Extreme Classification with Large-scale Structured Learning

Yiming Yang, Carnegie Mellon University
Joint work with Siddharth Gopal
(NIPS 2012, KDD 2013, ICML 2014)

Large-Scale Classification Challenges

- **Modeling challenge** - Leveraging the dependency structures among categories in very large *hierarchies* and *networks*
- **Computational challenge** – Making the *joint* optimization of all classifiers (hundreds of thousands) tractable
- **Evaluation challenge** – International benchmark evaluations, e.g., the *PASCAL Challenge* for *Large-Scale Hierarchical Text Categorization* (*LSHTC*), 2010 – present

Benchmark Evaluation Data Sets (Examples)

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

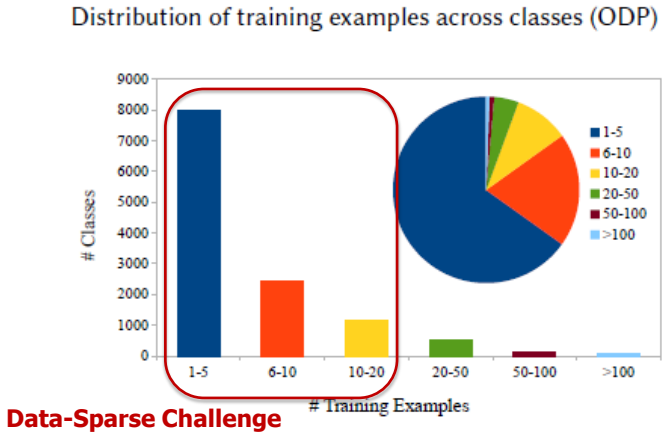
The PASCAL LSHTC Challenge

Large Taxonomies → Large # of Parameters

- **LWIKI, the largest dataset in the PASCAL LSHTC Challenge**
  - 478,020 Wikipedia categories in a directed graph
  - **614,428 categories** after adding "misc." nodes that pull out instances from internal nodes
  - 1,617,899 unique words in the vocabulary of web pages
- **The number of parameters in joint optimization**
  - **1,617,899 x 614,428 = 1 trillion (4 TB)**

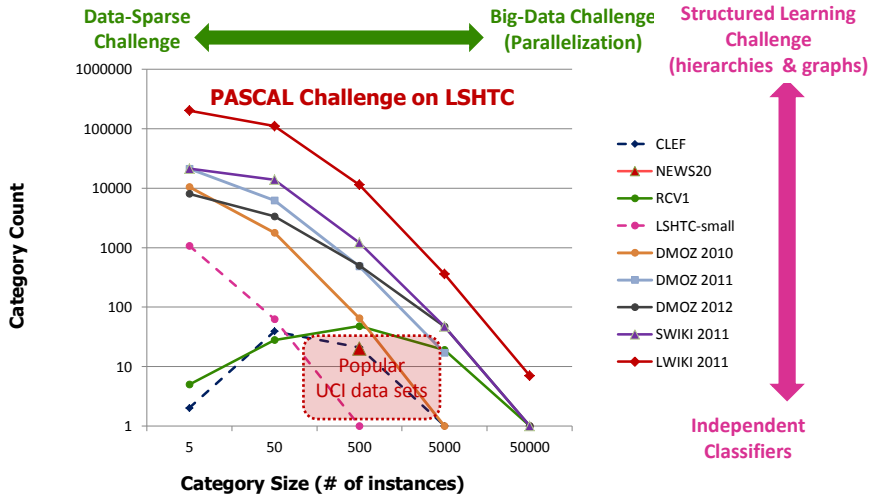
Large Taxonomies → Skewed Category Distribution

Majority of classes are rare



Data-Sparse Challenge

The landscape of research challenges



Outline

- ✓ Motivation & Related Work
- Our Work
  - Bayesian Hierarchical Logistic Regression (BHLR) (S Gopal, NIPS 2012)
  - Recursively Regularized SVM and LR (S Gopal & Y Yang, KDD 2013)

Notation

- $\mathcal{N}$ is the set of all the nodes in a hierarchy or a graph ;
- $\mathcal{T}$ is the set of leaf nodes (a subset of $\mathcal{N}$);
- n is a specific node'
- $\pi(n)$ is the parent of node n ;
- $C(n)$ is the set of the children of node n ;
- $D = \{(x_i, t_i)\}$ for $i = 1$ to N is a training set where instance x_i belongs to exactly one leaf-node category t_i in $\mathcal{T}$;
- $y_{in} = 1$ or -1, denoting whether instance x_i is assigned to node n .

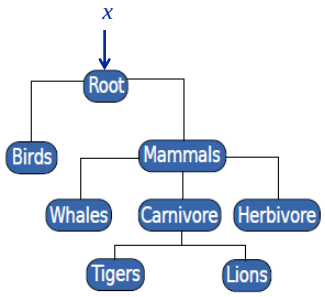
Hierarchical Bayesian Logistic Regression (HBLR)

- Training a softmax classifier for every internal node
- Propagating model parameters from parents to children as

$$w_n \sim N(w_{\pi(n)}, \Sigma_?)$$

- Jointly optimizing all w's and Σ's over the hierarchy
- At each leaf node, predicting class labels as

$$p(t \mid x) = \frac{\exp(w_t^T x)}{\sum_{t' \in T} \exp(w_t^T x)} \quad \text{where} \quad \sum_{t' \in T} p(t \mid x) = 1$$

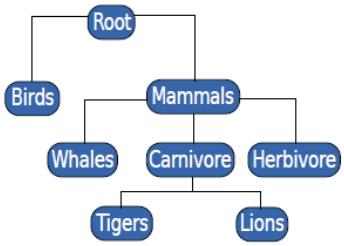


HBLR Model 1 (M1)

M1

$$w_n \sim N(w_{\pi(n)}, \Sigma_{\pi(n)}) \quad \forall n$$
$$\Sigma_{\pi(n)} = \alpha_{\pi(n)}^{-1} I$$
$$\alpha_n \sim \Gamma(a_n, b_n) \quad \forall n \notin \mathcal{T}$$

- Siblings share a common covariance matrix $\Sigma_{\pi(n)}$.
- $\Sigma_{\pi(n)}$ determines how close or far the siblings are from parent.
- Mammals in general are more varied than carnivores.

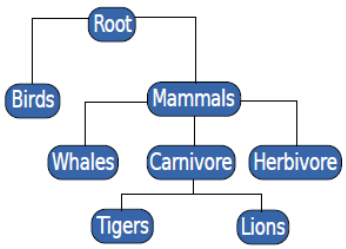


HBLR Model 2 (M2)

M2

$$w_n \sim N(w_{\pi(n)}, \Sigma_{\pi(n)}) \quad \forall n$$
$$\Sigma^{-1}_{\pi(n)} = diag(\alpha^{(1)}_{\pi(n)}, \alpha^{(2)}_{\pi(n)}, \ldots, \alpha^{(d)}_{\pi(n)})$$
$$\alpha^{(i)}_n \sim \Gamma(a^{(i)}_n, b^{(i)}_n) \quad i = 1..d, \forall n \notin \mathcal{T}$$

- Feature-specific variance.
- Birds and Mammals are close in some dimension such as 'eyes', 'claw', but not in other dimensions such as 'feathers'

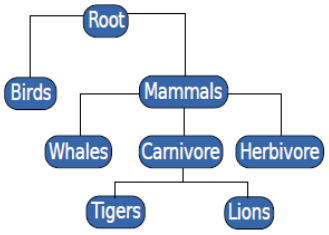


HBLR Model 3 (M3)

M3

$$w_n \sim N(w_{\pi(n)}, \Sigma_n) \quad \forall n$$
$$\Sigma_n = \alpha_n^{-1} I$$
$$\alpha_n \sim \Gamma(a_n, b_n) \quad \forall n$$

- Node-specific covariance matrix Σ_n.
- Each node individually determines how far/close it is to the parent.
- Whales is not a typical Mammal and is an 'outlier'.



Inference

The main inference problem in Bayesian methods,

$$P(\mathbf{W}, \boldsymbol{\alpha} | \mathcal{D}, \mathbf{a}, \mathbf{b}) \propto \overbrace{P(\mathcal{D}|\mathbf{W}, \boldsymbol{\alpha})}^{likelihood} \overbrace{P(\boldsymbol{\alpha}, \mathbf{W} \mid \mathbf{a}, \mathbf{b})}^{prior}$$

$$\mathbf{W} = \{ \mathbf{w}_n : n \text{ in } \mathcal{N} \}$$

For M2

$$\propto \overbrace{\prod_{(x,t)\in D} \frac{\exp(w_t^\top x)}{\sum_{t'\in\mathcal{T}} \exp(w_{t'}^\top x)}}^{likelihood} \overbrace{\prod_{n\notin\mathcal{T}} \prod_{i=1}^{d} \Gamma(\alpha_n^{(i)} | a_n^{(i)}, b_n^{(i)}) \prod_{n\in\mathcal{N}} N(w_n | w_{\pi(n)}, \Sigma_{\pi(n)})}^{prior}$$

No analytical solution. Resort to numerical methods.

Hierarchical Variational Inference

- MCMC (existing solution) is too slow when the number of model parameters is very large (e.g., a trillion)
- Variational inference is a better alternative (our algorithm is the first one for hierarchical LR models)
- Find the posterior Q that approximates the true posterior P as

$$\min_{\boldsymbol{\mu},\boldsymbol{\tau},\upsilon,\boldsymbol{\Psi}} KL(\overbrace{Q(\mathbf{W}, \boldsymbol{\alpha} | \boldsymbol{\mu}, \boldsymbol{\tau}, \upsilon, \boldsymbol{\Psi})}^{approx\ posterior} \,||\, \overbrace{P(\mathbf{W}, \boldsymbol{\alpha} | \mathcal{D}, \mathbf{a}, \mathbf{b})}^{true\ posterior})$$

- Iteratively optimize until convergence (to a local maxima)

Outline

- ✓ Motivation & Related Work
- ✓ Our Work
  - ✓ Bayesian Hierarchical Logistic Regression (BHLR) (S Gopal, NIPS 2012)
  - – Recursively Regularized SVM and LR (S Gopal & Y Yang, KDD 2013)

Joint Regularization of All Models ("Recursive Regularization")

- We minimize the *expected risk* by finding

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{w}} \lambda(\mathbf{w}) + C \times R_{emp}(\mathbf{w}, D_{train})$$

Regularization term Empirical risk

- Incorporate a hierarchy (H) into regularization

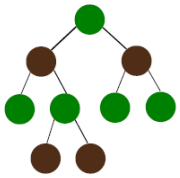
$$\lambda_H(\mathbf{W}) = \sum_{n \in \mathrm{N}} \| w_n - w_{\pi(n)} \|^2 \quad \text{where N is fullset of nodes}$$

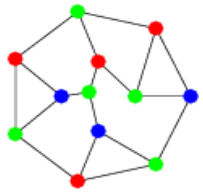
- Incorporate a graph (G) into regularization

$$\lambda_G(\mathbf{W}) = \sum_{(i.j) \in \mathrm{E}} \| w_i - w_j \|^2 \quad \text{where } \mathrm{E} = \{(i, j) : i, j \in N\}$$

Parallel Divide-&-Conquer Strategies

- **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

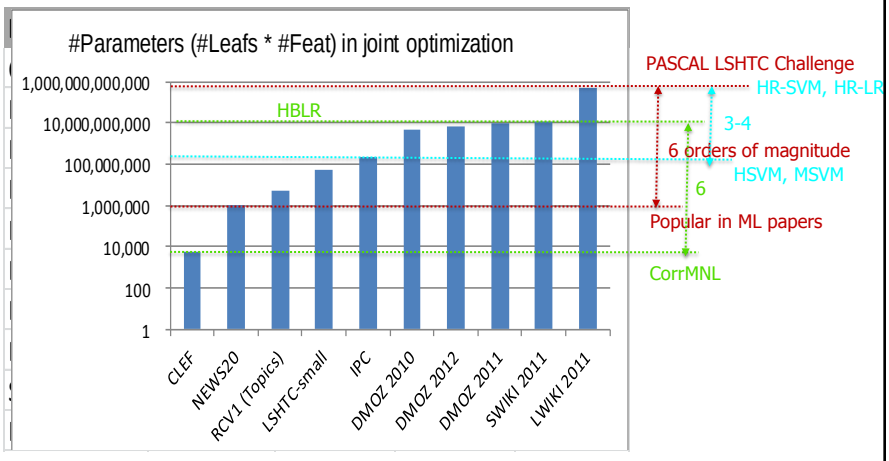




Evaluation: HBLR vs. Baseline Methods (8) + 3 New Methods (Ours)

- Hierarchical methods
  1. CorrMNL: [Shahbaba and Neal, 2007] Another Bayesian method but using MCMC sampling.
  2. HSVM: [Tsochantaridis et al., 2006] A Large-margin method with path dependent discriminant function.
  3. OT: [Zhou et al., 2011] A Large-margin method with orthogonality constraints.
  4. TD: [Yang et al., 2003] Top-down SVM with pachinko machine.
- Flat methods
  1. BSVM - Binary SVM, MSVM - Multiclass SVM, BLR - Binary logistic Regression MLR - Multiclass logistic Regression.

Data Sets (10) & Scalability of Methods



Comparison with Correlated Multinomial Logit

CorrMNL [Shahbaba and Neal, 2007] : A Bayesian version of correlated Multinomial Logit with MCMC Sampling for inference.

Dataset	Metric-type	Metric	CorrMNL	HBLR
CLEF	Effectiveness			
		Macro-F_1	55.59	**59.65**
		Micro-F_1	81.10	**81.41**
	Efficiency			
		Time (mins)	2270	3

Improvement in *Macro-F_1* : 7.3%

Improvement in scalability : 750x

Comparison with Hierarchical SVM

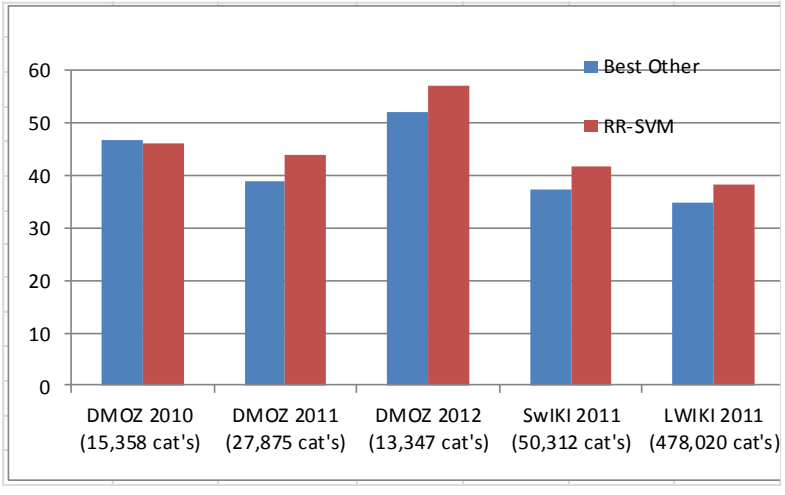
HSVM [Tsochantaridis et al., 2006] : A Large-margin discriminative method based on structured SVM.

Dataset	Metric-type	Metric	HSVM	HBLR
CLEF	Effectiveness			
		Macro-F_1	57.23	**59.65**
		Micro-F_1	79.92	**81.41**
	Efficiency			
		Time (mins)	3.19	3
LSHTC-small	Effectiveness			
		Macro-F_1	21.95	**30.81**
		Micro-F_1	39.66	**46.03**
	Efficiency			
		Time (mins)	289.60	5.2

Improvement in *Macro-*F_1 : 22.5%

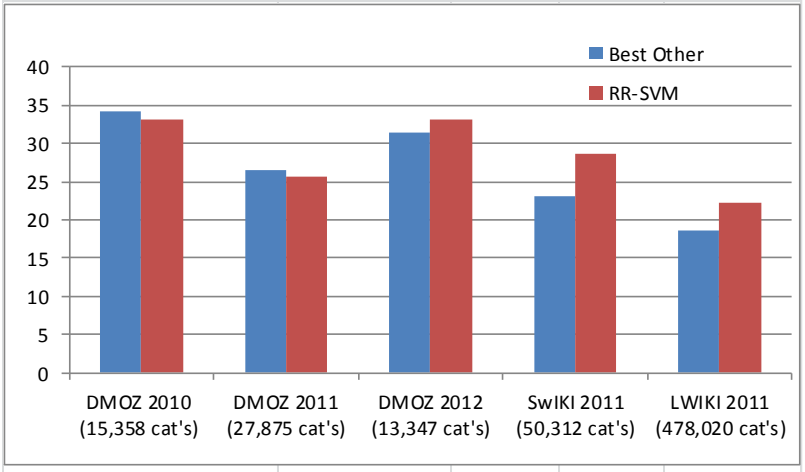
Improvement in scalability : 30x

RR-SVM vs. the state of the art (results in Micro-avg F1)

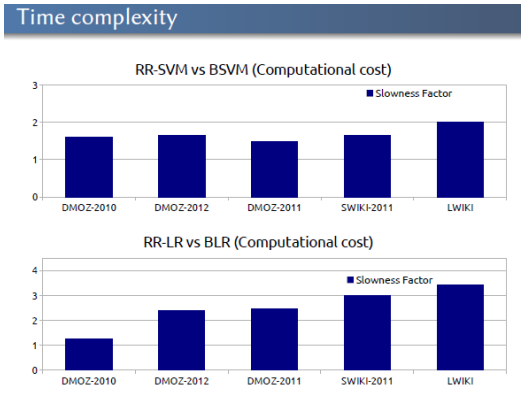


RR-SVM vs. the state of the art (results in Macro-avg F1)

Macro-F1	Best Other	RR-SVM	RR-LR	HBLR (M3)



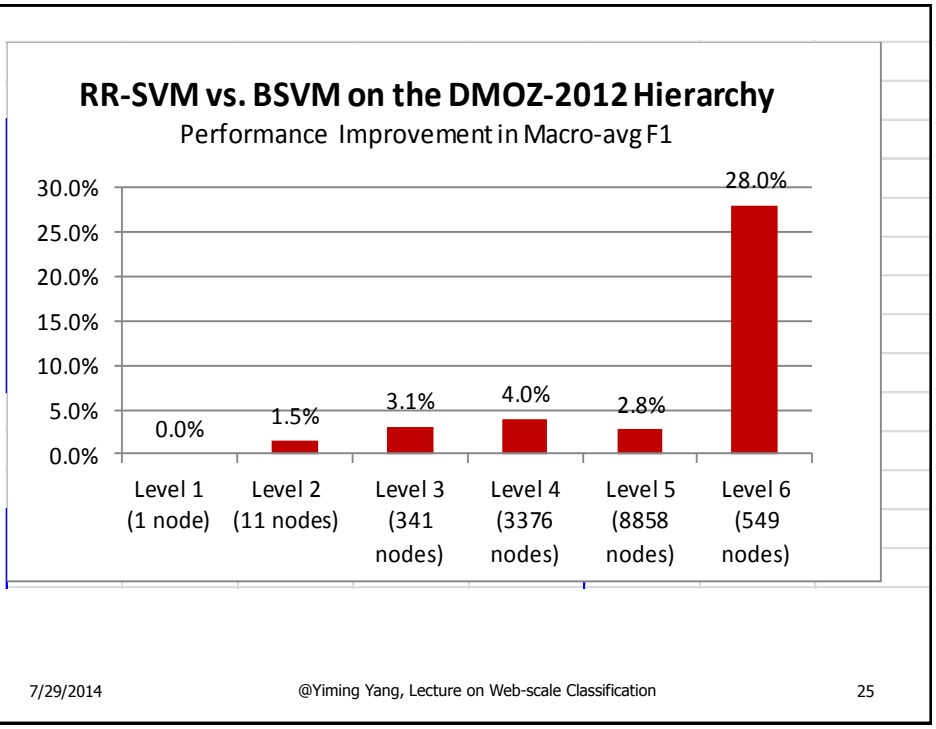
Computation Time in Training

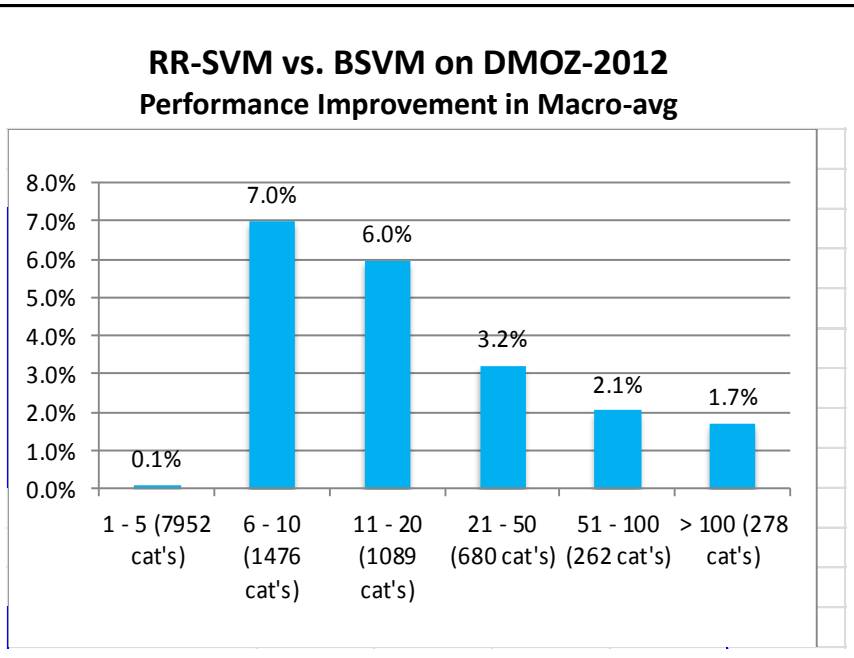


On the LWIKI dataset
-- jointly optimizing 614,428 classifiers with 1 trillion parameters

BSVM	19 hours
RR-SVM	37 hours
BLR	36 hours
RR-LR	121 hours

We used a Hadoop with 500+ cores (300 cores as Mappers and 220 cores as reducers).


RR-SVM vs. BSVM on the DMOZ-2012 Hierarchy
Performance Improvement in Macro-avg F1
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
0.0%
1.5%
3.1%
4.0%
2.8%
28.0%
Level 1 (1 node)
Level 2 (11 nodes)
Level 3 (341 nodes)
Level 4 (3376 nodes)
Level 5 (8858 nodes)
Level 6 (549 nodes)


RR-SVM vs. BSVM on DMOZ-2012
Performance Improvement in Macro-avg
8.0%
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
0.1%
7.0%
6.0%
3.2%
2.1%
1.7%
1 - 5 (7952 cat's)
6 - 10 (1476 cat's)
11 - 20 (1089 cat's)
21 - 50 (680 cat's)
51 - 100 (262 cat's)
> 100 (278 cat's)

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 presented novel methods & scalable algorithms for joint optimization of, e.g., 600+ thousands of classifiers with one trillion of model parameters (4 TB) in 37 hours.
- HBLR, RR-SVM and RR-LR obtained the best results on the largest data sets in benchmark evaluations (PASCAL/LSHTC).
- None of the other joint optimization methods have scaled to the largest problems in the PASCAL Challenges.