Extreme Classification with Large-scale Structured Learning

Yiming Yang, Carnegie Mellon University
Joint work with Siddharth Gopal (now Google)

Outline

• Introduction
• Related Work
• Proposed Approach
• Evaluation Plan
• Q/A
Problem Statements

- **Classification**: Automated assignment of category labels to objects (web pages, news articles, images, movies, music, shopping items, etc.)
- **Extreme Classification**: Very large number of categories (e.g., 500k for Wikipedia articles)
- **Structured Learning**: Using the dependency structures (hierarchies or graphs) among categories to *jointly optimize* the classification models.

10/6/2016

Yiming Yang, CPP Example
4.45 million articles in English Wikipedia, ~500,000 categories

Benchmark Evaluation Data Sets (Examples)

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Data Type</th>
<th>#Trn Classes</th>
<th>#Leaf Classes</th>
<th>Features</th>
<th>#TrnDocs</th>
<th>#TstDocs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEWS20</td>
<td>news stories</td>
<td>20</td>
<td>20</td>
<td>53,975</td>
<td>11,260</td>
<td>7,505</td>
</tr>
<tr>
<td>CLEF</td>
<td>X-ray images</td>
<td>63</td>
<td>63</td>
<td>89</td>
<td>10,000</td>
<td>1,006</td>
</tr>
<tr>
<td>RCV1 (Topics)</td>
<td>news stories</td>
<td>137</td>
<td>101</td>
<td>48,734</td>
<td>23,149</td>
<td>784,446</td>
</tr>
<tr>
<td>IPC</td>
<td>patents</td>
<td>552</td>
<td>451</td>
<td>541,869</td>
<td>46,324</td>
<td>28,926</td>
</tr>
<tr>
<td>LSHTC-small</td>
<td>web pages</td>
<td>1,563</td>
<td>1,139</td>
<td>51,033</td>
<td>4,463</td>
<td>1,858</td>
</tr>
<tr>
<td>DMOZ 2010</td>
<td>web pages</td>
<td>15,358</td>
<td>12,294</td>
<td>381,580</td>
<td>128,710</td>
<td>34,880</td>
</tr>
<tr>
<td>DMOZ 2011</td>
<td>web pages</td>
<td>35,448</td>
<td>27,875</td>
<td>348,548</td>
<td>383,408</td>
<td>103,435</td>
</tr>
<tr>
<td>DMOZ 2012</td>
<td>web pages</td>
<td>13,347</td>
<td>11,947</td>
<td>594,158</td>
<td>394,756</td>
<td>104,263</td>
</tr>
<tr>
<td>SWIKI 2011</td>
<td>Wikipedia</td>
<td>50,312</td>
<td>36,504</td>
<td>346,299</td>
<td>456,886</td>
<td>81,262</td>
</tr>
<tr>
<td>LWIKI 2011</td>
<td>Wikipedia</td>
<td>478,020</td>
<td>325,056</td>
<td>1,617,899</td>
<td>2,365,436</td>
<td>452,167</td>
</tr>
</tbody>
</table>

The PASCAL LSHTC Challenge

# of model parameters: 1,617,899 x 478,020 = 800 million
Large Taxonomies → Skewed Category Distribution (using DMOZ 2010 as an example)

Majority of classes are rare

Data-Sparse Challenge: We cannot ignore 99% of the categories.

Why Structured Learning?

– Leveraging the hierarchies or networked relations among categories to help the training of classifiers with sparse data

– In other words, we want to “borrow” data from the label-rich categories for the training of label-poor categories
Outline

• Introduction
• Related Work
• Proposed Approach
• Evaluation Plan
• Q/A

Related Work (Representative Methods)

1) Flat Classification Models (simplest, dominating)
2) Top-down Classification (earlier efforts)
3) Hierarchical Large-margin Models (later approaches)
Related 1: Flat Classification

- Train one binary classifier (e.g., SVM or LR) per category
- Pros: very fast
  - We used 256 processors in CMU clusters to train 325,056 SVM’s with 1.6 millions features for Wikipedia categories in LWIKI
  - 19 hours of training time, which means an average of 54 sec/classifier on a single machine
  - 5 to 10 minutes of the testing time over 452,167 doc’s, i.e., an average of 0.34 sec/doc on a single machine
- Cons: Ignore the dependencies among categories
  - The system may assign both “person” and “car” to the same object although we know that those labels are mutually exclusive.

Related Work 2: Top-Down Classification

[Dumais & Chen, 2000] [Yang et al., 2003] [Lie et al., 2005]

- Partition the training data into local subsets along the hierarchy;
- Train classifiers independently on all nodes on local subsets of data;
- Apply the classifiers to each test instance in a top-down fashion;
- Pros: Computationally efficient in both training and testing phases
- Cons: Once an error is made at a higher level, it will be wrong at all the lower levels.
Related Work 3: Hierarchical modeling with large-margin classifiers

[Tsochantridis et al., 2006], [Cai & Hofmann, 2004], [Rousu et al., 2006],
[Dekel et al., 2004], [Cesa-Bianchi et al., 2006]

- Jointly optimizing large-margin classifiers with a tree loss function
  
  \[
  \min_{w, \iota} \frac{1}{2} \sum_{j \in \text{tree}} \|w_j\|^2 + C \sum_{m=1}^{M} \xi_m
  \]

  Subject to:
  
  \( \forall (r,s) : \pi(r) = \pi(s) , y_r, y_s \in \text{SubTree} \Rightarrow f_r(x_r) \geq f_s(x_s) + 1 - \xi, \)
  
  \( \xi \geq 0 \quad \forall i = 1, \ldots, m \text{(\#training pairs)} \)

- **Pros**: leveraging hierarchical structures in joint optimization of classifiers
- **Cons**: not scale well

Outline

- Introduction
- Related Work
- **Proposed Approach**
- Evaluation Plan
- Q/A
Recursive Regularization of Classifiers
(Gopal & Yang, KDD 2013, [Best student paper runner up] )

• leverage both hierarchical and graphical dependencies among categories
• Allow various types of classifiers (SVM, LR, NNet, etc.) to be modeled in a unified framework
• Computationally scalable to extremely large problems (with 600k+ classifiers)

Expected Risk Minimization in a Classifier

\[ \hat{w} = \arg \min_w \lambda(w) + C \times R_{\text{emp}}(w, D_{\text{train}}) \]

Regularization term \hspace{1cm} Empirical risk

standard SVM: \[ \min_w \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} (1 - y_i w^T x_i)_+ \] Hinge Loss

RR-SVM: \[ \min_w \lambda(W) + C \sum_{n \in T} \sum_{i=1}^{N} (1 - y_{in} w_n^T x_i)_+ \]

\[ W = \{ w_1, \cdots, w_{|T|} \} \]
Two Kinds of Dependency Structure (H and G)

Key Idea: Structured Regularization

Structured Regularization

- Incorporate a hierarchy (H) into regularization
  \[ \lambda_H(W) = \sum_{n \in H} \| w_n - w_{\pi(n)} \|^2 \]
- Incorporate a graph \( G=(V, E) \) into regularization
  \[ \lambda_G(W) = \sum_{(i,j) \in E} \| w_i - w_j \|^2 \]
- Intuition: Related categories should have similar w’s.
Inference: Iterative Process

- Starting from arbitrary \( w \) at each node;
- Pick one node at a time to optimize its \( w \) while fixing the models (\( w \)'s) on all other nodes;
- Repeat the above step until all the nodes have a stable (and optimal) \( w \)'s.

Parallel Computing with Divide-\&-Conquer

- **Hierarchies**
  - Optimize odd and even levels alternately

- **Graphs**: First find a graph vertex coloring, and then
  - Pick a color
  - In parallel, optimize all nodes with that color
  - Repeat with a different color
Outline

- Introduction
- Related Work
- Proposed Approach
- Evaluation Plan
- Q/A

10 Benchmark Data Sets

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Data Type</th>
<th>#Trn Classes</th>
<th>#Leaf Classes</th>
<th>#Features</th>
<th>#TrnDocs</th>
<th>#TstDocs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEWS20</td>
<td>news stories</td>
<td>20</td>
<td>20</td>
<td>53,975</td>
<td>11,260</td>
<td>7,505</td>
</tr>
<tr>
<td>CLEF</td>
<td>X-ray images</td>
<td>63</td>
<td>63</td>
<td>89</td>
<td>10,000</td>
<td>1,006</td>
</tr>
<tr>
<td>RCV1 (Topics)</td>
<td>news stories</td>
<td>137</td>
<td>101</td>
<td>48,734</td>
<td>23,149</td>
<td>784,446</td>
</tr>
<tr>
<td>IPC</td>
<td>patents</td>
<td>552</td>
<td>451</td>
<td>541,869</td>
<td>46,324</td>
<td>28,926</td>
</tr>
<tr>
<td>LSHTC-small</td>
<td>web pages</td>
<td>1,563</td>
<td>1,139</td>
<td>51,033</td>
<td>4,463</td>
<td>1,858</td>
</tr>
<tr>
<td>DMOZ 2010</td>
<td>web pages</td>
<td>15,358</td>
<td>12,294</td>
<td>381,580</td>
<td>128,710</td>
<td>34,880</td>
</tr>
<tr>
<td>DMOZ 2011</td>
<td>web pages</td>
<td>35,448</td>
<td>27,875</td>
<td>348,548</td>
<td>383,408</td>
<td>103,435</td>
</tr>
<tr>
<td>DMOZ 2012</td>
<td>web pages</td>
<td>13,347</td>
<td>11,947</td>
<td>594,158</td>
<td>394,756</td>
<td>104,263</td>
</tr>
<tr>
<td>SWIKI 2011</td>
<td>Wikipedia</td>
<td>50,312</td>
<td>36,504</td>
<td>346,299</td>
<td>456,886</td>
<td>81,262</td>
</tr>
<tr>
<td>LWIKI 2011</td>
<td>Wikipedia</td>
<td>478,020</td>
<td>325,056</td>
<td>1,617,899</td>
<td>2,365,436</td>
<td>452,167</td>
</tr>
</tbody>
</table>
Baseline Methods (8)

- Hierarchical methods
  - CorrMNL: [Shahbaba and Neal, 2007] Another Bayesian method but using MCMC sampling.
  - HSVM: [Tsochantaridis et al., 2006] A large-margin method with path dependent discriminant function.
  - TD: [Yang et al., 2003] Top-down SVM with pachinko machine.
- Flat methods
  - BSVM - Binary SVM, MSVM - Multiclass SVM, BLR - Binary logistic Regression, MLR - Multiclass logistic Regression.

Scalability Analysis on Benchmark Datasets

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Data Type</th>
<th>#Trn Classes</th>
<th>#Leaf Classes</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEF</td>
<td>X-ray images</td>
<td>63</td>
<td>63</td>
<td>89</td>
</tr>
<tr>
<td>NEWS20</td>
<td>news stories</td>
<td>20</td>
<td>20</td>
<td>53,975</td>
</tr>
<tr>
<td>RCV1 (Topics)</td>
<td>news stories</td>
<td>137</td>
<td>101</td>
<td>48,734</td>
</tr>
<tr>
<td>LSHTC-small</td>
<td>web pages</td>
<td>1,563</td>
<td>1,139</td>
<td>51,033</td>
</tr>
<tr>
<td>IPC</td>
<td>patents</td>
<td>552</td>
<td>451</td>
<td>541,869</td>
</tr>
<tr>
<td>DMOZ 2010</td>
<td>web pages</td>
<td>15,358</td>
<td>12,294</td>
<td>381,580</td>
</tr>
<tr>
<td>DMOZ 2012</td>
<td>web pages</td>
<td>13,347</td>
<td>11,947</td>
<td>594,158</td>
</tr>
<tr>
<td>DMOZ 2011</td>
<td>web pages</td>
<td>35,448</td>
<td>27,875</td>
<td>348,548</td>
</tr>
<tr>
<td>SWIKI 2011</td>
<td>Wikipedia</td>
<td>50,312</td>
<td>36,504</td>
<td>346,299</td>
</tr>
<tr>
<td>LWIKI 2011</td>
<td>Wikipedia</td>
<td>478,020</td>
<td>325,056</td>
<td>1,617,899</td>
</tr>
</tbody>
</table>

#Parameters (#Leafs * #Feat) in joint optimization

- PASCAL LSHTC Challenge
- RR-SVM, RR-LR
- 3-4 orders of magnitude
- HSVM, MSVM
- Popular in ML papers
RR-SVM vs. the state of the art
(performance measured in micro/macro-avg F1)

RR-SVM vs. BSVM on DMOZ-2012
Performance Improvement in Macro-avg
Concluding Remarks

- Large-scale classification is an important part of machine learning in the big-data era.
- Large hierarchies and graphs of categories present significant challenges & opportunities for structured learning.
- We propose a new approach (RR-SVM and RR-LR) to the joint optimization of ~500K classifiers on PASCAL Challenge data sets.
- We aim to significantly improve the state of the art in extreme classification in both accuracy and scalability.

Outline

- Introduction (expected time: 5 min)
- Related Work (expected time: 13 min)
- Proposed Approach (expected time: 12+ min)
- Evaluation Plan (expected time: 5 min)
- Q/A (expected time: 5 min)