Random Forests

—

Parametrization, Tree Selection and Dynamic Induction

Simon Bernard

Document and Learning research team
LITIS lab.
University of Rouen, France

décembre 2014

Random Forests

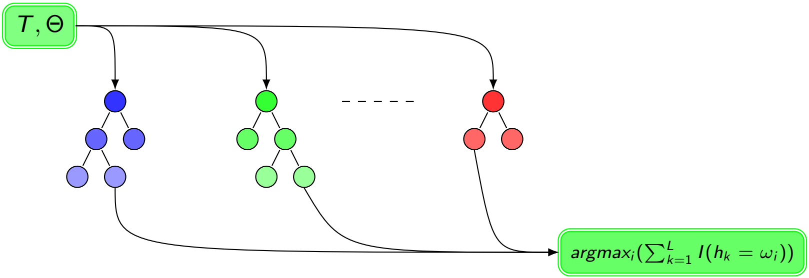
Definition [Breiman 2001][1]

Definition

A random forest is a classifier consisting of a collection of tree-structured classifiers, noted

$$\{\ h_k = h(x, \theta_k), \quad k = 1, ..., L\ \}$$

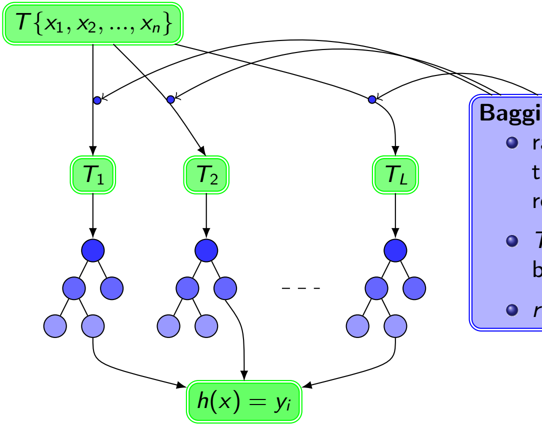
where the $\{\theta_k\}$ are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.



1. L. Breiman, *Random Forests*. Machine Learning, vol.45, num.1, pp 5–32, 2001

Reference algorithm Forest-RI
[Breiman 2001] [1]

Two randomization principles :



Bagging

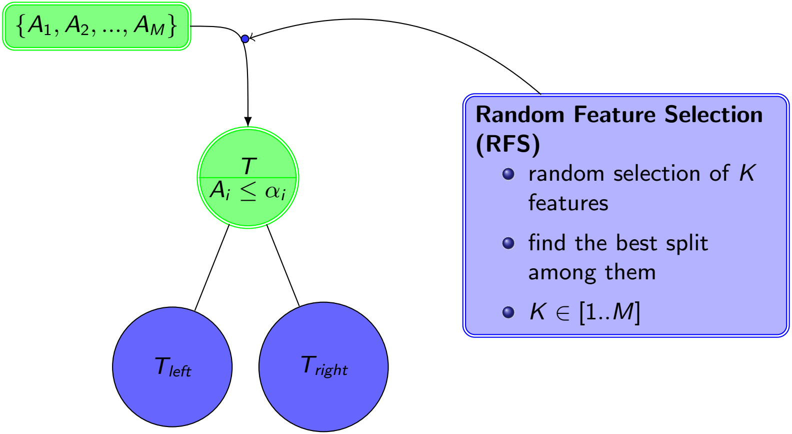
- random sampling of r training instances with replacement.
- T_i are called bootstrap samples.
- r usually equal to n.

1. L. Breiman, *Random Forests*. Machine Learning, vol.45, num.1, pp 5–32, 2001

Reference algorithm Forest-RI

[Breiman 2001] [1]

Two randomization principles :



1. L. Breiman, *Random Forests*. Machine Learning, vol.45, num.1, pp 5–32, 2001

Random Forests
Pros and Cons

Random Forest are efficient classifiers... [Fernández et al. JMLR, 2014]
But this efficiency is sometimes difficult to obtain

Example : the Madelon dataset (2600 instances, 500 features ($= M$), 2 classes)

Forest-RI : $K = 22$ ($\sqrt{M}$), $L = 300$
$\rightarrow$ test error rate $= 30.50\%$

Forest-RI : $K = 260$, $L = 300$
$\rightarrow$ test error rate $= 17.73\%$

Forest-RI : $K = 260$, $L = 100$ (tree selection)
$\rightarrow$ test error rate $= \mathbf{15.96}\%$

Why and How these performances fluctuate ?

Random Feature Selection

Control the Randomness

What do we know ?

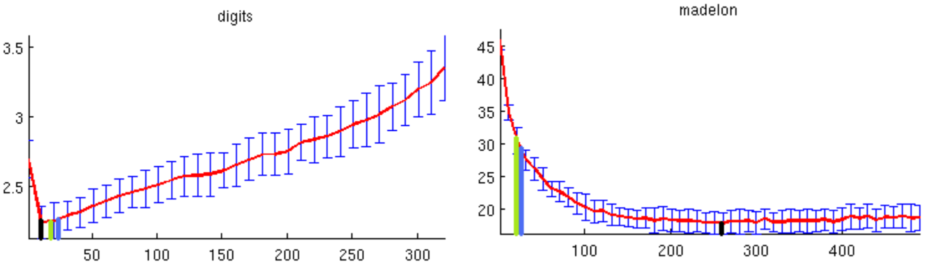
Control of the randomness : hyperparameter K

K : number of random features used for each node splitting

- Allow to control the "amount" of randomness used during the induction procedure

K	1	2	...	$M-1$	M
Randomness	*max*		$\leftarrow \oplus \ldots \ominus \rightarrow$		$\emptyset$

- Several arbitrary values in the literature : 1, $\sqrt{M}$, $\lceil log_2(M) \rceil$



Mean test error rate with respect to K
($K = \sqrt{M}$, $K = \lceil log_2(M) \rceil$, $K = K^*$)

Exhaustive search for K^*

Intuition : the best value for K depends on the relevancy of the features [2]

Experimental protocol :

- 20 datasets, 50 random splits Training/Test for each
- McNemar statistical test of significance
- Exhaustive search of K^* : all possible values between 1 and M are tested
  $\rightarrow$ K^* : the best value in average, over the 50 splits
- Measure the *information gain* for each feature (estimate the relevancy)

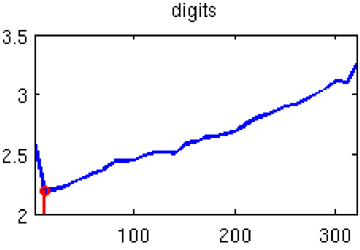
$$Gain(T, A_i) = \Delta I(T, A_i) = I(T) - I(T, A_i)$$

$$I(T) = \sum_{j=1}^{c} -\frac{n_{j.}}{n_{..}} \log_2 \frac{n_{j.}}{n_{..}} \qquad I(T, A_i) = \sum_{k=1}^{m_i} \frac{n_{.k}}{n_{..}} I(T_k))$$

2. Geurts et al, "Extremely Randomized Trees", Machine Learning, 2006

Exhaustive search for K^*

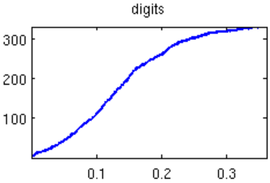
2 types of results :



Mean error rates w.r.t. K

17 of the 20 datasets

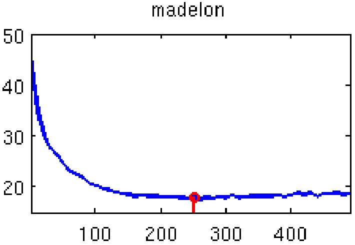
- $K = \sqrt{M}$ is always a good choice
- high proportion of relevant features



Cumulative number of features w.r.t. to information gain

Exhaustive search for K^*

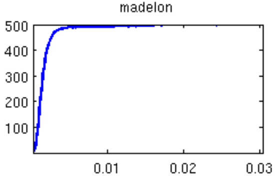
2 types of results :



Mean error rates w.r.t. K

3 of the 20 datasets

- None of the given settings is satisfying
- Very few relevant features



Cumulative number of features w.r.t. to information gain

The Forest-RK algorithm

An alternative with a new push-button algorithm, called Forest-RK [3]

Forest-RK

- Same as Forest-RI, *i.e.* combines Bagging and RFS.
- Except that the value of K is randomly set for each node.

  $\rightarrow$ K is not an hyperparameter of the method anymore.

  $\rightarrow$ increase diversity by allowing some trees to be "less random"

3. S. Bernard, L. Heutte, S. Adam, *Forest-RK : A New Random Forest Induction Method*, ICIC, 2008.

The Forest-RK algorithm
Evaluation

Forest-RK $\equiv$ Forest-RI/$K_{\sqrt{M}}$ on the 17 "regular" datasets

For the 3 "atypical" datasets (with very few relevant features) :

	DigReject	Madelon	Musk
Forest-RI/K^*	7.12 ± 0.34	17.73 ± 1.60	2.34 ± 0.30
Forest-RI/$K_{\sqrt{M}}$	7.58 ± 0.34	30.50 ± 1.94	2.40 ± 0.29
Forest-RK	7.16 ± 0.33	18.34 ± 1.52	2.34 ± 0.31
Test de McNemar			
RK vs RI/$K_{\sqrt{M}}$	RK	RK	$\equiv$
RK vs RI/K^*	$\equiv$	$\equiv$	$\equiv$

$\rightarrow$ Forest-RK $>$ Forest-RI/$K_{\sqrt{M}}$

Forest-RK $\equiv$ *Forest-RI/K^** **for all the 20 datasets**

Tree Selection and Dynamic Induction

Control the Diversity

What do we know ?

Generalization error convergence

For an increasing number of trees in the forest, generalization error rate converges to a minimum.[1 4 5]

- **Strength** : $s = E_{X,Y}[mr(X,Y)]$

  where $mr(X,Y) = P_\Theta(h(X,\Theta) = Y) - max_{j \neq Y} P_\Theta(h(X,\Theta) = j)$ is the margin of the forest

- **Correlation** : $\overline{\rho} = E_{\Theta,\Theta'}[\rho(rmg(\Theta,X,Y), rmg(\Theta',X,Y))]$

  where $rmg(\Theta,X,Y) = I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{\jmath}(X,Y))$ is the raw margin of a tree

- **Generalization error bound** :

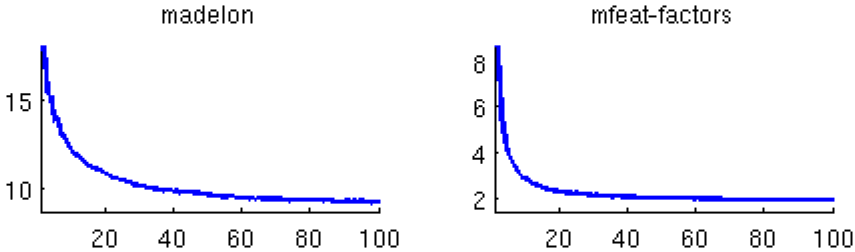
$$PE^* \leq \frac{\overline{\rho}(1 - s^2)}{s^2} \leq \frac{\overline{\rho}}{s^2} \tag{1}$$

---

1. L. Breiman, *Random Forests*. Machine Learning, vol.45, num.1, pp 5–32, 2001
4. Latinne et al., *Limiting the Number of Trees in Random Forests*, MCS, 2001.
5. Bernard et al., *Using Random Forests for Handwritten Digit Recognition*, ICDAR, 2007.

What do we know ?
Generalization error convergence

For an increasing number of trees in the forest, generalization error rate converges to a minimum.[1 4 5]



1. L. Breiman, *Random Forests*. Machine Learning, vol.45, num.1, pp 5–32, 2001
4. Latinne et al., *Limiting the Number of Trees in Random Forests*, MCS, 2001.
5. Bernard et al., *Using Random Forests for Handwritten Digit Recognition*, ICDAR, 2007.

Analysis of several sub-forests

What makes an ensemble of trees more accurate than another ?

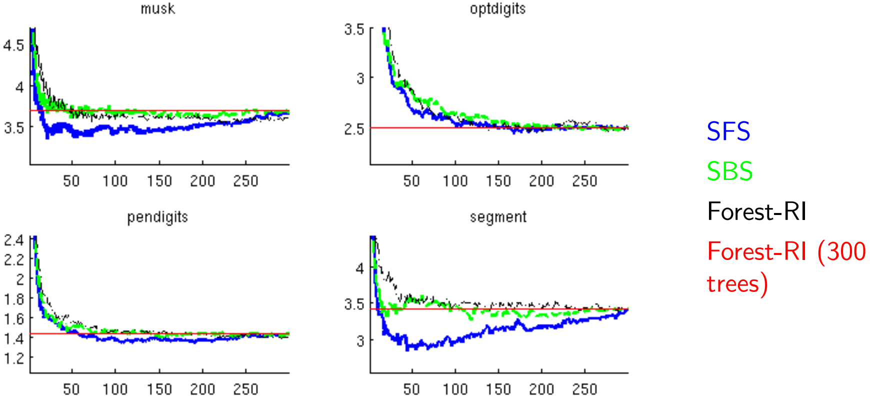
→ Generate different sub-forests and examine their performances

Experimental protocol :

- 20 datasets, 50 random splits Training/Test
- Sequential classifiers selection techniques :
  - Sequential Forward Search (SFS)
  - Sequential Backward Search (SBS)

  → Selection criteria : validation error rate

Analysis of several sub-forests
Results



Error rate w.r.t. the number of trees in the sub-forests

→ 18 datasets : at least one sub-forest significantly better than the forest
→ Sometimes only 10% of the trees can reach the performance of the forest

Analysis of several sub-forests

Results

Datasets	SFS error rates	SFS # trees	SBS error rates	SBS # tree	Forest-RI 300 trees
Diabetes	23.57	142	23.61	32	24.31
Digits	2.45	276	2.46	289	2.47
DigReject	7.53	289	7.54	245	7.55
Gamma	12.60	76	12.67	219	12.71
Isolet	6.93	213	6.73	249	7.03
Letter	6.21	265	6.27	228	6.28
Madelon	33.01	103	34.33	101	34.88
Mfeat-factors	3.98	153	4.18	236	4.28
Mfeat-fourier	17.37	297	17.33	292	17.46
Mfeat-karhunen	4.80	272	4.74	268	4.83
Mfeat-zernike	22.39	298	22.26	281	22.40
MNIST	5.92	210	5.93	288	5.95
Musk	3.35	21	3.61	49	3.70
OptDigits	2.43	168	2.47	274	2.50
Page-block	3.01	105	2.96	88	3.06
Pendigits	1.34	191	1.40	154	1.44
Segment	2.85	44	3.26	119	3.43
Spambase	5.30	47	5.45	161	5.51
Vehicle	24.75	79	25.35	53	26.17
Waveform	14.81	262	14.71	145	14.90

→ 18 datasets : at least one sub-forest significantly better than the forest

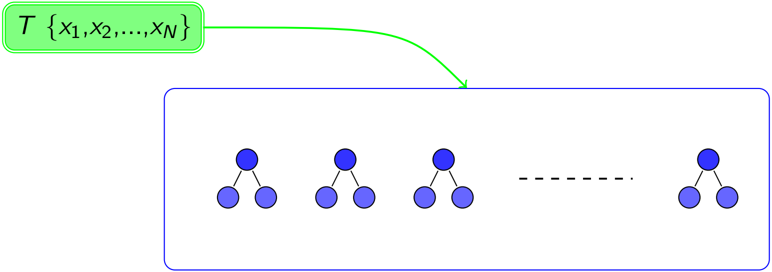
→ Sometimes only 10% of the trees can reach the performance of the forest

Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

→ New tree grown to suit the best possible to the current sub-forest.



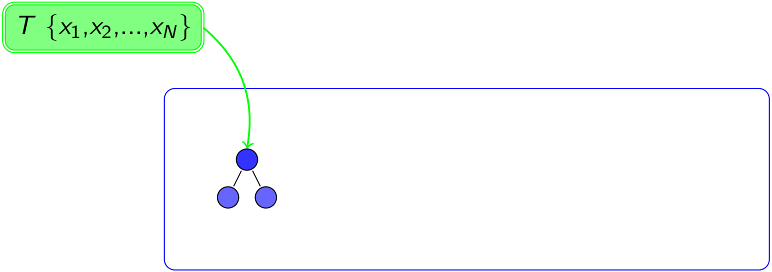
1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

$\rightarrow$ New tree grown to suit the best possible to the current sub-forest.



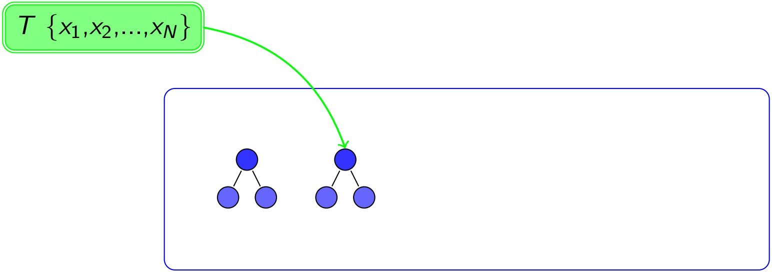
1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

$\rightarrow$ New tree grown to suit the best possible to the current sub-forest.



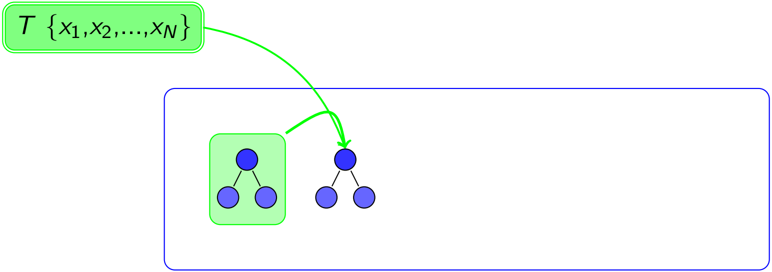
1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

$\rightarrow$ New tree grown to suit the best possible to the current sub-forest.



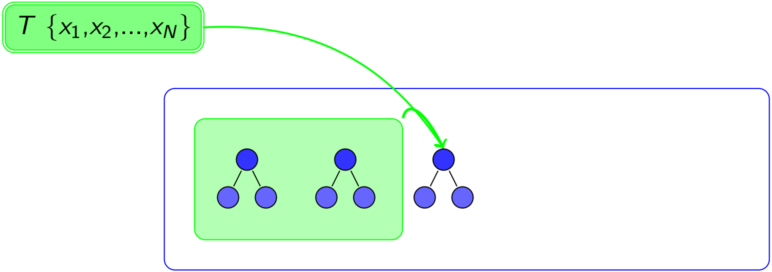
1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

→ New tree grown to suit the best possible to the current sub-forest.



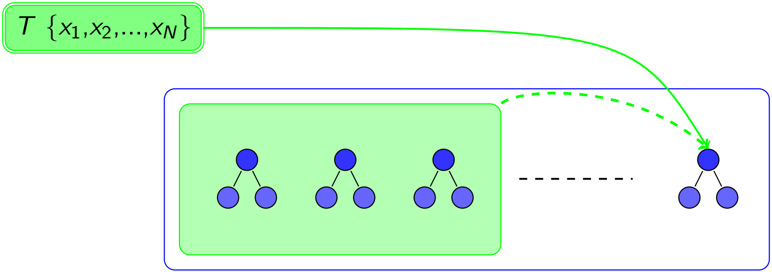
1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

→ New tree grown to suit the best possible to the current sub-forest.



1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

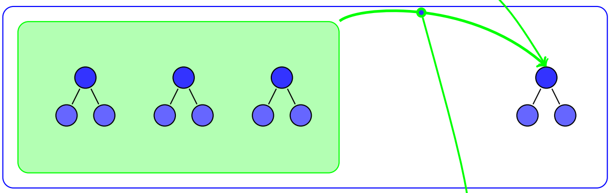
Dynamic Random Forest (DRF) [1]

Principe

Key idea : guide the tree induction

→ New tree grown to suit the best possible to the current sub-forest.

$T \; \{x_1, x_2, \ldots, x_N\}$



Weight the training instances :

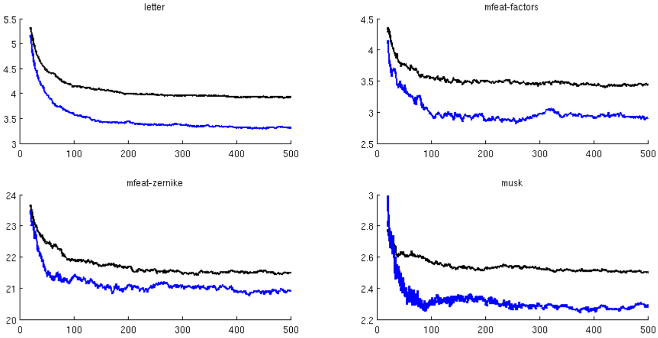
→ Increase weights of wrongly classified instances

→ Criterion : proportion of correct tree predictions

1. Bernard et al. *Dynamic Random Forests*, Pattern Recognition Letters, 2012

Dynamic Random Forest

Evaluation



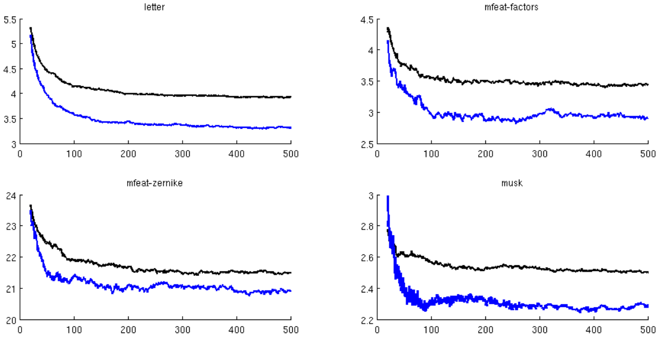
DRF

Forest-RK

Test error rate w.r.t. the number of trees

1. 500-trees Forests :
   $\rightarrow$ DRF significantly better than Forest-RK for 14 datasets
   $\rightarrow$ DRF $>$ Forest-RK $>$ *Forest-RI/K**

Dynamic Random Forest
Evaluation



DRF

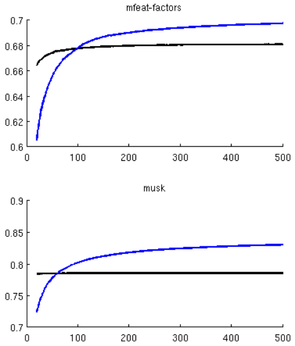
Forest-RK

Test error rate w.r.t. the number of trees

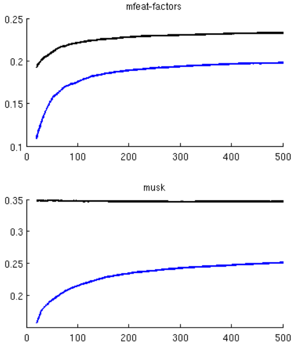
2. Generalization error convergence
   $\rightarrow$ Not mathematically proven anymore, but experimentally observed

Strength and Correlation in Dynamic Random Forest

The *strength* must be maximized and the *correlation* must be minimized (for minimizing $\frac{\overline{\rho}}{s^2}$)



Strength w.r.t. the number of trees



Correlation w.r.t. the number of trees

DRF, Forest-RK

Works in progress with Random Forests

In the chronological order...

1. **Dynamic Random Forest** : weighting the features to guide the Decision Tree induction
2. **One-Class Random Forest**[2] : Random Forests for One-Class classification
3. **Random Forests with Random Hierarchies** : Randomization principle for Hierarchical Multilabel classification
4. **Cost-Sensitive Random Forests** : Random Forests for Cost-Sensitive classification with multi-objective evolutionnary techniques

---

2. Désir et al. *One-Class Random Forests*, Pattern Recognition, 2013