Focused Crawling with Scalable Ordinal Regression Solvers

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

Dept. of Computer Science and Automation, Indian Institute of Science, INDIA

{rashmin,saketha,krishnan,chiru,mnm}@csa.iisc.ernet.in

Overview

- Propose a clustering based, scalable, OR formulation:
  - Classifies data clusters, instead of data points
  - Instance of SOCP with one SOC constraint.
- Develop a fast solver, for proposed OR formulation:
  - Exploit the special structure — SOCP with only one SOC constraint
  - Fast iterative algorithm similar in spirit to Platt's SMO [3] for Quadratic Programs (QP)
- Pose problem of Focused crawling as large OR problem:
  - Exploits the link structure in web — increase chance of crawling relevant pages
  - Avoids need for topic taxonomy and negative class data

Clustering based OR formulation

- Class conditional densities are modeled using mixture distributions with spherical covariance.
- Moments $(\mu, \sigma^2 \mathbf{I})$ efficiently estimated using a scalable clustering algorithm like BIRCH, in linear time $T_{clust} = O(datapoints)$.
- Constrain that most of class i data points lie between $\mathbf{w}^\top \mathbf{x} - b_{i-1} = 0$ and $\mathbf{w}^\top \mathbf{x} - b_i = 0$: $\mathbf{P}(\mathbf{w}^\top \mathbf{X} - \mathbf{b_i} \le -1 + \xi)$ and $\mathbf{P}(\mathbf{w}^\top \mathbf{X} - \mathbf{b_{i-1}} \ge 1 - \xi^*)$ **is high**
- Each such cluster constraint can be written as a single SOC constraint
- Following the arguments in [4], the clustering based large margin OR formulation is:

$$
\begin{aligned}
\min_{\mathbf{w},\mathbf{b},\xi_i^j,\xi_i^{*j}} \quad & \sum_{i=1}^{r}\sum_{j=1}^{k_i} \xi_i^j + \xi_i^{*j} \\
\text{s.t.} \quad & \mathbf{w}^\top \mu_i^j - b_i \le -1 + \xi_i^j - \kappa \sigma_i^j W, \\
& \mathbf{w}^\top \mu_i^j - b_{i-1} \ge 1 - \xi_i^{*j} + \kappa \sigma_i^j W, \\
& \xi_i^j \ge 0,\ \xi_i^{*j} \ge 0,\ \forall\, i, j,\ \|\mathbf{w}\|_2 \le W, \\
& b_i - b_{i-1} > 0,\ i = 2, \ldots, r-1
\end{aligned} \qquad (1)
$$

- **Instance of SOCP with single SOC constraint**
- Using cluster moment information in input space and employing the kernel trick, the formulation can be extended to non-linear cases.
- The kernelized dual can be written as:

$$
\begin{aligned}
\max_{\alpha,\alpha^*,\rho} \quad & \mathbf{d}^\top (\alpha + \alpha^*) - \rho W \\
\text{s.t.} \quad & \sqrt{(\alpha^* - \alpha)^\top \mathbf{K} (\alpha^* - \alpha)} \le \rho, \\
& 0 \le \alpha \le 1, 0 \le \alpha^* \le 1 \\
& s_i^* \le s_i,\ \forall\, i = 1, \ldots, r-2, s_{r-1}^* = s_{r-1}
\end{aligned} \qquad (2)
$$

  where α, α^*, ρ are the Lagrange multipliers, $\mathbf{K}$ is the gram matrix for cluster centers, $\mathbf{d}$ is calculated using κ, σ_i^j and Gaussian kernel parameter, $s_i = \sum_{k=1}^{i} \sum_{j=1}^{n_k} \alpha_k^j$ and $s_i^* = \sum_{k=2}^{i+1} \sum_{j=1}^{n_k} \alpha_k^{*j}$
- **The size and no. constraints in the above dual SOCP formulation (2) are O(clusters) rather than O(datapoints).** Hence formulation scales well for very large datasets
- Dual SOCP can be solved using generic solvers like SeDuMi[1]. **Overall training time is linear** $\mathbf{T_{train}} = \mathbf{T_{clust}} + \mathbf{T_{SOCP}} = \mathbf{O(datapoints)}$
- The number of support vectors is at max. no. clusters rather than no. datapoints. Hence the prediction time is expected to be low for the proposed formulation.

Fast solver for single-cone constraint SOCP

- At every iteration, KKT conditions are evaluated. If the optimal solution is found then the algorithm terminates. Else the **maximum KKT violating pair is chosen and their values are incremented or decremented by a quantity** $\mathbf{\Delta}\alpha$ such that the constraint $s_{r-1} = s_{r-1}^*$ always holds.
- $\Delta\alpha$ chosen such that value of the objective function (2) has maximum decrease. **Expressed as following 1-d minimization problem**:

$$
\begin{aligned}
\min_{\Delta\alpha} \quad & \sqrt{a(\Delta\alpha)^2 + 2b(\Delta\alpha) + c} - e\Delta\alpha \\
\text{s.t.} \quad & lb \le \Delta\alpha \le ub
\end{aligned} \qquad (3)
$$

- The values a, b, c, lb, ub can be easily calculated from the parameters of the problem, at every step, in $O(clusters)$ calculations.
- **1-d minimization problem (3) has a closed form solution.**
- Values of α and α^* updated accordingly using $\Delta\alpha$ and the procedure is repeated in the next iteration.
- **The fast solver avoids the use of any generic optimization tools and can be shown to be more scalable than SeDuMi.**

Focused Crawling as OR problem

- Focused crawling - Given a topic (seed pages) find out relevant pages from the web
- Requires low bandwidth, low disk space and small updation cycles
- Focused crawling was coined by Chakrabarti et.al. [6]
  - A classifier is trained to determine the relevance of newly crawled web pages.
  - For any topic, the negative set is very large and diverse, which makes it difficult to construct the training set.
  - Topic taxonomy is used to choose negative set.
  - Out links from relevant pages are crawled with higher priority
- **Exploit link structure in web**
  - Some off-topic pages often lead to topic pages.
  - Binary classifier classifies off-topic pages as negative class pages.
  - Grangier and Bengio [7] observe that any document is semantically closer to documents hyperlinked with it, than to documents which are not.
  - Pages which are one link away are semantically closer to seed pages than pages that are two links away.
  - **Rank the documents based on their link distance to the topic pages.**

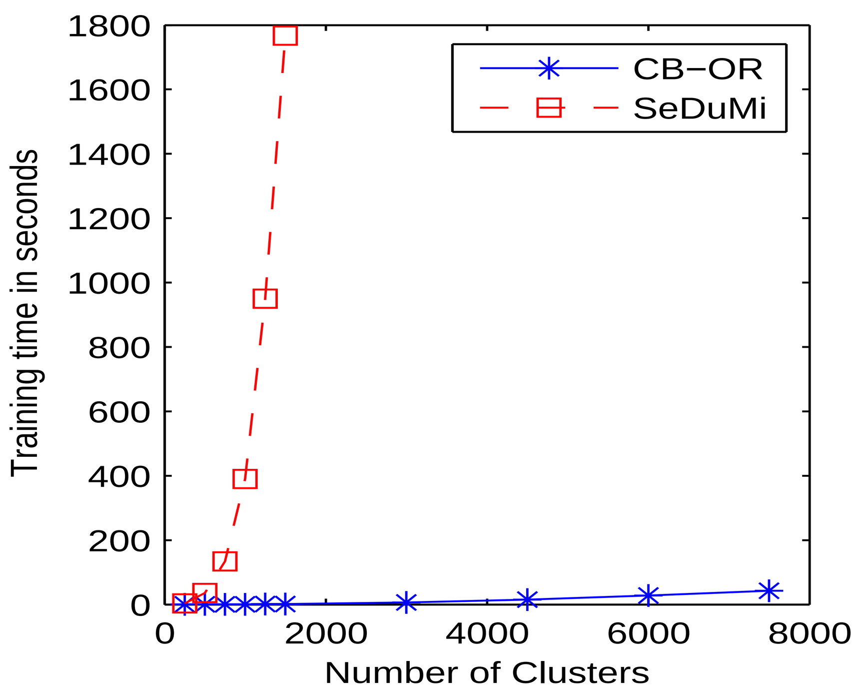
- Pose focused crawling as an OR problem instead of a classification problem
- The training data for different ordinal class can be generated from the seed set using back link information given by Google API.
- Out links are crawled with priority based on the rank predicted by the Ordinal Regressor.

Experimental Results

- Scaling experiment results on CH-California housing and CS-Census datasets are shown below. They compare the training time (in sec) and test error rate (err) of **SMO-OR** [5], proposed formulation (2) solved using SeDuMi (denoted by **SeDuMi**) and the fast solver (denoted by **CB-OR**):

	S-Size	CB-OR	SMO-OR	SeDuMi
		sec (err)	sec (err)	sec
CH	10,320	.5 (.623)	551.9 (.619)	112
	13,762	1.5 (.634)	1033.2 (.616)	768.8
	15,482	8.4 (.618)	1142 (.617)	×
	17,202	14.3 (.621)	1410 (.617)	×
	20,230	**10.4** (.62)	**1838.5** (.62)	×
CS	5,690	.3 (.109)	893 (.128)	20.4
	11,393	.7 (.112)	5281.6 (.107)	108.8
	15,191	1 (.108)	9997.5 (.107)	271.1
	22,331	**1.5** (.119)	×	**435.7**

- Figure below compares the scalability of **SeDuMi** and **CB-OR** on synthetic datasets with varying number of clusters. Clearly **CB-OR** scales better than **SeDuMi**:



- Following table describes training set sizes for different category.

Category	Seed	1	2	3	4
NASCAR	1705	1944	1747	1464	1177
Soccer	119	750	1109	1542	3149
Cancer	138	760	895	858	660
Mutual Funds	371	395	540	813	1059

- Following table gives a comparison of the performance of FOCUS with the baseline crawler.

Dataset	#Good/#Bad	Baseline	OR
NASCAR	11530/19646	.3698	.6977
Soccer	10167/9131	.34	.4952
Cancer	6616/12397	.4714	.58
Mutual Fund	9960/10992	.526	.5969

References

[1] Zhang, T., Ramakrishnan, R., & Livny, M., BIRCH: an efficient data clustering method for very large databases. *ICMD, 1996*.

[2] Jos F. Sturm. Optimization software over symmetric cones. http://sedumi.mcmaster.ca/

[3] John Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel MethodsSupport Vector Learning, 1999*.

[4] Saketha Nath, J., Bhattacharyya, C., & Murty, M. N. Clustering based large margin classification: a scalable approach using socp formulation. *SIGKDD, 12, 2006*.

[5] Chu, W., & Keerthi, S. S. New approaches to support vector ordinal regression. *ICML, 2005*.

[6] Chakrabarti, S., van den Berg, M. & Dom, B. Focused Crawling: A New Approach for Topic-Specific Resource Discovery. *WWW, 1999*.

[7] Grangier, D. and Bengio, S. Exploiting Hyperlinks to Learn a Retrieval Model. *NIPS, 2005*

[1] http://sedumi.mcmaster.ca/

International Conference on Machine Learning, June $20-24$ 2007, Oregon