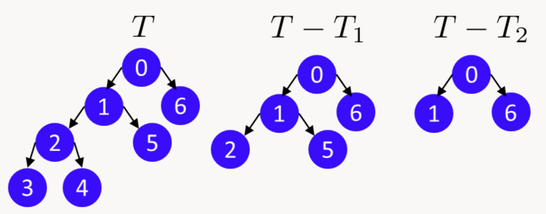
Cost-complexity pruning of random forests

ISMM 2017, 13th International Symposium on Mathematical Morphology, Fontainebleau, France, May 15 - 17, 2017

Why perform pruning ?

- Out-of-bag samples are un-used samples from the Bootstrap Aggregation procedure in random forests.
- We study the effect of using the out-of-bag samples to improve the generalization error first of the decision trees, and second the random forest by post-pruning.

Decision Tree Pruning



$\text{Leaves}(T) = \{3, 4, 5, 6\}$

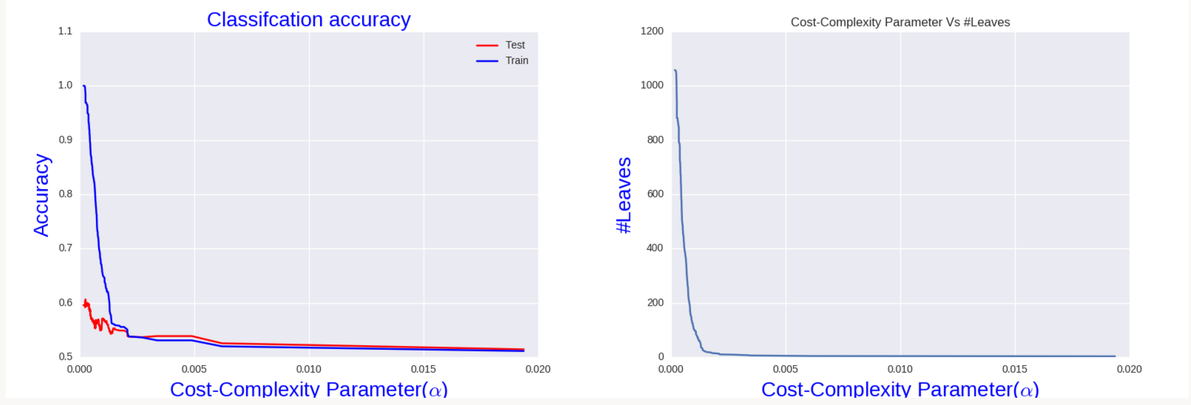
$\text{Internal nodes} = \{1, 2\}$

$\text{Sequence of subtrees} : T \supseteq T - T_2 \supseteq T - T_1$

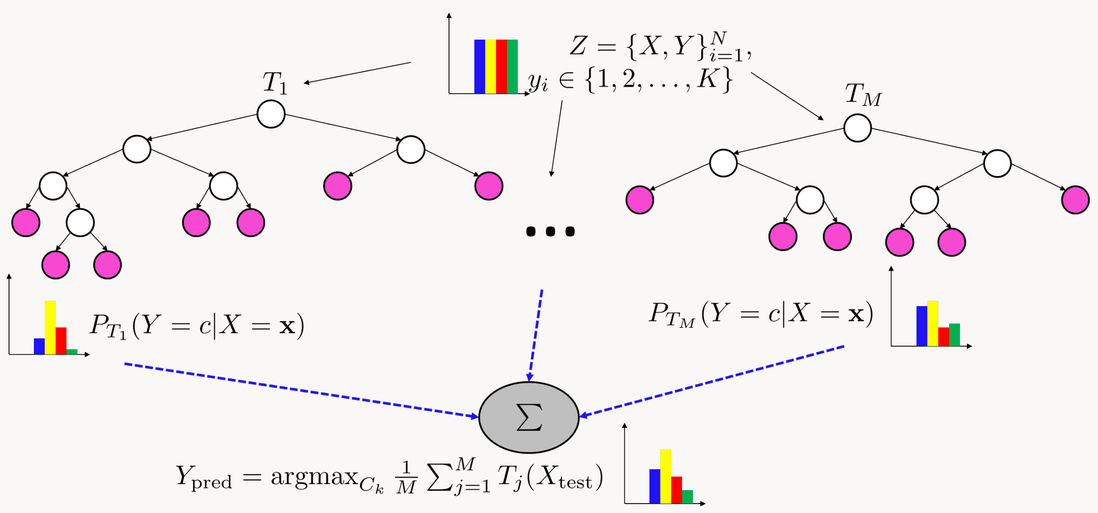
$\text{Cost complexity values} : g(t) = 0, \alpha_2, \alpha_1$

$\text{Final set of trees and parameters} : \mathcal{T}, \mathcal{A}$

- Cost-Complexity function : $g(t) = \frac{R(t) - R(T_t)}{|\text{Leaves}(T_t)| - 1}$
- $R(T) = \sum_{t \in \text{Leaves}(T)} r(t) \cdot p(t) = \sum_{t \in \text{Leaves}(T)} R(t)$
- $R(T)$ is the training error, $\text{Leaves}(T)$ is #leaves of tree T
- $r(t) = 1 - \max_k p(C_k)$ is the misclassification rate and $p(t) = n_t / N$ is the number of samples in node n_t to total training samples N.



Random Forests



Decision tree ensembles : Random Forests (RF), Extremely Randomize Trees (ET), Bagged trees (BT)
Randomization Methods : Boostrap Aggregation, Random Feature selection, Random Threshold section

Out-Of-Bag Cost Complexity Pruning

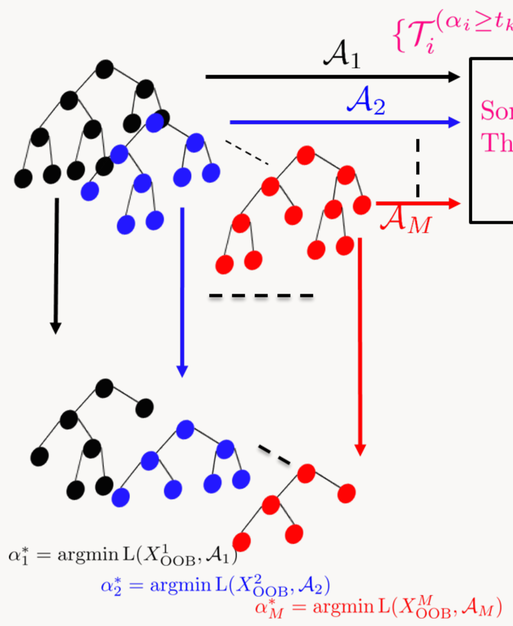
Independent tree pruning :

$$\mathcal{T}_j^* = \text{argmin}_{\alpha \in \mathcal{A}_j} \mathbb{E}\left[\| Y_{\text{OOB}} - \mathcal{T}_j^{(\alpha)}(X_{\text{OOB}}^j) \|^2 \right]$$

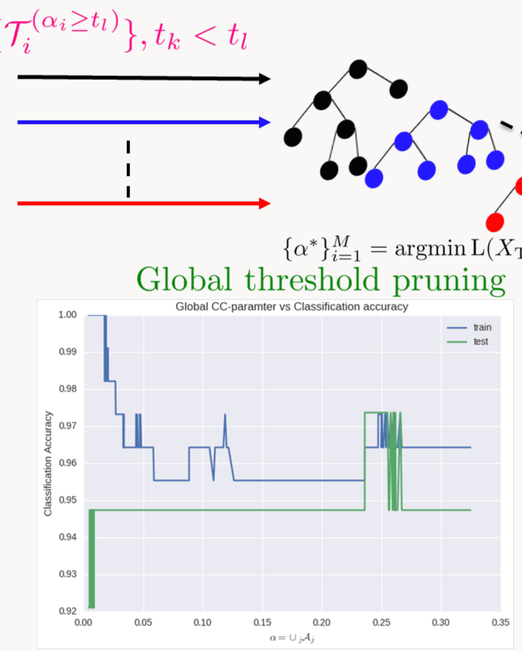
Global threshold pruning :

$$\{\mathcal{T}_j^*\}_{j=1}^{M} = \text{argmin}_{\alpha \in \cup_j \mathcal{A}_j} \mathbb{E}\left[\| Y_{\text{train}} - \frac{1}{M} \sum_{j=1}^{M} \mathcal{T}_j^{(\alpha)}(X_{\text{OOB}}^j) \|^2 \right]$$

Overview of method



Independent tree pruning



Global CC-pruning CV-accuracy plot

Results and Analysis

Plots of $\mathcal{A}_j \forall j$ for RFs, ETs, BTs :

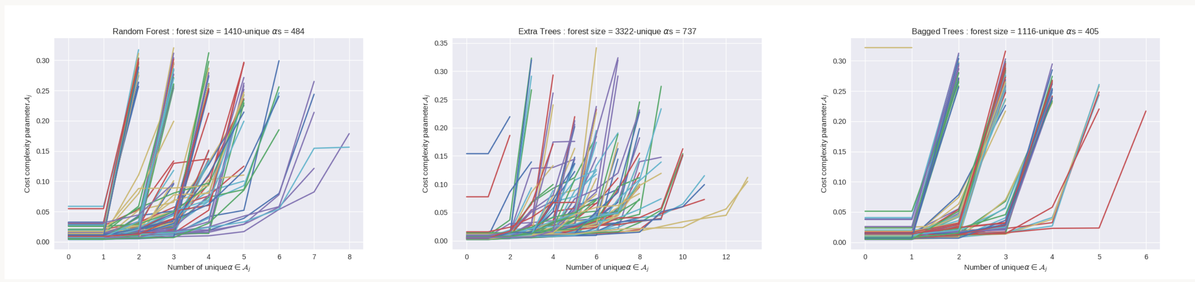
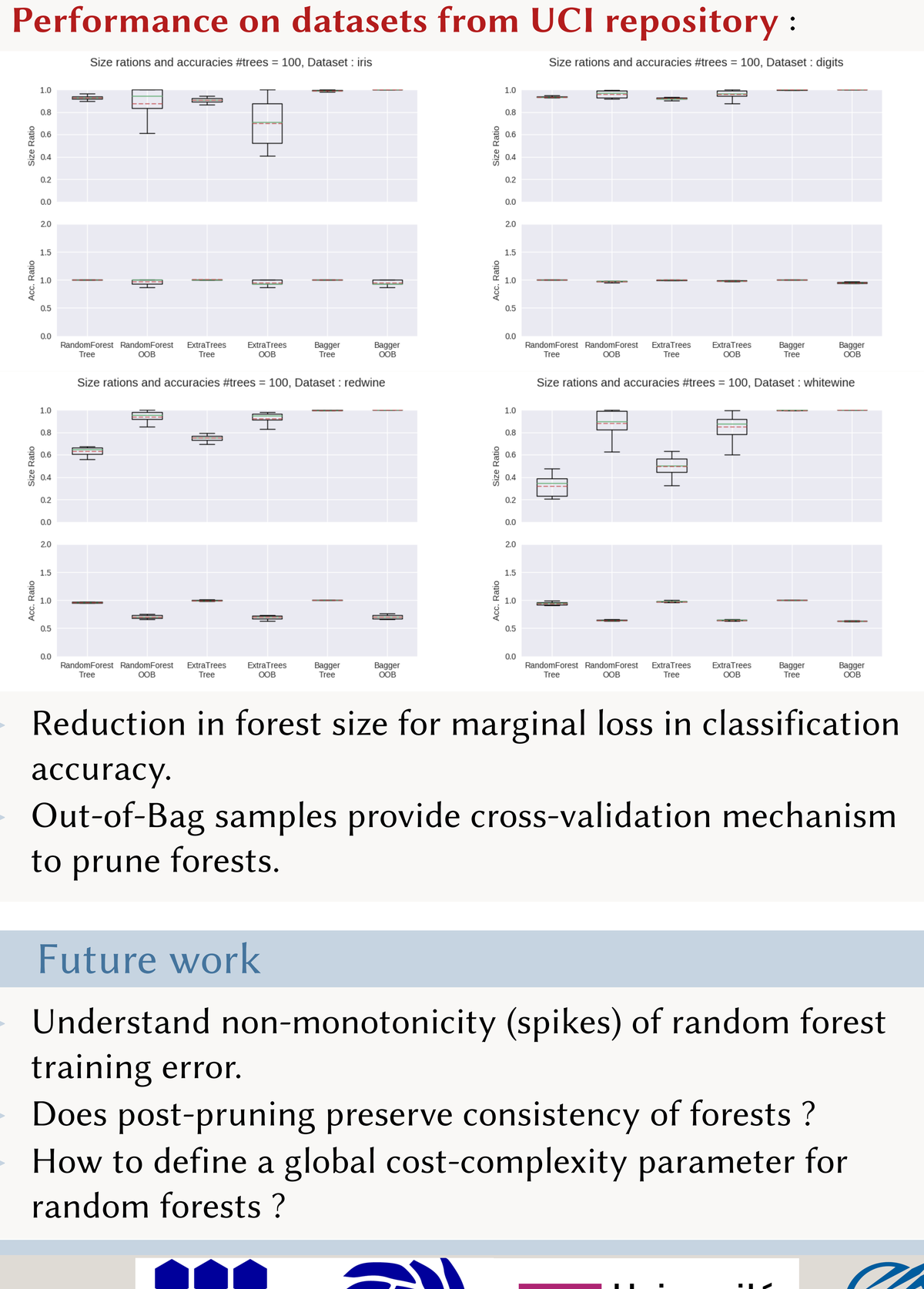


Figure: RFs and ETs provide a larger subset of CC-parameter values $\mathcal{A}_j$ and thus subtrees $\mathcal{T}_j$ for the cross-validation step.

Performance on datasets from UCI repository :



- Reduction in forest size for marginal loss in classification accuracy.
- Out-of-Bag samples provide cross-validation mechanism to prune forests.

Future work

- Understand non-monotonicity (spikes) of random forest training error.
- Does post-pruning preserve consistency of forests ?
- How to define a global cost-complexity parameter for random forests ?

B Ravi Kiran[*], **Jean Serra**[+]
beedotkiran@gmail.com, https://beedotkiran.github.io/forest.html
Université Lille 3, CRIStAL, Université[*], Ecole des Mines de Paris, CMM [+]