Wavelets decomposition of Random Forest

Oren Elisha
Bernried 2017
Joint work with Prof Shai Dekel

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

\[ x \in \Omega \]

\[ f(x) \]

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

<table>
<thead>
<tr>
<th>Wine ID</th>
<th>Alcohol</th>
<th>Acidity</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>0.5</td>
<td>7</td>
</tr>
</tbody>
</table>

#...

New sample

<table>
<thead>
<tr>
<th>Alcohol</th>
<th>Acidity</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0.2</td>
<td>..</td>
</tr>
</tbody>
</table>

\( \hat{x} \)

model

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

\( \hat{y} \)
Why Random Forest

- 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 that 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 $\Omega$ recursively, with low order Local Polynomials $Q_\Omega$ to minimize:

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

$x = (x_1, \ldots, x_n) \rightarrow \tilde{f}(x) := Q_{\Omega'}(x)$,
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 $\Omega'$ be a child of $\Omega$ in a tree $T$, $\Omega' \subset \Omega$

The Geometric Wavelet associated with $\Omega'$

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

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

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

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:

$$\| \psi_{\Omega_{k_1}} \|_2 \geq \| \psi_{\Omega_{k_2}} \|_2 \geq \| \psi_{\Omega_{k_3}} \|_2 \cdots$$
Geometric Wavelets

Adaptive $M$-term geometric wavelet sum

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

Classical Wavelets properties

- Details between low and high resolutions
- Multi-resolution representation
- Enables sparse representation for appropriate data.
- Vanishing moments $f_{\Omega} \in \prod_r \Rightarrow Q_{\Omega} = Q_{\Omega} = f_{\Omega}$ 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 creates random splits.

So we have $\tilde{f}(x) := \sum_j w_j \tilde{f}_j(x)$
e.g. with $w_j = 1 / J$

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 $\hat{f}_j(x)$ 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

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

A wavelet representation of the entire random forest

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

Order the wavelet components of the random forest by

$$w_j(\Omega_{k_1}) \| \psi_{\Omega_{k_1}} \|_2 \geq w_j(\Omega_{k_2}) \| \psi_{\Omega_{k_2}} \|_2 \cdots$$

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

\[ f_M(x) := \sum_{m=1}^{M} w_j(\Omega_{km}) \psi_{\Omega_{km}}(x). \]
For a function $f \in L_{r} (\Omega_{0})$, $0 < \tau \leq \infty$, $h \in \mathbb{R}^{n}$ and $r \in \mathbb{N}$, we recall the $r^{th}$ order difference operator

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

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** $\Omega$ **is defined by**

$$
\omega_{r} (f, t) := \sup_{|\xi| \leq t} \| \Delta^{r}_{\xi} (f, \Omega, \cdot) \|_{L_{r} (\Omega)}, \quad t > 0,
$$

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

$$
\omega_{r} (f, \Omega) := \omega_{r} (f, \text{diam}(\Omega)).
$$

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

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

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 $\alpha$ Besov semi-norm associated with the forest is

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

Let $\mathcal{F} = \{ T_{j} \}_{j=1}^{J}$ 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 $\Omega \cap \Omega' \neq \emptyset$, we have

$$|\Omega'| \leq \rho |\Omega|,$$

where $|E|$ 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 $|f|_{\mathcal{B}^{2,\tau}(\mathcal{F})} < \infty$, where

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

Then,

$$E_{M} := \| f - f_{M} \|_{p} \leq C(p, \alpha, J, \rho) M^{-\alpha} |f|_{\mathcal{B}^{2,\tau}(\mathcal{F})}.$$
Measuring the smoothness

Using \( \int_1^M m^{-u} dm = (M^{1-u} - 1)/(1 - u) \), we estimate \( \alpha_j \) by

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

Example \( \alpha = 1/2 \)
Variable importance

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

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:

\[
y \sim \begin{cases} 
    Ber(0.7), & x_2 = 0, \\
    Ber(0.3), & x_2 = 1. 
\end{cases}
\]  \hspace{1cm} (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". σ = 20
## Compression

<table>
<thead>
<tr>
<th>Task regression (R) classification (C)</th>
<th>Pruning Min-D [59]</th>
<th>Pruning Mean-D [59]</th>
<th>Wavelets – 90% error saturation</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#trees</td>
<td>#nodes</td>
<td>#trees</td>
<td>#nodes</td>
</tr>
<tr>
<td>C Record linkage</td>
<td>1</td>
<td>123</td>
<td>1</td>
<td>123</td>
</tr>
<tr>
<td>R CT Slice*</td>
<td>2</td>
<td>77042</td>
<td>2</td>
<td>76396</td>
</tr>
<tr>
<td>C Titanic</td>
<td>3</td>
<td>711</td>
<td>10</td>
<td>2248</td>
</tr>
<tr>
<td>C Balanced scale</td>
<td>1</td>
<td>185</td>
<td>1</td>
<td>185</td>
</tr>
<tr>
<td>R Concrete</td>
<td>19</td>
<td>2297</td>
<td>8</td>
<td>966</td>
</tr>
<tr>
<td>C Magic Gamma</td>
<td>9</td>
<td>26793</td>
<td>5</td>
<td>14961</td>
</tr>
<tr>
<td>R Airfoil</td>
<td>5</td>
<td>4533</td>
<td>3</td>
<td>7487</td>
</tr>
<tr>
<td>R California Housing</td>
<td>4</td>
<td>65436</td>
<td>9</td>
<td>149863</td>
</tr>
<tr>
<td>C EEG</td>
<td>7</td>
<td>17845</td>
<td>11</td>
<td>28355</td>
</tr>
<tr>
<td>R Parkinson</td>
<td>18</td>
<td>103822</td>
<td>19</td>
<td>110187</td>
</tr>
<tr>
<td>R Wine quality</td>
<td>14</td>
<td>39350</td>
<td>13</td>
<td>36439</td>
</tr>
<tr>
<td>R Year Prediction</td>
<td>21</td>
<td>1065779</td>
<td>24</td>
<td>1220158</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>Wavelet error</th>
<th>RF error</th>
<th>Pruned RF error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original spiral set</td>
<td>12.2 ± 0.9%</td>
<td>14.4 ± 1.1%</td>
<td>15.9 ± 0.8%</td>
</tr>
<tr>
<td>Set with amplified mis-labeling</td>
<td>13.9 ± 1.2%</td>
<td>17.8 ± 1.3%</td>
<td>22.7 ± 1.6%</td>
</tr>
</tbody>
</table>
Thank you