Focused Crawling with Scalable Ordinal Regression Solvers

Rashmin Babaria, J Saketha Nath, Krishnan S, KR Sivaramakrishnan, Chiranjib Bhattacharyya, M N Murty

Department of Computer Science and Automation
Indian Institute of Science, INDIA

ICML-2007
Focused Crawling & Large scale OR

**Focused Crawling**
- Given a topic (seed pages) find out relevant pages from the web
- Pose Focused Crawling as a large scale OR problem

**Ordinal Regression**
- Fast OR training algorithm — scales to millions of datapoints
  - Fast algorithm to solve an SOCP with one SOC constraint
- Low prediction time
Baseline OR Formulation [Chu & Keerthi, 2005]
Clustering based scalable OR Formulation

- Describe data using clusters instead of data points
Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
  - Class conditional distributions — mixture models with spherical covariance
Clustering based scalable OR Formulation

- Describe data using clusters instead of data points
  - Class conditional distributions — mixture models with spherical covariance
- Using second order moments \((\mu, \sigma^2 I)\), classify clusters
Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
  - Class conditional distributions — mixture models with spherical covariance
- Using second order moments \((\mu, \sigma^2 \mathbf{I})\), classify **clusters**
- Proposed formulation will have constraints **per cluster**
Clustering based scalable OR Formulation

- Describe data using **clusters** instead of data points
  - Class conditional distributions — mixture models with spherical covariance
- Using second order moments \((\mu, \sigma^2 I)\), classify **clusters**
- Proposed formulation will have constraints **per cluster**
- Size of optimization problem \(O(\text{clusters})\) rather than \(O(\text{datapoints})\)
Proposed OR formulation’s solution
Proposed OR formulation’s solution
Proposed OR formulation’s solution
Proposed OR formulation

Features:

- **SOCP Problem with one SOC constraint**

  \[ T_{\text{train}} = T_{\text{clust}} + T_{\text{SOCP}} = O(n) \]

  - Cluster moments estimated using BIRCH [Zhang et.al., 1996]
    \[ T_{\text{clust}} = O(n) \]
  - SOCP solved using SeDuMi\(^a\). \( T_{\text{SOCP}} \) is independent of \( n \)

- Can be Kernelized — using input space cluster moments
  - No. of Support Vectors at max. \( k \) — low prediction time

\(^a\)http://sedumi.mcmaster.ca/
Clustering + SOCP gives speedup

**Table:** Training times (sec) with SeDuMi and SMO-OR [Chu & Keerthi, 2005] on synthetic dataset.

<table>
<thead>
<tr>
<th>S-Rate</th>
<th>S-Size</th>
<th>SMO-OR</th>
<th>SeDuMi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.002</td>
<td>10,000</td>
<td>182</td>
<td>1</td>
</tr>
<tr>
<td>0.0025</td>
<td>12,500</td>
<td>260</td>
<td>1</td>
</tr>
<tr>
<td>0.003</td>
<td>15,000</td>
<td>340</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>1,500,000</td>
<td>×</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>5,000,000</td>
<td>×</td>
<td>36</td>
</tr>
</tbody>
</table>

**Table:** Training times (sec), test error rate with SeDuMi and SMO-OR [Chu & Keerthi, 2005] on CS-Census dataset.

<table>
<thead>
<tr>
<th>S-Size</th>
<th>SMO-OR</th>
<th>SeDuMi</th>
</tr>
</thead>
<tbody>
<tr>
<td>sec (err)</td>
<td>sec</td>
<td></td>
</tr>
<tr>
<td>CS 5,690</td>
<td>893 (.128)</td>
<td>20.4 (.109)</td>
</tr>
<tr>
<td>11,393</td>
<td>5281.6 (.107)</td>
<td>108.8 (.112)</td>
</tr>
<tr>
<td>15,191</td>
<td>9997.5 (.107)</td>
<td>271.1 (.108)</td>
</tr>
<tr>
<td>22,331</td>
<td>×</td>
<td>435.7 (.119)</td>
</tr>
</tbody>
</table>
Large number of clusters is still challenging

**Table:** Training times (sec), test error rate with **SeDuMi** and **SMO-OR** [Chu & Keerthi, 2005] on CH-California Housing dataset.

<table>
<thead>
<tr>
<th>S-Size</th>
<th>SMO-OR</th>
<th>SeDuMi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sec (err)</td>
<td>sec</td>
</tr>
<tr>
<td></td>
<td>112 (.623)</td>
<td>768.8 (.634)</td>
</tr>
<tr>
<td>CH</td>
<td>10,320</td>
<td>551.9 (.619)</td>
</tr>
<tr>
<td></td>
<td>13,762</td>
<td>1033.2 (.616)</td>
</tr>
<tr>
<td></td>
<td>15,482</td>
<td>1142 (.617)</td>
</tr>
<tr>
<td></td>
<td>17,202</td>
<td>1410 (.617)</td>
</tr>
<tr>
<td></td>
<td>20,230</td>
<td>1838.5 (.62)</td>
</tr>
</tbody>
</table>

× denotes failure.
**Key Idea:**
- Exploit special SOCP form — SOCP problem with one SOC constraint
  - Erdougan et.al., 2006 — specialized solvers scale better
- Fast algorithm similar in spirit to Platt’s SMO for QP

**Features:**
- More scalable than generic solvers
- Easy to implement, uses no optimization tools
Rewrite Dual as follows:

\[
\min_{\alpha, \alpha^*} \quad W \sqrt{(\alpha^* - \alpha)^\top K (\alpha^* - \alpha)} - d^\top (\alpha + \alpha^*)
\]

\[
s.t. \quad 0 \leq \alpha \leq 1, \quad 0 \leq \alpha^* \leq 1
\]

\[
s_i^* \leq s_i, \quad \forall \ i = 1, \ldots, r - 2, \quad s_{r-1}^* = s_{r-1}
\]

\(K\) is Gram matrix for cluster centers

\[
s_i = \sum_{k=1}^{i} \sum_{j=1}^{n_k} \alpha_k^j \quad \text{and} \quad s_i^* = \sum_{k=2}^{i+1} \sum_{j=1}^{n_k} \alpha_k^* j
\]
CB-OR Solver

Minimization wrt. two multipliers

\[
\begin{align*}
\min_{\Delta \alpha} & \quad \sqrt{a(\Delta \alpha)^2 + 2b(\Delta \alpha) + c - e\Delta \alpha} \\
\text{s.t.} & \quad lb \leq \Delta \alpha \leq ub
\end{align*}
\]

Has closed form solution:

\[
\Delta \alpha = \begin{cases} 
\frac{\sqrt{ac-b^2}}{a}u_b + \frac{b}{a}l_b & \text{if } ac - b^2 > 0, a - e^2 > 0 \\
\frac{b}{a}u_b + \frac{-b}{a}l_b & \text{if } ac - b^2 = 0, a - e^2 > 0 \\
ub & \text{if } e - \sqrt{a} \geq 0 \\
lb & \text{if } e + \sqrt{a} \leq 0
\end{cases}
\]
CB-OR Solver

CB-OR Algorithm

Step 1  Pick two most KKT violators
Step 2  Solve the 1-d minimization problem
Step 3  Update unknowns
Step 4  Check for KKT violators. If none terminate. Else Step 1
CB-OR — Evaluation

Figure: Dashed line represents training time with \texttt{SeDuMi} and continuous line that with \texttt{CB-OR} on a synthetic dataset.
Table: Comparison of training times (in sec) with CB-OR, SMO-OR and SeDuMi on benchmark datasets. The test set error rate is given in brackets. (CH-California Housing, CS-Census datasets).

<table>
<thead>
<tr>
<th>CH</th>
<th>S-Size</th>
<th>CB-OR</th>
<th>SMO-OR</th>
<th>SeDuMi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sec (err)</td>
<td>sec (err)</td>
<td>sec</td>
<td></td>
</tr>
<tr>
<td>10,320</td>
<td>.5 (.623)</td>
<td>551.9 (.619)</td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>13,762</td>
<td>1.5 (.634)</td>
<td>1033.2 (.616)</td>
<td>768.8</td>
<td></td>
</tr>
<tr>
<td>15,482</td>
<td>8.4 (.618)</td>
<td>1142 (.617)</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>17,202</td>
<td>14.3 (.621)</td>
<td>1410 (.617)</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>20,230</td>
<td><strong>10.4 (.62)</strong></td>
<td><strong>1838.5 (.62)</strong></td>
<td>×</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CS</th>
<th>S-Size</th>
<th>CB-OR</th>
<th>SMO-OR</th>
<th>SeDuMi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sec (err)</td>
<td>sec (err)</td>
<td>sec</td>
<td></td>
</tr>
<tr>
<td>5,690</td>
<td>.3 (.109)</td>
<td>893 (.128)</td>
<td>20.4</td>
<td></td>
</tr>
<tr>
<td>11,393</td>
<td>.7 (.112)</td>
<td>5281.6 (.107)</td>
<td>108.8</td>
<td></td>
</tr>
<tr>
<td>15,191</td>
<td>1 (.108)</td>
<td>9997.5 (.107)</td>
<td>271.1</td>
<td></td>
</tr>
<tr>
<td>22,331</td>
<td><strong>1.5 (.119)</strong></td>
<td>×</td>
<td>435.7</td>
<td></td>
</tr>
</tbody>
</table>
Focused Crawling

- Given a topic (seed pages) find out relevant pages from the web.
- Requires low bandwidth and low disk space.
- Small updation cycle.
Baseline Focused Crawler [Chakrabarti et al., 1999]
Topic Taxonomy
Topic Taxonomy

Some obvious choices for the negative class
Topic Taxonomy

Some other members of negative class

Some obvious choices for the negative class
Topic Taxonomy

Some other members of negative class

Too many categories (and pages)!

Some obvious choices for the negative class
Topic Taxonomy

Some obvious choices for the negative class

Some other members of negative class

Too many categories (and pages)!
How many pages should be selected?
Topic Taxonomy

Some other members of negative class

Too many categories (and pages)!
How many pages should be selected?
AND how to select the pages?

Some obvious choices for the negative class
Topic Taxonomy

- Some other members of negative class
  - Too many categories (and pages)!
  - How many pages should be selected?
  - Maintain balance between classes.

- Some obvious choices for the negative class
  - Computers
  - Tennis
  - Basketball
  - Motorsports
  - Business
  - Billiards
  - Bike Racing
  - Karting
  - Auto Racing
  - ATV Racing
  - SnowMobile Racing
  - Formula One
  - Nascar
  - Rallying
Topic Taxonomy

Some other members of negative class

Too many categories (and pages)!
How to select the pages?
Random Sampling.

Some obvious choices for the negative class
Exploit link structure

- Grangier and Bengio observe that hyperlinked documents are semantically closer.
- One link away pages are more similar to seed pages compare to two link away pages.
Link structure in web
Link structure in web
Link structure in web
Focused Crawling as OR problem — exploit link structure
Focused Crawling as OR problem — exploit link structure
Focused Crawling as OR problem — exploit link structure

Level 1 - Page has many links to level 0 pages (Hub)

Level 0 - Pages belong to topic
Focused Crawling as OR problem — exploit link structure

Level 1 - Page has many links to level 0 pages (Hub)

Level 2 - Some of the links on this page will lead to topic pages.

Level 0 - Pages belong to topic
Baseline Focused Crawling architecture
Proposed Focused Crawling architecture

- Seed set
- Google API
- Level pages
- Training
- Parameters
- Ordinal Regression
- Regression
- Newly Crawled page
- pick best
- Crawl Database
- Frontier URLs priority queue
- URLs
Focused Crawling is a large scale OR problem

<table>
<thead>
<tr>
<th>Category</th>
<th>Seed</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASCAR</td>
<td>1705</td>
<td>1944</td>
<td>1747</td>
<td>1464</td>
<td>1177</td>
</tr>
<tr>
<td>Soccer</td>
<td>119</td>
<td>750</td>
<td>1109</td>
<td>1542</td>
<td>3149</td>
</tr>
<tr>
<td>Cancer</td>
<td>138</td>
<td>760</td>
<td>895</td>
<td>858</td>
<td>660</td>
</tr>
<tr>
<td>Mutual Funds</td>
<td>371</td>
<td>395</td>
<td>540</td>
<td>813</td>
<td>1059</td>
</tr>
</tbody>
</table>
NASCAR harvest rate
Cancer harvest rate

![Graph showing cancer harvest rate over total pages crawled with two lines: OR and LR.](image-url)
Mutual Funds harvest rate

![Graph showing harvest rate over total pages crawled]

- **OR**
- **LR**
## Harvest rate comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASCAR</td>
<td>0.3698</td>
<td>0.6977</td>
</tr>
<tr>
<td>Cancer</td>
<td>0.4714</td>
<td>0.58</td>
</tr>
<tr>
<td>Mutual Fund</td>
<td>0.526</td>
<td>0.5969</td>
</tr>
<tr>
<td>Soccer</td>
<td>0.34</td>
<td>0.4952</td>
</tr>
</tbody>
</table>
Conclusions

- Proposed a scalable clustering based OR formulation
  - Training time \( O(\text{datapoints}) \)
  - Support Vectors \( O(\text{clusters}) \)
- Exploited special structure of the formulation to develop a fast solver, CB-OR
  - Scalable to tens of thousands of clusters
- We formulated focused crawling as large scale ordinal regression
  - No need for negative class definition
  - Independent of topic taxonomy
  - OR captures link structure of web graph.
Focused crawler code available at

http://mllab.csa.iisc.ernet.in/downloads/focusedcrawler.html
Acknowledgments

This project is partially supported by AOL India Pvt Ltd and DST, Government Of India (DST/ECA/CB/660)
Questions?