustration: 7-segment display (Breiman et al., 1984)





Strongly relevant features can not be masked $(K > 1, D = p)$

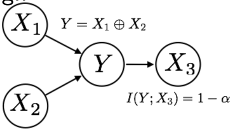
Thm. $\forall K : X$ strongly relevant
$\Rightarrow Imp_K(X) > 0$.



In the case of strictly positive distributions, non random trees always find a superset of the **minimal-optimal** solution.

Note that X_i strongly relevant and X_j weakly relevant does not imply that $Imp(X_i) \geq Imp(X_j)$, even for strictly positive distribution.

E.g.:



$\Rightarrow Imp(X_1)(= Imp(X_2)) < Imp(X_3)$ when $\alpha < \frac{1}{2}$

Impact of pruning $(K = 1, D < p)$

Pruning all trees up to depth $D = d \leq p$ limits importances to **the first d terms** of the decomposition.

$$Imp^{D=d}(X_m) = \sum_{k=0}^{d-1} \frac{1}{\binom{p}{k}(p-k)} \sum_{B \in \mathcal{P}_k(V^{-m})} I(X_m; Y|B)$$

In general, we have $Imp^{D<p}(X_m) \neq Imp^{D=p}(X_m)$ and $Imp^{D<p}(X_m)$ now depends on the number of irrelevant variables.

Pruned totally random trees still solves the **all-relevant problem**, i.e.,

$$X_m \text{ irrelevant iff } Imp^{D<p}(X_m) = 0$$

when either:

- For all relevant variables $X \in V$, $deg(X) < D$
- The number r of relevant variables is $\leq D$

Non asymptotic setting: finite number of trees

In general, a **single tree** can not identify all relevant features, even the strongly relevant ones.

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

Note: All strongly relevant features will however be detected with a single tree with $K = p$ if the following property holds for all $X_1, X_2 \in V, B \subseteq V, b$:

$$Y \perp\!\!\!\perp X_1 | B = b \text{ and } Y \perp\!\!\!\perp X_2 | B = b \Rightarrow Y \perp\!\!\!\perp X_1, X_2 | B = b$$

When $\mathbf{N_T} < \infty$, $Imp_{N_T}(X) > 0$ implies that X is relevant but the opposite is not true.

Non asymptotic setting: finite number of samples

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

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

To avoid false positives, one should use:

- Pruned trees ($D < p$), not to estimate mutual informations from too few samples
- Non totally random trees ($K > 1$), to avoid splits on irrelevant features at the top nodes

From previous analyses, decreasing D or increasing K will however increase the number of false negatives (and also affect the final ranking).

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

1. Tree-based variable importance scores

2. Towards a better understanding of the MDI measure

3. Towards large-scale feature selection

- 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: correctness when $T = \infty$

When $\mathbf{K} = \mathbf{1}$, unpruned trees grown from q random features are strictly equivalent to trees pruned to depth q grown using all features:

- $Imp_{q,\infty}(X) = \sum_{k=0}^{q-1} \frac{1}{\binom{p}{k}(p-k)} \sum_{B \in \mathcal{P}_k(V^{-m})} I(X_m; Y|B)$
- If $deg(X) < q$ for all relevant features X (e.g., when there are q or less relevant features):

$$Imp_{q,\infty}(X) > 0 \text{ iff } X \text{ is relevant.}$$

When $\mathbf{K} > \mathbf{1}$ and there are q or less relevant features:

$$X \text{ strongly relevant } \Rightarrow Imp_{q,\infty}(x) > 0.$$

RS for feature selection: convergence

When $q \ll p$, finding all relevant features may require very large T, depending on variable degrees.

> The probability to sample one feature X of degree $k < q$ together with its minimal conditioning is $\frac{\binom{p-k-1}{q-k-1}}{\binom{p}{q}}$
>
> E.g.: $p = 10000, q = 50, k = 1 \Rightarrow \frac{\binom{p-k-1}{q-k-1}}{\binom{p}{q}} = 2.5 \cdot 10^{-5}$. In average, at least $T = 40812$ trees are required to find X.

When the number of relevant features $r \ll p$, many trees will be grown only from irrelevant features

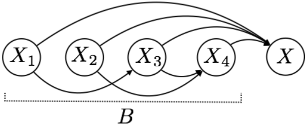
> E.g.: $p = 10000, q = 50, r = 10 \Rightarrow 95\%$ of the trees will not see any relevant features

How to improve convergence at fixed memory size?

Thm. Let B be a minimal subset of V such that $Y \not\perp\!\!\!\perp X|B$ for a relevant X:

- All $X_i \in B$ are relevant and $deg(X_i) \leq |B|$.
- For a PC distribution[1], there exists an ordering $\{X_1, \ldots, X_k\}$ of the variables in B such that

$$\forall 1 \leq i \leq k : \exists B' \subseteq \{X_1, \ldots, X_{i-1}\} : Y \not\perp\!\!\!\perp X_i|B'$$



$\Rightarrow$ Suggests that RS convergence can be improved by enforcing the selection of previously found relevant features.

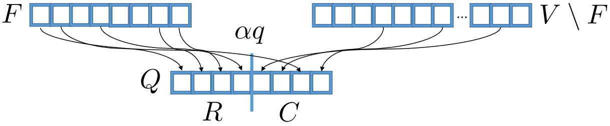
[1] A strictly positive distribution is PC iff it satisfies the composition property, i.e., for any disjoint subsets R, T, R of V, we have:

$$Y \perp\!\!\!\perp T|R \text{ and } Y \perp\!\!\!\perp U|R \Rightarrow Y \perp\!\!\!\perp T \cup U|R.$$

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 $F$
```



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

SRS correctness when $T = \infty$

If there are **less than q relevant features**, SRS provides the same guarantee as RS, whatever α.

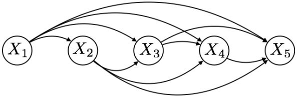
- With $K = 1$, F will contain all relevant features
- With $K > 1$, F will be a superset of the strongly relevant features

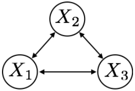
If there are **more than q relevant features**, $\alpha > 0$ introduces some **masking** effect. But we still have the following result:

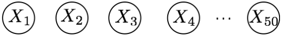
> When $K = 1$, F will contain all relevant features X such that $deg(X) < (1 - \alpha)q$

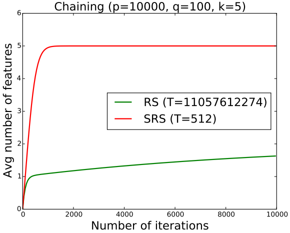
One should thus choose q according to the expected number of relevant features and α according to the expected maximum feature degree.

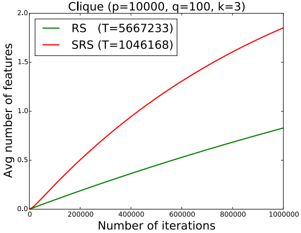
SRS convergence in specific scenarios (exact computations)

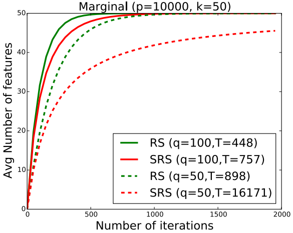












Average number $\overline{N}$ of iterations to find all k variables (with $k \ll q$):

- Chaining: $\overline{N}_{RS} \simeq (\frac{p}{q})^k$ and $\overline{N}_{SRS} \simeq k\frac{p}{q}$
- Clique: $\overline{N}_{RS} \simeq k(\ln k + 1)(\frac{p}{q})^k$ and $\overline{N}_{SRS} \simeq (\frac{p}{q})^k$

Assumptions: $K = q$, all relevant variables are strongly relevant and the same variable used at all nodes of a given level

Practical implementation

```
1. Let $F = \emptyset$
2. Repeat $T$ times:
    2.1 Let $Q = R \cup C$, where ...
    2.2 Grow a tree only using features in $Q$
    2.3 Add to $F$ all features that get non-zero importance
3. Return $F$
```

In practice, even irrelevant features can get non-zero importance

Practical implementation:

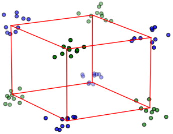
- A random probe is added to the q features at each iteration
- F contains all features that
  - were sampled more than L times in Q sets
  - were more important than the random probe in at least β percent of the trees

  (In the experiments: $L = 5$, $\beta = 95\%$)
- Output: ranking of the features according to importance

Experiments: protocol

Madelon data (Guyon et al., 2007)

- 1500 samples (|LS|=1000, |TS|=500)
- 20 relevant features: 5 features that define Y, 5 random linear combinations of the first 5, and 10 noisy copies of the first 10
- Increasing number of irrelevant features: 480, 1480, 2980, 5480



Parameters: $q = 50$, $K = q$, no bootstrap, threshold randomization (Geurts et al., 2006)

Evaluation:

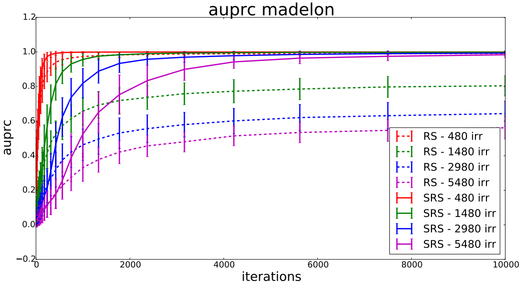
- Average over 50 random LS/TS splits
- Evolution of TS accuracy with number of iterations
- Evolution of the area under the precision-recall curve (auprc) with number of iterations, when features are ranked according to importances

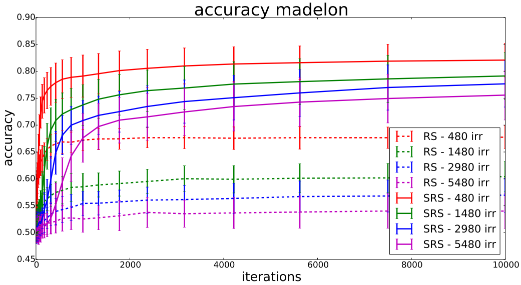
Experiments: results

Important improvement of both auprc and accuracy with SRS

The lower q/p, the larger the improvement

Only SRS always eventually perfectly ranks the features





Conclusions

Future works on SRS:

- More experiments on real data
- 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.

http://www.montefiore.ulg.ac.be/~geurts

Importances of strongly versus weakly features

The importance of strongly relevant features is not higher than the importance of weakly relevant features, even in the case of strictly positive distribution.

Example: $Y = X_1 \oplus X_2$, $X_3 \not\perp\!\!\!\perp X_1$, $X_3 \perp\!\!\!\perp X_2$, and $Y \not\perp\!\!\!\perp X_3$ (X_1 and X_2 are strongly relevant and X_3 is weakly relevant):

p	X_1	X_2	X_3	Y
$\frac{(1-\epsilon)}{4}$	0	0	0	1
$\frac{\epsilon}{4}$	0	0	1	1
$\frac{\epsilon}{4}$	0	1	0	0
$\frac{(1-\epsilon)}{4}$	0	1	1	0
$\frac{\epsilon}{4}$	1	0	0	0
$\frac{(1-\epsilon)}{4}$	1	0	1	0
$\frac{(1-\epsilon)}{4}$	1	1	0	1
$\frac{\epsilon}{4}$	1	1	1	1

With $\epsilon = 0.05$, we have: $I(Y; X_1) = I(Y; X_2) = 0$, $I(Y; X_1, X_2) = 1$,

$I(Y; X_3) = 1 - (1-\epsilon)\log(1-\epsilon) - \epsilon\log(\epsilon) = 0.714$, $I(Y; X_3|X_1) = 0.714$,

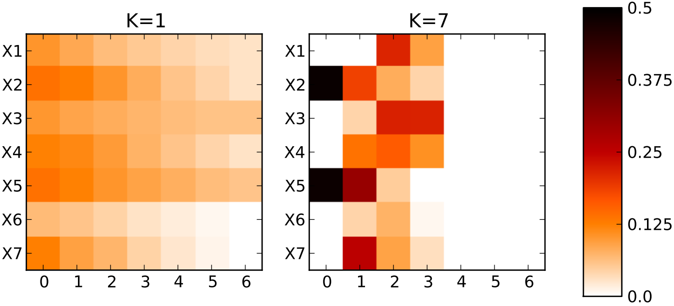
$I(Y; X_1|X_2, X_3) = I(Y; X_2|X_1, X_3) = 0.286.$, $I(Y; X_1|X_3) = I(Y; X_2|X_3) = 0$ And thus:

$Imp(X_1) = Imp(X_2) = \frac{1}{6}(1+0) + \frac{1}{3}0.286 = 0,262$

$Imp(X_3) = \frac{1}{3}0.714 + \frac{1}{6}(0.714 + 0.714) = 0,476$ X_3 is thus more important than X_1 and X_2 despite being weakly relevant.

Illustration: decomposition ($K = 1$ and $K = 7$)

Imp	$K = 1$	$K = 7$
X_1	0.412	0.306
X_2	0.581	0.799
X_3	0.531	0.475
X_4	0.542	0.412
X_5	0.656	0.835
X_6	0.225	0.120
X_7	0.372	0.372



Learning with a memory constraint

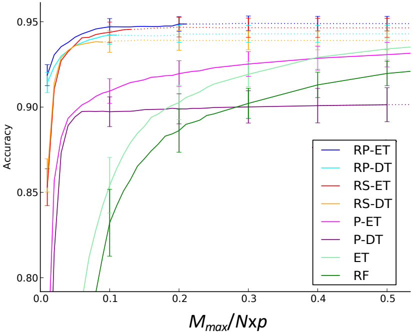
Simplistic constrained setting:

Memory size M_{max} of computing node(s) is small with respect to dataset size $N \times p$.

Straightforward ensemble-based solutions:

Build a forest where each tree is trained from a subset of:

- $N_s \leq M_{max}/p$ samples (**P**asting, Breiman, 1999, Chawla *et al.*, 2004)
- $N_f \leq M_{max}/N$ features (**R**andom **S**ubspace, Ho, 1998)
- N_s samples and N_f inputs such that $N_s N_f \leq M_{max}$ (**R**andom **P**atches, Louppe et al., 2012)



Learning with a memory constraint

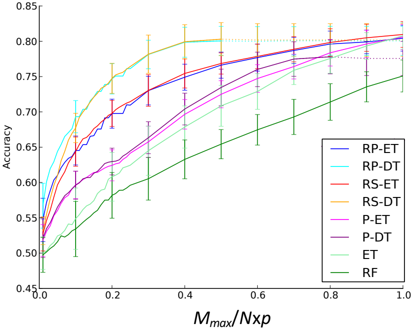
Simplistic constrained setting:

Memory size M_{max} of computing node(s) is small with respect to dataset size $N \times p$.

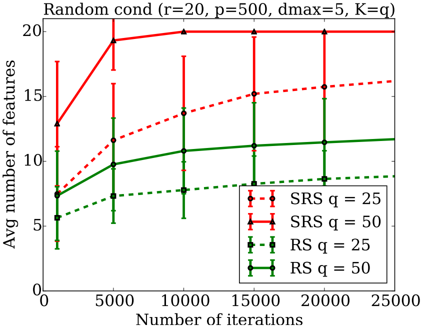
Straightforward ensemble-based solutions:

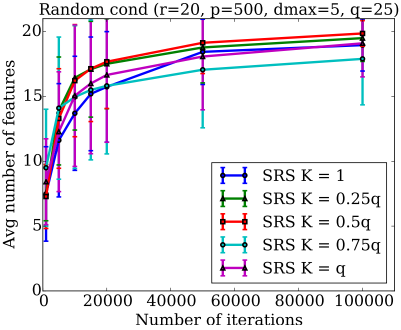
Build a forest where each tree is trained from a subset of:

- $N_s \leq M_{max}/p$ samples (**P**asting, Breiman, 1999, Chawla *et al.*, 2004)
- $N_f \leq M_{max}/N$ features (**R**andom **S**ubspace, Ho, 1998)
- N_s samples and N_f inputs such that $N_s N_f \leq M_{max}$ (**R**andom **P**atches, Louppe et al., 2012)



SRS convergence: numerical simulations in mixed setting



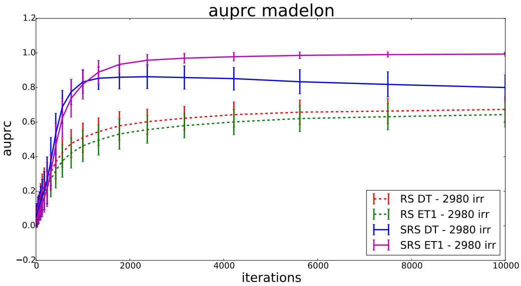


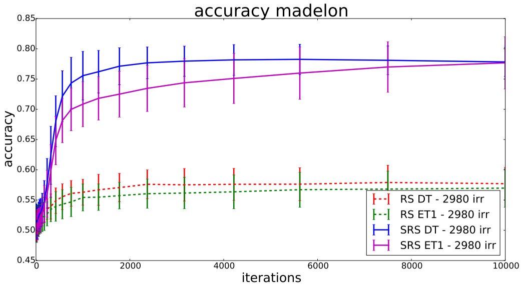
- Degrees and minimal conditionings of relevant features are selected at random, with maximum degree 5.
- $p = 500$ features, $r = 20$ relevant ones, $q = 25, 50$, $K \in \{1, 0.25q, 0.5q, 0.75q, q\}$

Experiments: results

DT versus ET

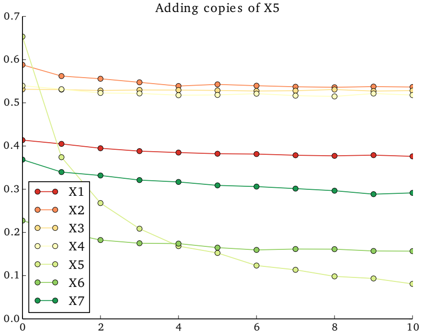
From theory, DT should find all strongly relevant features

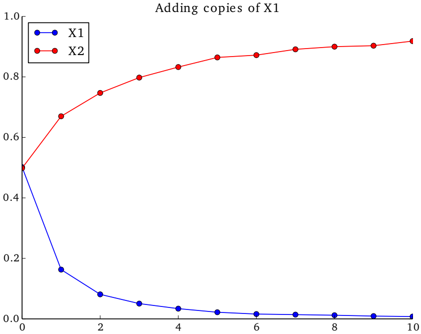




Redundant variables $(K = 1, D = p)$

Adding copies of an existing variable decreases the relevance of both copies and increases the relevance of variables in interactions with these variables


Adding copies of X5
X1
X2
X3
X4
X5
X6
X7


Adding copies of X1
X1
X2