Wavelets decomposition of Random Forest

Oren Elisha
Bernried 2017
Joint work with Prof Shai Dekel

Based on "Wavelet decompositions of Random Forests-smoothness analysis, sparse approximation and applications." *Journal of Machine Learning Research* 17.198 (2016): 1-38.

Motivation

- Improving machine learning algorithms using wavelets
- Main tasks
  - Prediction (classification, regression)
  - Feature importance
  - Model compression
- Domains
  - Image processing
  - Computer Vision
  - Ranking
  - NLP
  - Other

Example: Wine quality data set

$$x \in \Omega$$

$$f(x)$$

$$\{x_i, y_i\}_{i=1}^{m}$$

Wine ID	Alcohol	Acidity	…	Score
1	8	0.5	..	7
#…				

New sample

$$\hat{x}$$

Alcohol	Acidity	…
7	0.2	..

model

$$\tilde{f}(x)$$

Score $\hat{y}$

Why Random Forest

Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

- Evaluate 179 classifiers arising from 17 families (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest- neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods)
- Using open source models the are implemented in Weka, R, C and Matlab
- Use 121 data sets, which represent the whole UCI data base (excluding the large-scale problems)

The classifiers most likely to be the bests are the random forest (RF) versions, the best of which (implemented in R and accessed via caret) achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets.

Decision Trees

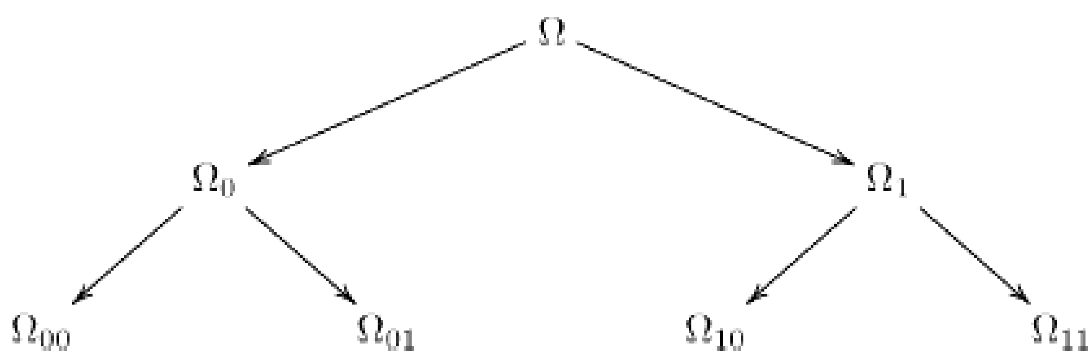
In the functional setting we are given a function

$$f \in L_2(\Omega), \quad \Omega \subset \mathbb{R}^n.$$

In applications, point values (or even "density")

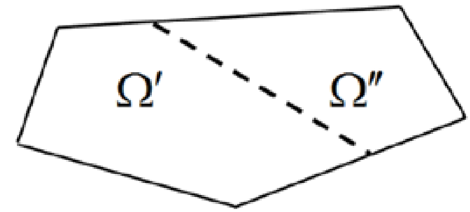
$$f(x_i), \quad x_i \in \Omega, \quad i \in I$$

We apply recursive subdivision of the data

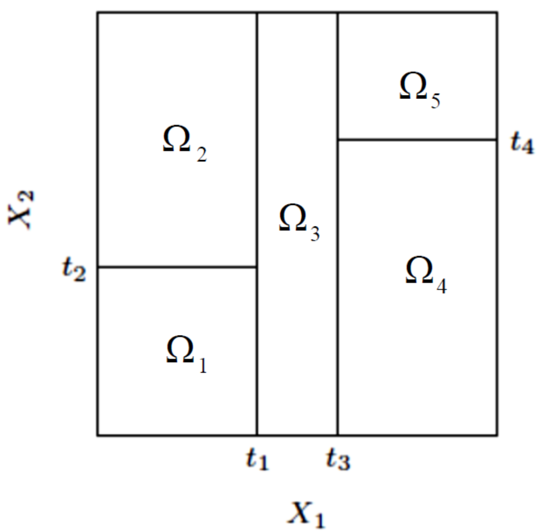


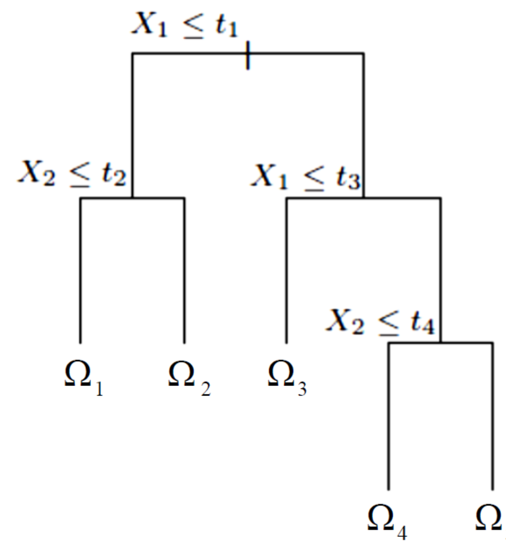
Decision Trees

Invoking a partition for each node Ω recursively, with low order Local Polynomials Q_Ω to minimize:



$$\sum_{x_i \in \Omega'} \left| f(x_i) - Q_{\Omega'} \right|^2 + \sum_{x_i \in \Omega''} \left| f(x_i) - Q_{\Omega''} \right|^2, \quad \Omega' \cup \Omega'' = \Omega$$

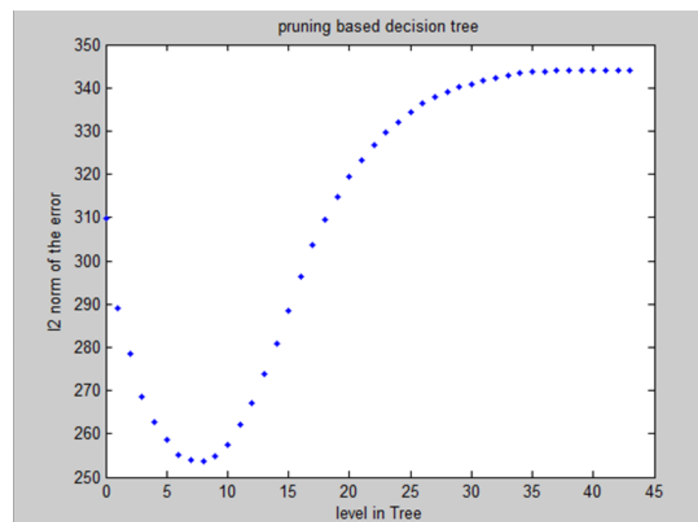




$$x = (x_1, \ldots, x_n) \quad \rightarrow \qquad \tilde{f}(x) := Q_{\Omega'}(x),$$

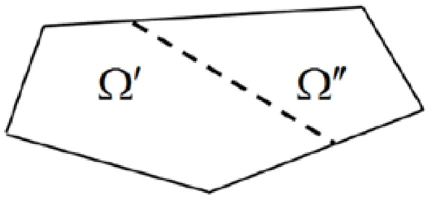
Some considerations for decision trees

- Impact of dimensionality
  - Curse of dimensionality (more samples are required for high dimensional data)
  - Computational Complexity (Approximation with lower degree of polynomials)
  - Restricted subdivisions (e.g. main axis only)
- Greedy nature of decision trees
  - Stopping criteria and Pruning (over-fitting)
  - Sensitivity to noise
  - Generalization error



Geometric Wavelets (Dekel and Leviatan, 2005)

Let Ω' be a child of Ω in a tree $\mathcal{T}$, $\quad \Omega' \dot{\subset} \Omega$



The Geometric Wavelet associated with Ω'

$$\psi_{\Omega'} := \psi_{\Omega'}(f) := \mathbf{1}_{\Omega'}(Q_{\Omega'} - Q_{\Omega})$$

$$f = \sum_{\Omega \in \mathcal{T}} \psi_{\Omega} \quad with \;\; \psi_{\Omega_0} := Q_{\Omega_0} := \min_{Q \in \Pi_r} \int_{\Omega_0} (f - Q)^2$$

With norms $$\|\psi_{\Omega'}\|_2^2 = \int_{\Omega'} (Q_{\Omega'}(x) - Q_{\Omega}(x))^2 \, dx,$$

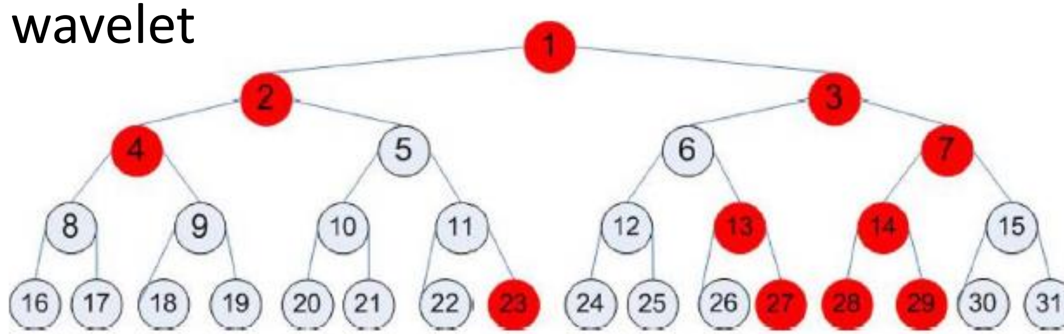
Or in the discrete case $$\|\psi_{\Omega'}\|_2^2 = \sum_{x_i \in \Omega'} |Q_{\Omega'}(x_i) - Q_{\Omega}(x_i)|^2,$$

To enable the Sorting: $$\left\|\psi_{\Omega_{k_1}}\right\|_2 \geq \left\|\psi_{\Omega_{k_2}}\right\|_2 \geq \left\|\psi_{\Omega_{k_3}}\right\|_2 \cdots$$

Geometric Wavelets

Adaptive M-term geometric wavelet sum

$$\tilde{f}_M(x) = \sum_{i=1}^{M} \psi_{\Omega_{k_i}}(x)$$



Classical Wavelets properties

- Details between low and high resolutions
- Multi-resolution representation
- Enables sparse representation for appropriate data.
- Vanishing moments $f_\Omega \in \Pi_r \Rightarrow Q_\Omega = Q_{\Omega'} = f_\Omega \;\; and \;\; \psi_{\Omega'} = 0$
- M- term representation (using wavelets norm)
- Correspondence with smoothness space

Distinctive properties

- Adaptive partitions creates non linearity (the decomposition depends on the function)
- No ortho basis

4096-term Bi-orthogonal Wavelets Approximation PSNR=29.22

FIG. 4.3. *Dyadic biorthogonal wavelet approximation of the "peppers" image with* $n = 4096$, *PSNR=29.22.*

2048-term Geometric Wavelets Approximation PSNR=31.32

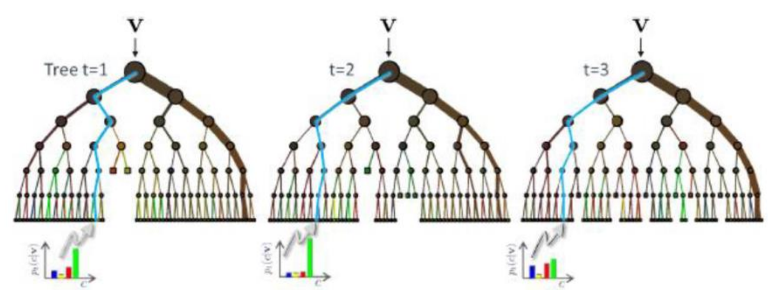
FIG. 4.2. *Geometric wavelet approximation of the "peppers" image with $n = 2048$, PSNR=31.32.*

Random forests

- 'Best' decision tree: NP-hard problem!
- Goal: overcome the 'greedy nature' of a single tree.
- 'Over each random subset we create a tree $\mathcal{T}_j$
- Diversity
  - Bagging': For each j, we select a random subset X^j consisting of 80% of the input data points.
  - For each tree Randomized attributes
  - Some methods cerates random splits.

So we have $$\tilde{f}(x) := \sum_j w_j \tilde{f}_j(x)$$

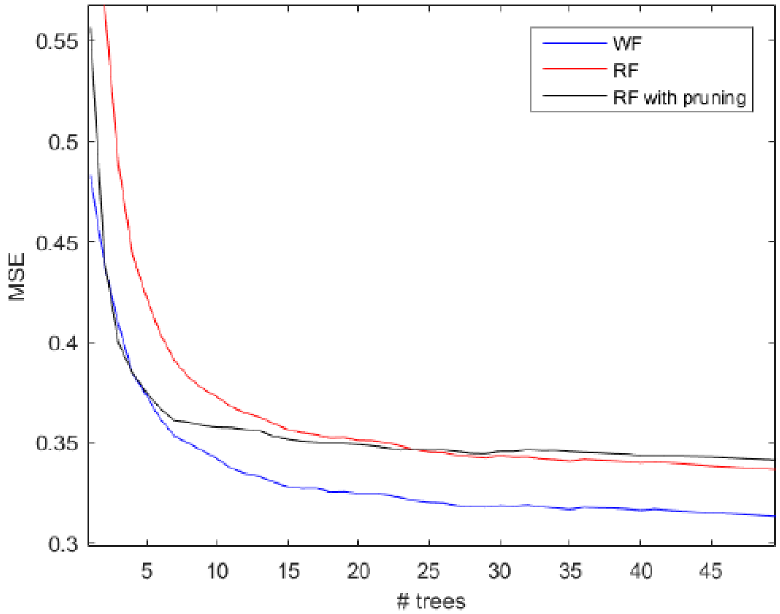
e.g. with $$w_j = 1/J$$



Source: Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.

Convergence of forest

- Leo Breiman, “random forests”, 2001:
  - For a large number of trees, it follows from the Strong Law of Large Numbers that as the number of trees increases, for almost surely, the generalization error of $\tilde{f}_J(x)$ converges
- This is the reason that random forests do not overfit as more trees are added.



Wavelet decomposition of a random forest

Create a wavelet decomposition of each tree in the random forest

$$\tilde{f}_j = \sum_{\Omega \in \mathcal{T}_j} \psi_\Omega, \quad j = 1, \ldots, N.$$

A wavelet representation of the entire random forest

$$\tilde{f}(x) = \sum_{j=1}^{N} \sum_{\Omega \in \mathcal{T}_j} w_j \psi_\Omega(x)$$

Order the wavelet components of the random forest by

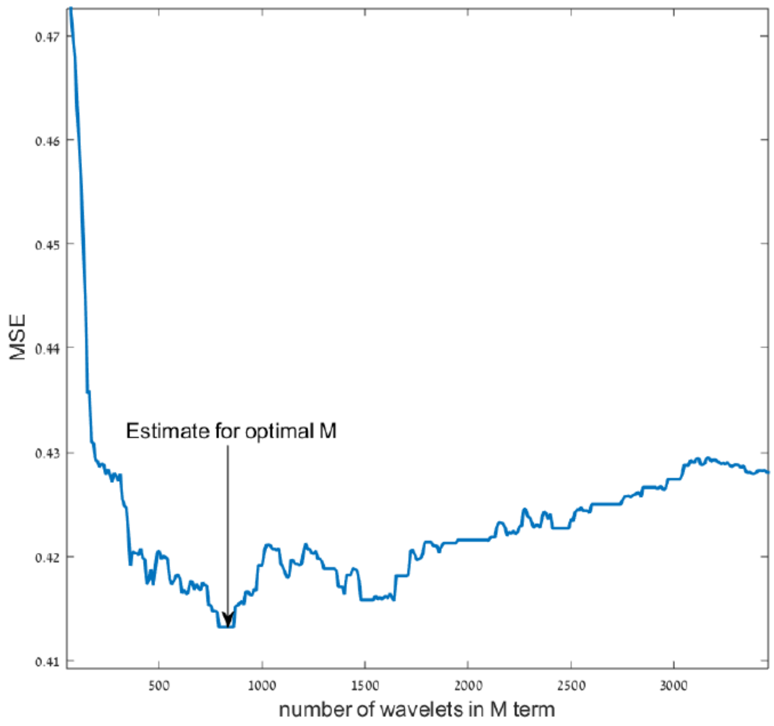
$$w_{j(\Omega_{k_1})} \left\| \psi_{\Omega_{k_1}} \right\|_2 \geq w_{j(\Omega_{k_2})} \left\| \psi_{\Omega_{k_2}} \right\|_2 \cdots$$

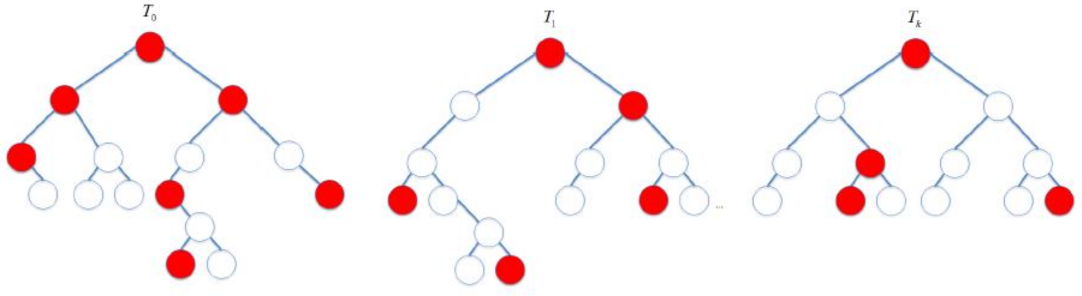
The M-term approximation of a random forest is

$$\tilde{f}_M(x) = \sum_{m=1}^{M} w_{j(\Omega_{k_m})} \psi_{\Omega_{k_m}}(x).$$

Wavelet decomposition of a random forest

$$f_M(x) := \sum_{m=1}^{M} w_{j(\Omega_{k_m})} \psi_{\Omega_{k_m}}(x).$$





Definitions

For a function $f \in L_\tau(\Omega_0)$, $0<\tau\le\infty$ $h\in\mathbb{R}^n$ and $r\in\mathbb{N}$, we recall the r^{th} order difference operator

$$\Delta_h^r(f,x) := \Delta_h^r(f,\Omega,x) := \begin{cases} \sum_{k=0}^{r}(-1)^{r+k}\binom{r}{k} f(x+kh) & [x,x+rh]\subset\Omega, \\ 0 & \text{otherwise,} \end{cases}$$

where $[x,y]$ denotes the line segment connecting any two points $x,y\in\mathbb{R}^n$. The ***modulus of smoothness of order r*** over Ω is defined by

$$\omega_r(f,t)_\tau := \sup_{|h|\le t}\left\|\Delta_h^r(f,\Omega,\cdot)\right\|_{L_\tau(\Omega)}, \qquad t>0,$$

where for $h\in\mathbb{R}^n$, $|h|$ denotes the norm of h. We also denote

$$\omega_r(f,\Omega)_\tau := \omega_r(f,diam(\Omega))_\tau .$$

For $0<p<\infty$ and $\alpha>0$, we set $\tau=\tau(\alpha,p)$, to be $1/\tau := \alpha+1/p$. For a given function $f\in L_p(\Omega_0)$, $\Omega_0\subset\mathbb{R}^n$ and tree $\mathcal{T}$, we define the associated B-space smoothness in $\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})$, $r\in\mathbb{N}$ by

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})} := \left(\sum_{\Omega\in\mathcal{T}}\left(|\Omega|^{-\alpha}\omega_r(f,\Omega)_\tau\right)^\tau\right)^{1/\tau},$$

where, $|E|$ denotes the volume of E. For a given forest $\mathcal{F}=\{\mathcal{T}_j\}_{j=1}^J$ and weights $w_j=1/J$, the α Besov semi-norm associated with the forest is

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} := \frac{1}{J}\left(\sum_{j=1}^J |f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T}_j)}^\tau\right)^{1/\tau}.$$

Jackson-type estimate

Let $\mathcal{F} = \{\mathcal{T}_j\}_{j=1}^{J}$ be a forest. Assume there exists a constant $0 < \rho < 1$, such that for any domain $\Omega \in \mathcal{F}$ on a level l and any domain $\Omega' \in \mathcal{F}$, on the level $l+1$, with $\Omega \bigcap \Omega' \neq \emptyset$, we have

$$|\Omega'| \leq \rho |\Omega|, \quad \text{where } |E| \text{ denotes the volume of } E \subset \mathbb{R}^n.$$

Denote formally $f = \sum_{\Omega \in \mathcal{F}} w_{j(\Omega)} \psi_\Omega$, and assume that $|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} < \infty$, where

$$\frac{1}{\tau} = \alpha + \frac{1}{p}.$$

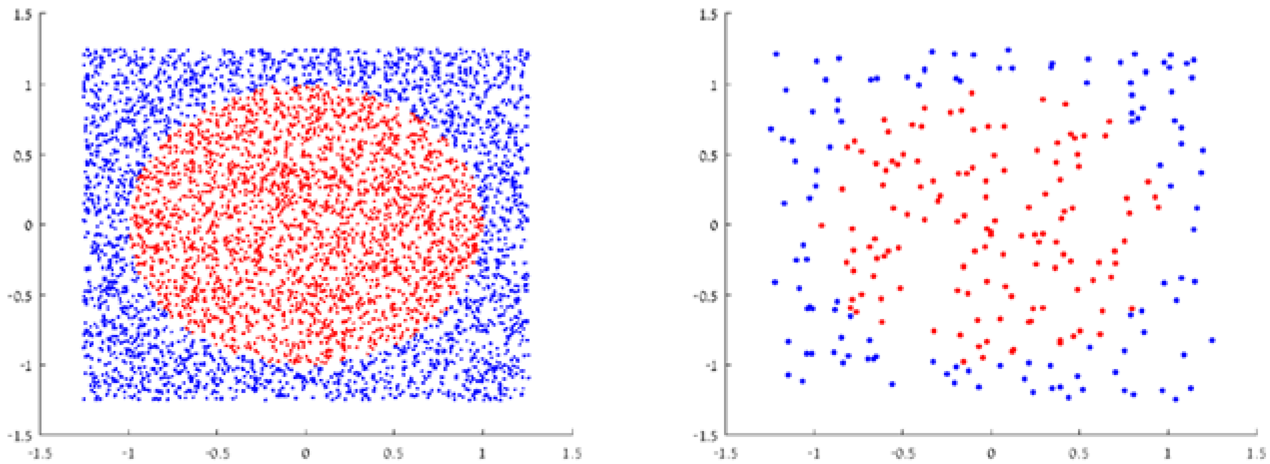
Then,

$$E_M := \|f - f_M\|_p \leq C(p, \alpha, J, \rho) M^{-\alpha} |f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})}.$$

Measuring the smoothness

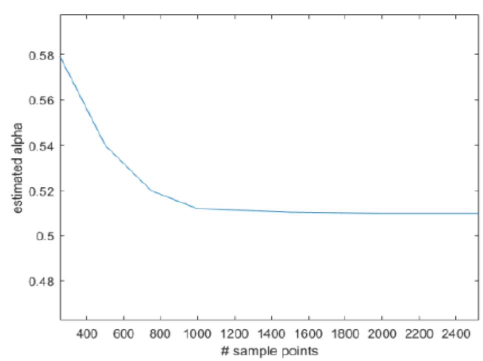
using $\int_1^M m^{-u} dm = \left(M^{1-u} - 1\right)/(1-u)$, we estimate α_j by

$$\min_{\alpha_j} \left| \frac{M^{1-\alpha_j} - 1}{1-\alpha_j} \sigma_{j,1} - \sum_{m=1}^{M-1} \sigma_{j,m} \right| .$$



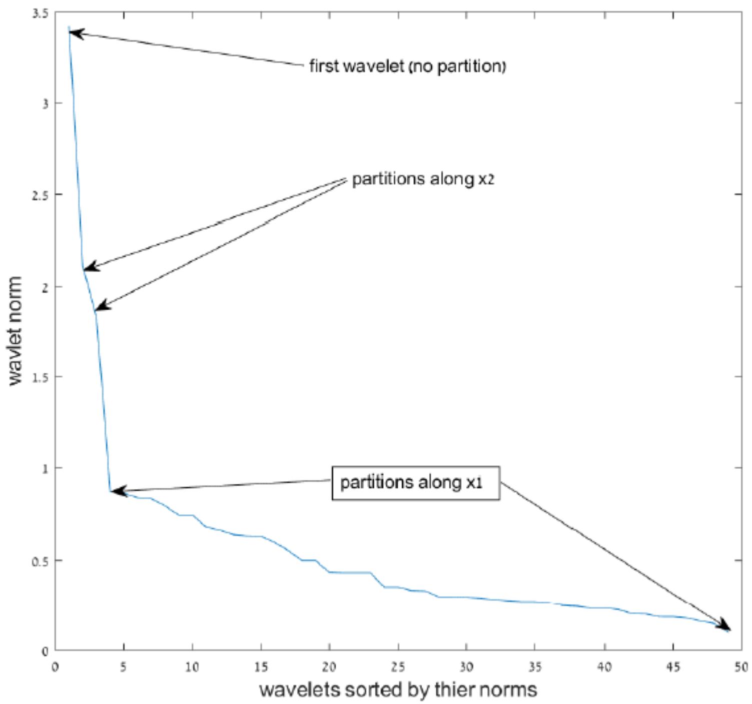
(a) 5,000 sampled points (b) 250 sampled points

Example $\alpha = 1/2$



Variable importance

$$S_i^{\mathcal{T}} := \frac{1}{J}\sum_{j=1}^{J}\sum_{\Omega\in\mathcal{T}_j\cap V_i}\|\psi_\Omega\|_2^{\tau}, \quad i = 1,\dots,n,$$

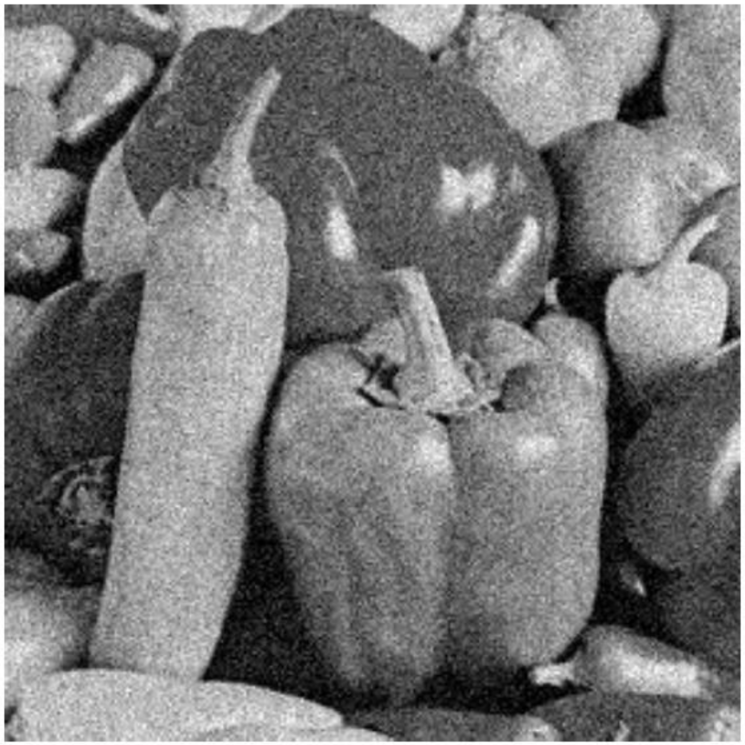


To demonstrate this problem, we follow the experiment suggested in (Strobl et. al. 2006). We set a number of samples to $m = 120$, where each sample has two explanatory independent variables: $x_1 \sim N(0,1)$ and $x_2 \sim Ber(0.5)$. A correlation between $y = f(x_1, x_2)$ and x_2 is established by:

$$y \sim \left\{ \begin{array}{ll} Ber(0.7), & x_2 = 0, \\ Ber(0.3), & x_2 = 1. \end{array} \right\} \tag{23}$$

Applications and empirical results

16 trees - 21734 significant wavelets



(a) Image with noise , 256×256,

PSNR=22.22

(b) Denoised image, 256×256,

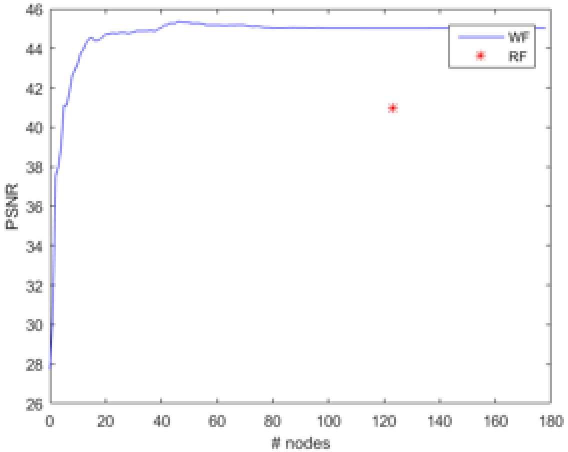
PSNR=30.7

Figure 5.1 Image denoising. "Peppers" . $\sigma = 20$

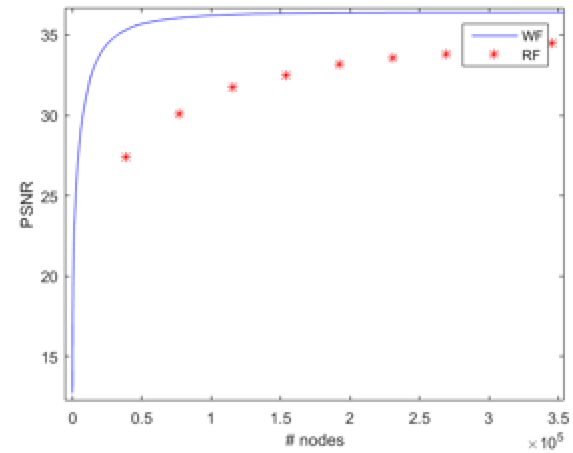
Compression

Task regression (R) classification (C)		Pruning Min-D [59]		Pruning Mean-D [59]		Wavelets – 90% error saturation			α
		#trees	#nodes	#trees	#nodes	#trees	#prediction nodes	#overall nodes	
C	Record linkage	1	123	1	123	1	3	6	0.99
R	CT Slice*	2	77042	2	76396	2	4212	5141	0.51
C	Titanic	3	711	10	2248	1	19	34	0.42
C	Balanced scale	1	185	1	185	1	40	55	0.34
R	Concrete	19	2297	8	966	3	54	64	0.32
C	Magic Gamma	9	26793	5	14961	3	823	1657	0.25
R	Airfoil	5	4533	3	7487	3	1734	1929	0.23
R	California Housing	4	65436	9	149863	4	5469	7292	0.2
C	EEG	7	17845	11	28355	6	9489	12808	0.15
R	Parkinson	18	103822	19	110187	12	19110	20947	0.11
R	Wine quality	14	39350	13	36439	12	21615	29089	0.07
R	Year Prediction	21	10657799	24	12201588	19	9296363	9300284	0.02

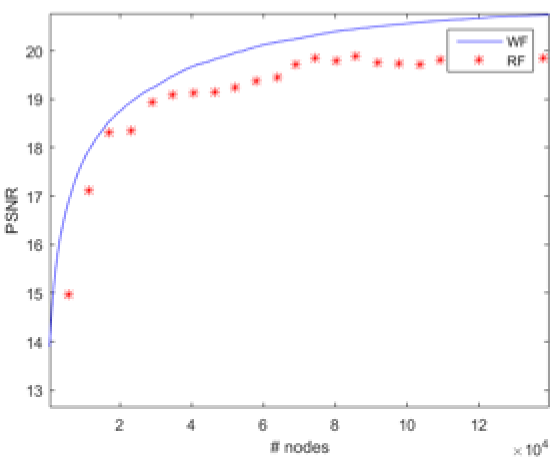
Compression



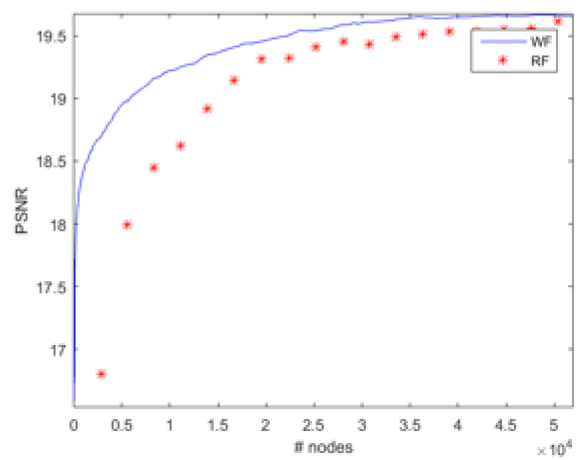
(a) Record linkage dataset ($\alpha = 0.99$)



(b) CT Slice dataset ($\alpha = 0.51$)

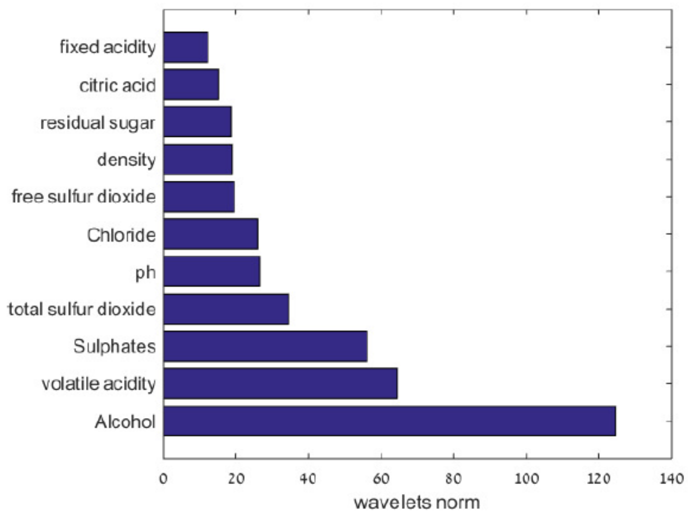


(a) Parkinson dataset ($\alpha = 0.11$)

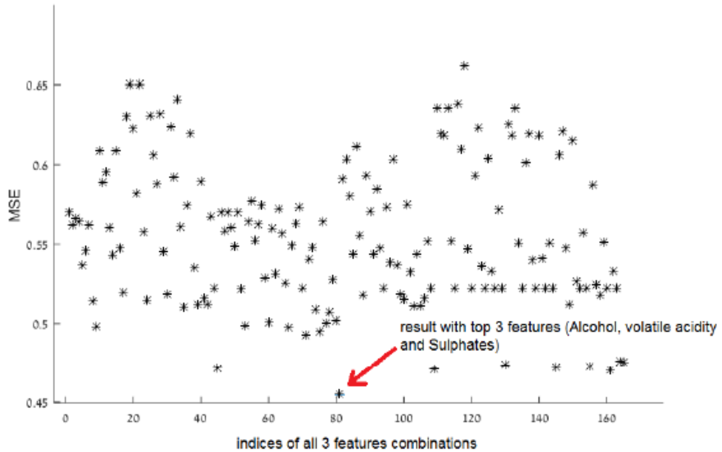


(b) Wine quality dataset ($\alpha = 0.7$)

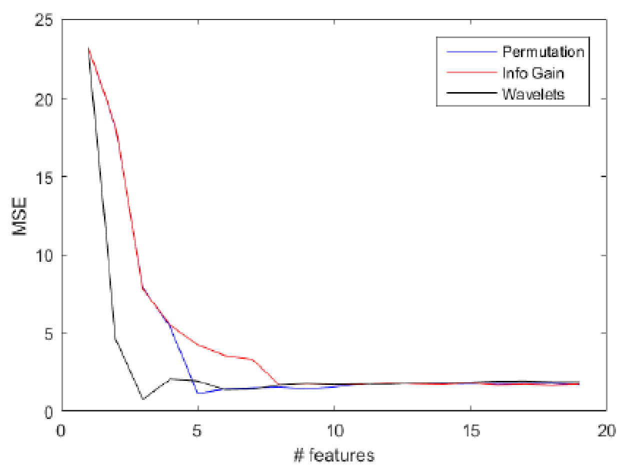
Variable importance



(a) Wavelet-based feature importance histogram

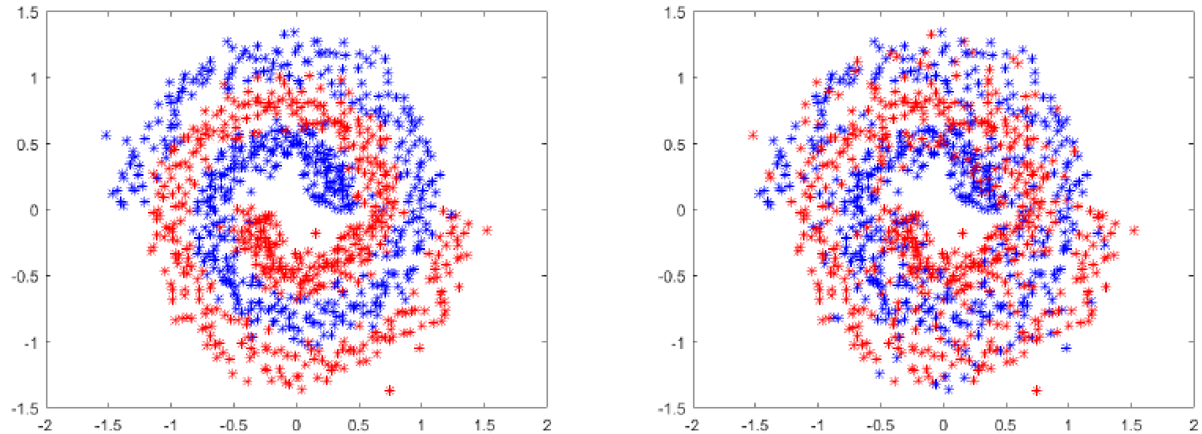


(b) Error of RFs constructed over all possible 3 feature subsets



(a) "Parkinson" dataset, $\epsilon = 1.74$.

Overcoming mislabeling in prediction



(a) Original set (b) Set with amplified mis-labeling

Figure 12: 'Spirals' dataset (Spiral dataset)

approach is more significant in the second case with more 'false labeling' in the training set.

Table 2: 'Spirals' dataset - Classification results.

	Wavelet error	RF error	Pruned RF error
Original spiral set	$12.2 \pm 0.9\%$	$14.4 \pm 1.1\%$	$15.9 \pm 0.8\%$
Set with amplified mis-labeling	$13.9 \pm 1.2\%$	$17.8 \pm 1.3\%$	$22.7 \pm 1.6\%$

Thank you