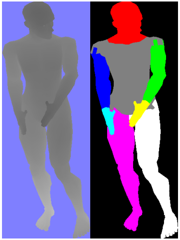
$L1$-based compression of random forest model

Arnaud Joly, François Schnitlzer, Pierre Geurts, Louis Wehenkel
a.joly@ulg.ac.be - www.ajoly.org

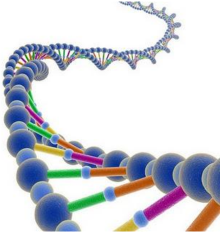
Departement of EE and CS & GIGA-Research

Université de Liège
GIGA

High dimensional supervised learning applications



3D Image segmentation



Genomics

Electrical grid

x=original image
y=segmented image

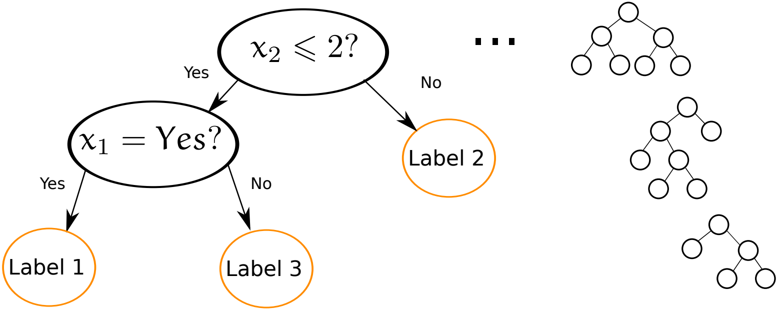
x=DNA sequence
y=phenotype

x=system state
y=stability

From 10^5 to 10^9 dimensions.

Tree based ensemble methods

From a dataset of input-output pairs $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathcal{X} \times \mathcal{Y}$, we approximate $f : \mathcal{X} \to \mathcal{Y}$ by learning an ensemble of M decision trees.



The estimator $\hat{f}$ of f is obtained by averaging the predictions of the ensemble of trees.

High model complexity $\rightarrow$ Large memory requirement

The complexity of tree based methods is measured by the number of internal nodes and increases with

- the ensemble size M;
- the number of samples n in the dataset.

The variance of individual trees increases with the dimension p of the original feature space $\rightarrow$ $M(p)$ should increase with p to yield near optimal accuracy.

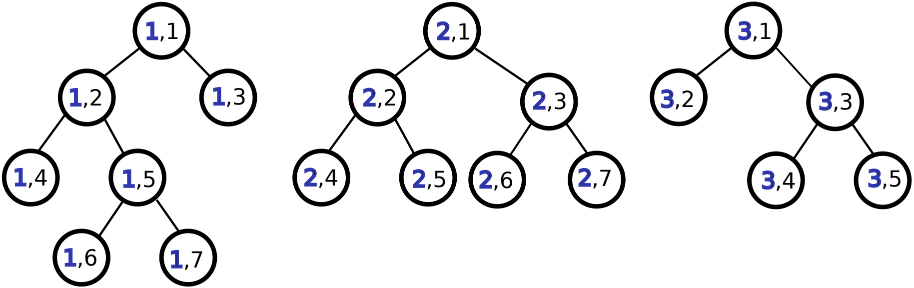
Complexity grows as $nM(p)$ $\rightarrow$ may require huge amount of storage.

Memory limitation will be an issue in high dimensional problems.

$L1$-based compression of random forest model (I)

We first learn an ensemble of M extremely randomized trees (Geurts, *et al.*, *2006*) . . .

Example



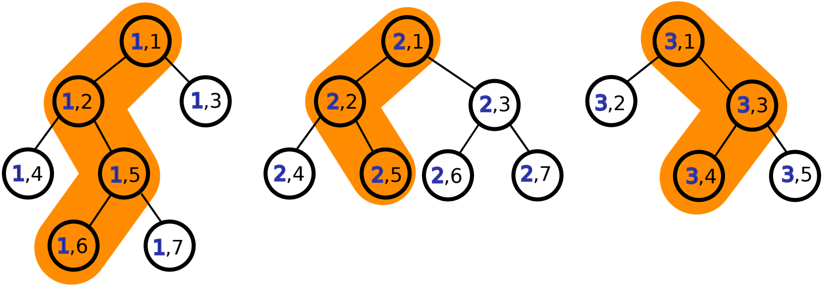
. . . and associate to each node an indicator function $1_{m,l}(x)$ which is equal to 1 if the sample (x, y) reaches the l-th node of the m-th tree, 0 otherwise.

$L1$-based compression of random forest model (II)

The node indicator functions $1_{m,l}(x)$ may be used to lift the input space $\mathcal{X}$ towards its induced feature space $\mathcal{Z}$

$$z(x) = (1_{1,1}(x), \ldots, 1_{1,N_1}(x), \ldots, 1_{M,1}(x), \ldots, 1_{M,N_M}(x))\,.$$

Example for one sample x_s



$$z(x_s) = (1\;1\;0\;0\;1\;1\;0 \mid 1\;1\;0\;0\;1\;0\;0 \mid 1\;0\;1\;1\;0)$$

$L1$-based compression of random forest model (III)

A variable selection method (regularization with $L1$-norm) is applied on the induced space $\mathcal{Z}$ to compress the tree ensemble using the solution of

$$\left(\beta_j^*(t)\right)_{j=0}^{q} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{q} \beta_j \, z_j(x_i) \right)^2$$

$$\text{s.t. } \sum_{j=1}^{q} |\beta_j| \leq t.$$

Pruning: a test node is deleted if all its descendants (including the test node itself) correspond to $\beta_j^*(t^*) = 0$.

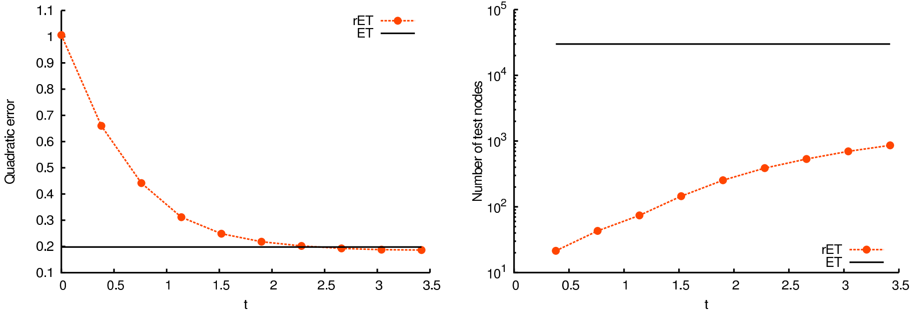
Overall assessment on 3 datasets

Datasets	Error		Complexity		
	ET	rET	ET	rET	ET/rET
Friedman1	0.19587	0.18593	29900	885	34
Two-norm	0.04177	0.06707	4878	540	9
SEFTi	0.86159	0.84131	39436	2055	19

[ET] Extra trees;
[rET] L1-based compression of ET.

Table: Parameters of the Extra-Tree method: $M = 100$; $K = p$; $n_{\min} = 1$ on Friedman1 and Two-norm, $n_{\min} = 10$ on SEFTi.

An increase of t decreases the error of rET until t = 3 with drastic pruning



(a) Estimated risk (b) Complexity

Friedman1 : $M = 100$, $K = p = 10$ and $n_{\min} = 1$

Managing complexity in the extra tree method

Bound M

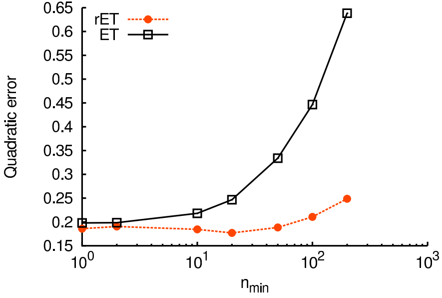
Restrict the size M of the tree based ensemble.

Pre-pruning

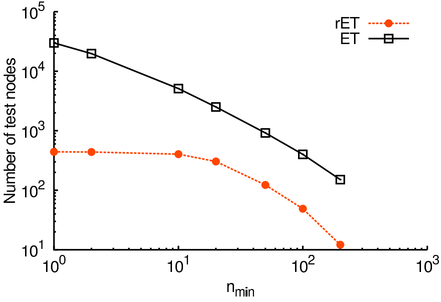
Pre-pruning reduces the complexity of tree based methods by imposing a condition to split a node *e.g.*

- minimum number of samples n_{min} in order to split,
- minimum decrease of an impurity measure,
- ...

The accuracy and complexity of an rET model does not depend on $n_{\min}$, for $n_{\min}$ small enough ($n_{\min} < 10$)



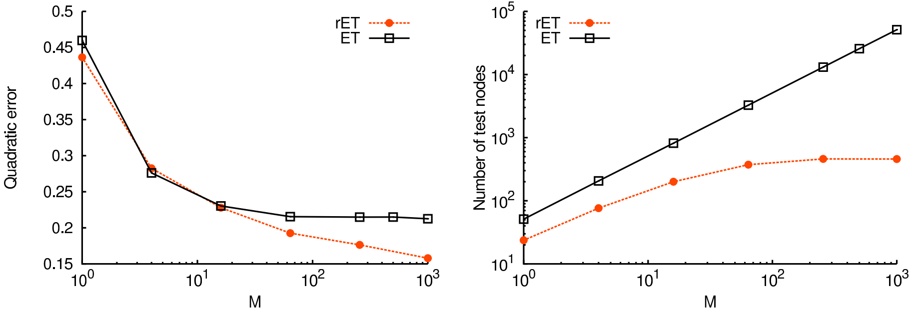
(c) Estimated risk



(d) Complexity

Friedman1 : $M = 100$, $K = p = 10$ and $t = t^*_{cv}$

After variance reduction has stabilized ($M \simeq 100$), further increasing M keeps enhancing the accuracy of the rET model without increasing complexity



(e) Estimated risk (f) Complexity

Friedman1 : $n_{\min} = 10$, $K = p = 10$ and $t = t^*_{cv}$

Conclusion & perspectives

1. Drastic pruning while preserving accuracy.
2. Strong compressibility of the tree ensemble suggests that it could be possible to design novel algorithms suited to very high dimensional input space.
3. Future research will target similar compression ratio without using the complete set of node indicator functions of the forest model.

$L1$-based compression of random forest model

Arnaud Joly, François Schnitlzer, Pierre Geurts, Louis Wehenkel
a.joly@ulg.ac.be - www.ajoly.org

Departement of EE and CS & GIGA-Research

Appendix

Overall assessment on 3 datasets

Datasets	Error			Complexity		
	ET	rET	Lasso	ET	rET	Lasso
Friedman1	0.19587	0.18593	0.282441	29900	885	4
Two-norm	0.04177	0.06707	0.033500	4878	540	20
SEFTi	0.86159	0.84131	0.988031	39436	2055	14

[ET] Extra trees;
[rET] L1-based compression of ET.

Table: Overall assessment (parameters of the Extra-Tree method: $M = 100$; $K = p$; $n_{\min} = 1$ on Friedman1 and Two-norm, $n_{\min} = 10$ on SEFTi).