Accelerating AdaBoost using multi-armed bandits

Róbert Busa-Fekete and Balázs Kégl
{busarobi,balazs.kegl}@gmail.com

Motivation

AdaBoost.MH with trees or products is state-of-the-art but slow if the number of features is large.

Idea: learn the “utility” of features using multi-armed bandits (MAB) → the base learner can be aided in finding good base classifiers fast.

**AdaBoost.MH**

\[
\text{AdaBoost.MH}(X, Y, W^{(l)}, \text{BASE}(\cdot, \cdot), T)
\]

1. for \( t = 1 \) to \( T \)
2. \( h^{(l)}(\cdot) \leftarrow \frac{d^{(l)}(\cdot)}{d^{(l)}(\cdot) + \gamma^{(l)}(\cdot)} \leftarrow \text{BASE}(X, Y, W^{(l)}) \)
3. for \( i = 1 \) to \( n \) for \( \ell = 1 \) to \( K \)
4. \( w^{(t)}_{i, \ell} \leftarrow w^{(t)}_{i, \ell} \frac{\exp(-h^{(t)}(x)_i)}{\sum_{\ell'=1}^{K} \sum_{j=1}^{n} w^{(t)}_{j, \ell'} \exp(-h^{(t)}(x)_j)} \)
5. return \( f(T)(\cdot) = \sum_{t=1}^{T} h^{(t)}(\cdot) \)

Base learners

The goal of the base learner is to maximize the edge

\[
\gamma = \sum_{i=1}^{n} w_i h(x)_i y_i.
\]

**Decision Stump**:

\[
\varphi_{j,b}(x) = \begin{cases} 
1 & \text{if } x^{(j)} \geq b, \\
-1 & \text{otherwise.}
\end{cases}
\]

**Product**:

\[ h(x) = \prod_{j=1}^{m} \varphi_{j,b_j}(x) \]

**Decision Tree**:

\[
\varphi_{j_1,b_1}(x), \varphi_{j_2,b_2}(x), \ldots \]

Multi-armed bandits

Each feature corresponds to an arm.

Each time a feature is selected, its reward

\[
r_{j}^{(t)} = 1 - \sqrt{1 - \frac{2}{T}}
\]

is recorded.

We select \( k \) features that maximize the sum of the reward and a confidence interval term

\[
c_j^{(t)} = \sqrt{\frac{2 \ln t}{T}}.
\]

Validation curves

Running time in minutes

<table>
<thead>
<tr>
<th>Base learner</th>
<th>Stump</th>
<th>Product</th>
<th>Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running time</td>
<td>274</td>
<td>456</td>
<td>384</td>
</tr>
</tbody>
</table>

Results

Hyperparameters:
- Number of terms in PRODUCT: \( m = 3 \)
- Number of leaves in DECISION TREE: \( N = 8 \)
- Number of features chosen by UCB(\( k \)): \( k = 50 \)
- Number of iterations (validated): \( T = 8000 \)

<table>
<thead>
<tr>
<th>learner \ data set</th>
<th>Churn (AUC)</th>
<th>Appetency (B. Acc.)</th>
<th>Upselling (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validation</strong></td>
<td><strong>Evaluation</strong></td>
<td><strong>Validation</strong></td>
<td><strong>Evaluation</strong></td>
</tr>
<tr>
<td>Stump</td>
<td>0.7424</td>
<td>0.6633</td>
<td>0.7051</td>
</tr>
<tr>
<td>Product</td>
<td>0.6702</td>
<td>0.6314</td>
<td>0.6999</td>
</tr>
<tr>
<td>Tree</td>
<td>0.7088</td>
<td>0.6819</td>
<td>0.7424</td>
</tr>
<tr>
<td>Winner</td>
<td>–</td>
<td>0.7611</td>
<td>–</td>
</tr>
</tbody>
</table>