FastXML: A Fast, Accurate and Stable Tree-classifier for eXtreme Multi-label Learning

Yashoteja Prabhu (IIT Delhi)  Manik Varma (Microsoft Research)

Extreme Multi-Label Learning

- Learning with millions of labels
- New paradigm for ranking and recommendation

\[ f: X \rightarrow 2^Y \]

X: Users  Y: Items

Our Contributions

- We formulate a novel node partitioning objective which directly optimizes an nDCG based ranking loss.
- We propose an efficient optimization algorithm for the novel formulation.
- We train on 4M points, 1M categories and 160K features data set in 8 hours on a single core desktop.
- We improve prediction accuracy significantly over the state of the art Extreme Multi-Label learning algorithms.

Prediction in Logarithmic Time

FastXML

Overview

Formulation

\[
\min_{\mathbf{w}, \mathbf{r}, i} \| \mathbf{w} \|_1 + \sum_i C_i (\delta_i) \log(1 + e^{-\delta_i \mathbf{w}^T \mathbf{x}_i}) + C_r \sum_i L_i^\pm (\delta_i, \mathbf{r}^\pm)
\]

where

\[
L_i^\pm (\delta_i, \mathbf{r}^\pm) = -\frac{1 \pm \delta_i}{2} \sum_j \log \left( 1 \pm \frac{N_j y_{ij}}{1 + \mathbf{r}_j^\pm} \right)
\]

Alternating Minimization

- Initialize \( \mathbf{w} \) to \( 0 \) and \( \delta_i \) to \( +1 \) or \(-1 \) at random
- Repeat until convergence
  - \( \mathbf{r}^\pm = \text{rank} \left( \sum_i \delta_i \mp 1 \right) N_j y_{ij} \)
  - \( \delta_i = \text{sign} \left( \mathbf{r}_i^\pm - v_i^\pm \right) \)
  - \( v_i^\pm = C_i (\pm 1) \log(1 + e^{\mp \mathbf{w}^T \mathbf{x}_i}) - C_r \sum_j \left( \frac{N_j y_{ij}}{1 + \mathbf{r}_j^\pm} \right) \)
- \( \mathbf{w} = \text{Solve} \left( \ell^1 \text{ Logistic Regression} \left( \{ \mathbf{x}_i, \delta_i \} \right) \right) \)

Results – Small Data Sets

Results – Large Data Sets

<table>
<thead>
<tr>
<th>Data Set Statistics</th>
<th># of Trees (M)</th>
<th># of Test Points (M)</th>
<th># of Dimensions (M)</th>
<th># of Labels (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>1.89</td>
<td>0.47</td>
<td>1.62</td>
<td>0.33</td>
</tr>
<tr>
<td>Ads-430K</td>
<td>1.12</td>
<td>0.50</td>
<td>0.088</td>
<td>0.43</td>
</tr>
<tr>
<td>Ads-1M</td>
<td>3.92</td>
<td>1.56</td>
<td>0.16</td>
<td>1.08</td>
</tr>
</tbody>
</table>

Precision at K

Precision vs No. of Trees

Training Time (Hr)

Variants of FastXML

Random Tree Selection

Speedup vs Cores (Ads-1M)