Efficiency and Accuracy Issues in LSH(T)C

Eric Gaussier
(R. Babbar, I. Partalas, M.-R. Amini, C. Amblard)

Lab. d'Informatique de Grenoble
Universite Grenoble-Alpes, Grenoble

2 April 2015

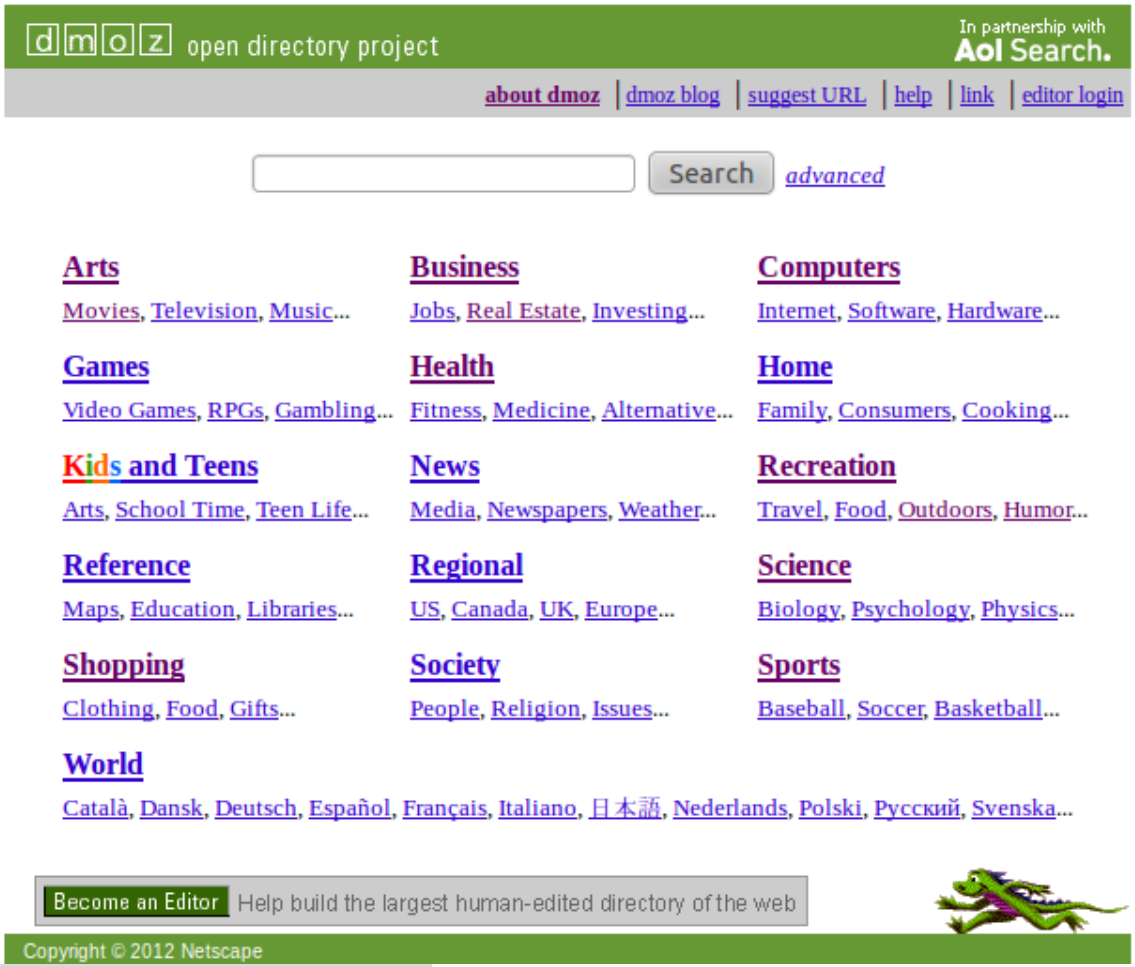
Need for Classification in Big Data

- New data become available:
  - 10,000 articles in **Wikipedia** every day
  - 100 hours every minute in **YouTube**
  - 20,000 scientific articles in **PubMed** every week
- Need for automated methods to categorize these data for:
  - Annotation (enrichment) and archiving purposes
  - Access purposes (ease of retrieval and browsing)

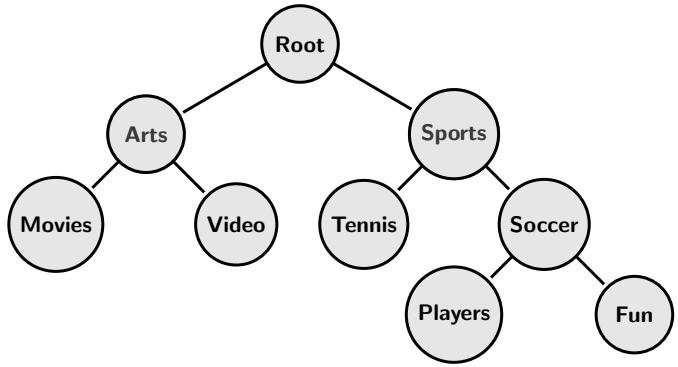
Data Organization

More and more large category systems, with hierarchical structure to organize data

- Directory Mozilla



- ca. 5×10^6 sites
- ca. 10^6 categories
- ca. 10^5 editors



Other examples of large scale taxonomies

- Wikipedia: over 600,000 categories organized in a graph (tree backbone)
- Medical Subject Heading[1]: over 27,000 categories organized in a graph (tree backbone)
- International Patent Collection[2]: 60,000 categories in a tree hierarchy
- Amazon (product hierarchy), Yahoo! Directory, ...

[1] https://www.nlm.nih.gov/mesh/
[2] http://www.wipo.int/classifications/ipc/en/

Evaluation challenges for large-scale classification

Recent challenges have been organized to push the state-of-the-art:

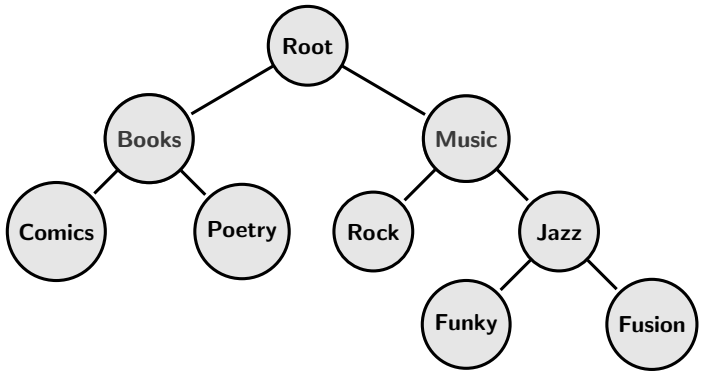
- Large Scale Hierarchical Text Classification[3] (2009-2014) : For large-scale text classification in tree and DAG structures
- BioASQ Challenge[4] (2012-2014) : Classification of abstracts of bio-medical data from National Library of Medicine using the Medical Subject Headings Hierarchy
- Large Scale Visual Recognition Challenge [5] (2010-2013) : Classification of Images in Large-scale setup

[3] http://lshtc.iit.demokritos.gr
[4] http://www.bioasq.org
[5] http://www.image-net.org/challenges/LSVRC/2013/

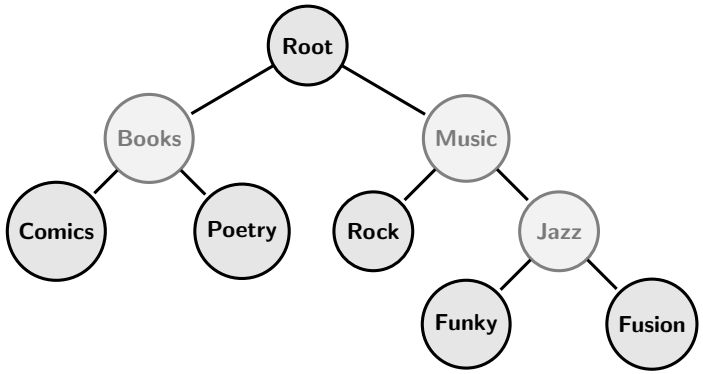
Approaches for classification in large-scale taxonomies

Linear classifiers learned from data

- Hierarchical
  - Top-down - solve individual classification problems at every node (prone to cascading errors) [Liu et al., 2005, Bennett and Nguyen, 2009]
  - Big-bang - solve the problem at once for entire tree (not suitable for large-scale problems) [Cai and Hofmann, 2004, Dekel, 2009, Gopal et al., 2012]



- Flat - ignore the taxonomy *altogether* (higher training and prediction complexities) [Bengio et al., 2010, Gao and Koller, 2011, Perronnin et al., 2012]



- Mildly Hierarchical - use the hierarchical structure *partially* [Malik, 2009]
  - Not clear which layers to remove

Challenges in large-scale classification (1/3)

Which approach to use, flat or hierarchical? Which hierarchy?

Hierarchical vs flat

- *Hierarchical classification*: according to [Liu et al., 2005], outperforms flat classification in terms of classification accuracy and training time on large-scale taxonomies
- *Flat classification*: according to [Bengio et al., 2010], outperforms hierarchical classification on Imagenet dataset
- Build a taxonomy to achieve logarithmic prediction time

Which hierarchy to use?

- Taxonomies are designed by humans for humans - not meant to (fully) optimize classification accuracy
- [Bengio et al., 2010] automatically builds a taxonomy to achieve logarithmic prediction time

Challenges in large-scale classification (2/3)

Scale of datasets

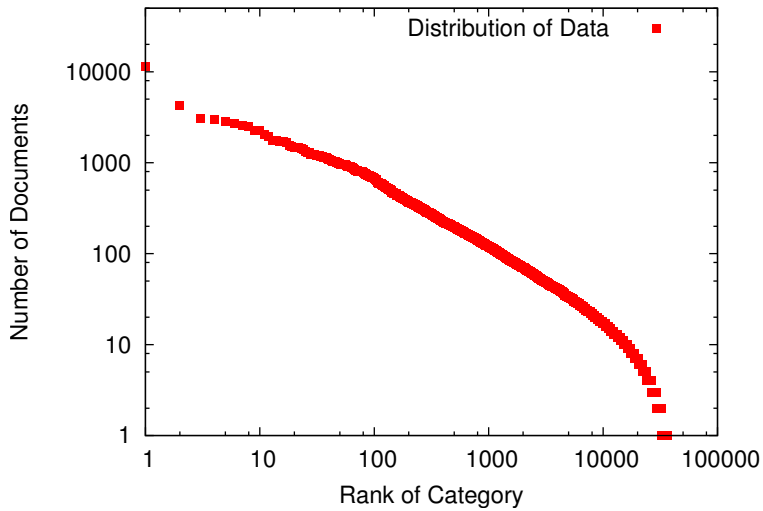
	#Categories	#Features	#Documents	Tree Depth
DMOZ	27,875	594,158	497,992	5
Wiki Small	36,504	346,299	538,148	10
Wiki Large	325,056	1,617,899	2,817,603	14
IPC	451	1,123,497	46,324	3

- $27{,}875 \times 594{,}158 = 16{,}562{,}154{,}250$ ($\approx$ 16 Billion) parameters when learning One-vs-Rest flat classifier for DMOZ dataset, 525,907,777,344($\approx$ 500 Billion) parameters for Wikipedia dataset (LSHTC)
- 123GB of disk space when using Liblinear to store DMOZ model, 2TB disk space for Wikipedia
- Time complexity for hierarchical lower than for flat ($\mathcal{O}(\log K)$ vs $\mathcal{O}(K)$); What about space complexity?

Challenges in Large-scale classification (3/3)

Rare Category Phenomena

- Distribution of documents among categories in Wikipedia subset exhibits *power-law* phenomena
  - Approximately 15,000 of the 36,000 categories have ≤ 5 documents
  - Approximately 4,000 of the 36,000 categories have just 1 document
- Difficulty to learn on such rare categories



The "hierarchical vs flat" debate (1)

Is it better to use a hierarchical or a flat strategy?

To answer this question

- General framework: flat special case of hierarchical strategy
- Consider linear classifiers (large scale), potentially in feature spaces
- Upper bound on generalization error (concentration inequalities)

Notations (multi-class classification)

- *Training Set:* $\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m}, \mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y^{(i)} \in \mathcal{Y}\}$
- *Linear classifiers* ($\mathcal{F}$): $f(\mathbf{x}, y) = \langle \mathbf{x}, \mathbf{w}_y \rangle$, $\mathbf{w}_y \in \mathbb{R}^d$;
- *Decision function:* $g_f(\mathbf{x}, y) \geq 0$ $(g_f(\mathbf{x}, y) = f(\mathbf{x}, y) - \max_{y' \neq y} f(\mathbf{x}, y'))$
- *Generalization error:* $\mathcal{E}(g_f) = \mathbb{E}_{\sim P(x,y)}[g_f(\mathbf{x}, y) < 0]$
- *Empirical error:* $\mathcal{E}_{emp}(g_f) = \frac{1}{m} \sum_i \mathbf{1}(g_f(\mathbf{x}^{(i)}, y^{(i)}) < 0)$

The "hierarchical vs flat" debate (2)

Some more notations (hierarchical extension)

- *Label Hierarchy:* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; $\mathcal{A}(v)$, $\mathcal{N}(v)$, $\mathcal{D}(v)$ (ancestors, sisters, daughters)
- *Target Label Set:* $\mathcal{Y}=\{u \in \mathcal{V} : \nexists v \in \mathcal{V}, (u,v) \in \mathcal{E}\} \subseteq \mathcal{V}$
- *Kernel-based classifiers ($\mathcal{F}$):* $f(\mathbf{x}, y) = \langle \phi(\mathbf{x}), \mathbf{w}_y \rangle$
- *Decision function:* $g_f(\mathbf{x}, y) \geq 0$
  $(g_f(\mathbf{x}, y) = \min_{v \in \mathcal{P}(y)} (f(\mathbf{x}, v) - \max_{v' \in \mathcal{N}(v)} f(\mathbf{x}, v'))$
- *Complexity of function class:* Rademacher complexity (McDiarmid concentration inequality)

The "hierarchical vs flat" debate (3)

Theorem: [Babbar et al., 2013] Let $\mathcal{S} = ((\mathbf{x}^{(i)}, y^{(i)}))_{i=1}^{m}$ *be a dataset of m examples drawn i.i.d. according to a probability distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$*, and let* $\mathcal{A}$ *be a Lipschitz function with constant L dominating the* $0/1$ *loss; further let* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a PDS kernel and let* $\Phi : \mathcal{X} \to \mathbb{H}$ *be the associated feature mapping function. Assume that there exists* $R > 0$ *such that* $K(\mathbf{x}, \mathbf{x}) \leq R^2$ *for all* $\mathbf{x} \in \mathcal{X}$*. Then, for all* $1 > \delta > 0$*, with probability at least* $(1 - \delta)$ *the following hierarchical multiclass classification generalization bound holds for all* $g_f \in \mathcal{G}_{\mathcal{F}_B}$ *:*

$$\mathcal{E}(g_f) \leq \frac{1}{m}\sum_{i=1}^{m}\mathcal{A}(g_f(\mathbf{x}^{(i)}, y^{(i)})) + \frac{8BRL}{\sqrt{m}}\sum_{v \in V \setminus \mathcal{Y}} |\mathfrak{D}(v)|(|\mathfrak{D}(v)| - 1) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

where $|\mathfrak{D}(v)|$ *denotes the number of daughters of node* v*.*

Indicates which strategy to adopt on a given taxonomy and suggests ways to improve taxonomies (DMOZ, wikipedia, IPC)

The "hierarchical vs flat" debate (4)

Interpretation

- The complexity term privileges hierarchical structures whereas empirical error term privileges flat structures (error propagation), esp. when classes are well balanced
- Trade-off between empirical error and complexity terms

illustration

Dataset	# Tr.	# Test	# Classes	# Feat.	**CR**	**Emp ER**
LSHTC2-3	38,725	10,102	3,956	145,354	**0.004**	**2.65**
LSHTC2-4	27,924	7,026	2,544	123,953	**0.005**	**1.8**
LSHTC2-5	68,367	17,561	7,212	192,259	**0.002**	**2.12**
IPC	46,324	28,926	451	1,123,497	**0.02**	**12.27**

Table : **CR**: complexity ratio H/F; **Emp ER** emp. error ratio H/F

The "hierarchical vs flat" debate (5)

Validation through classification results

	LSHTC2-3		LSHTC2-4		LSHTC2-5		IPC	
	MLR	SVM	MLR	SVM	MLR	SVM	MLR	SVM
FL	0.528	0.535	0.497	0.501	0.542	0.547	**0.546**†	**0.446**†
RN	0.493	0.517	0.478	0.484	0.532	0.536	0.547	0.458
FH	**0.484**†	**0.498**†	**0.473**†	**0.476**†	**0.526**†	**0.527**†	0.552	0.465
PR-B	**0.481**	**0.495**	**0.466**	**0.465**	**0.522**	**0.522**	0.546	0.450
PR-M	**0.480**	**0.493**	**0.469**	**0.472**	**0.522**	**0.523**	0.544	0.450

Table : Error results across all datasets. Bold typeface is used for the best results. Statistical significance (using micro sign test (s-test) Yang and Liu [1999]) is denoted with † for p-value$<$0.05

Pruning a hierarchy (speed/accuracy trade-off)

1. Above result suggests that one can gain in accuracy by *flattening* a hierarchy
2. Above bound suggests ingredients that can be used to determine whether a node should be removed or not:
   - Number of categories, number of examples in different category sets
   - Dimension of feature space in different category sets
   - Confusion between categories
3. Learn a binary (meta-)classifier on some taxonomies (if pruning a node v leads to significantly better accuracy, v is labelled 1, 0 otherwise)
4. Apply meta-classifier to new taxonomies (number of nodes pruned depends on the policy - conservative or not)

Results on Classification Error

	LSHTC2-3		LSHTC2-4		LSHTC2-5		IPC	
	MLR	SVM	MLR	SVM	MLR	SVM	MLR	SVM
FL	0.528	0.535	0.497	0.501	0.542	0.547	0.546	**0.446**
RN	0.493	0.517	0.478	0.484	0.532	0.536	0.547	0.458
FH	0.484	0.498	0.473	0.476	0.526	0.527	0.552	0.465
PR-M	**0.480**†	**0.493**†	**0.469**†	**0.472**†	**0.522**†	**0.523**	**0.544**	0.450
PR-B	**0.481**	**0.495**	**0.466**	**0.465**	**0.522**	**0.522**	0.546	0.450

Table : Error results across all datasets, bold typeface is used for the best results. Statistical significance (using micro sign test(s-test) [Yang and Liu, 1999]) is denoted with † for p-value$<$0.05

- Pruning the taxonomy using the proposed meta-learning strategy improves classification accuracy
- Another strategy based on directly using the rademacher-based bound can also be applied for pruning

Space complexity of flat and hierarchical methods

For power law distributed category systems (incl. all textual category systems), the space complexity of the hierarchical approaches is lower than the one of the flat approaches

[Babbar et al., 2014a]

Space-complexity for Top-down Classification

- Using Heap's law, it can be shown that distribution of features exhibit a fit to power-law l, i.e., $d_{l,r} \approx d_{l,1} r^{-\beta_l}$
- Size of the top-down model is given by:

$$Size_{hier} = \sum_{l=1}^{L-1} \sum_{r=1}^{B_l} b_{l,r} d_{l,r} \approx \sum_{l=1}^{L-1} \sum_{r=1}^{B_l} b_{l,r} d_{l,1} r^{-\beta_l}$$

where $b_{l,r}$ represent the branching factor for the r-th ranked category, and B_l the total number of categories at level l.

Proposition ([Babbar et al., 2014a])

For a hierarchy of categories of depth L and K leaves, let $\beta = \min_{1 \leq l \leq L} \beta_l$ and $b = \max_{l,r} b_{l,r}$. Then,

$$For\ \beta > 1, if\ \beta > \frac{K}{K - b(L-1)} (> 1),\ then\ \ Size_{hier} < Size_{flat}$$

$$For\ 0 < \beta < 1, if\ \frac{b^{(L-1)(1-\beta)} - 1}{(b^{(1-\beta)} - 1)} < \frac{1-\beta}{b} K, then\ Size_{hier} < Size_{flat}$$

Datasets and Space-complexities

Dataset	Training/Test	Categories	Features	Tree Depth
LSHTC1-large	93,805/34,880	12,294	347,255	6
LSHTC2-a	25,310/6,441	1,789	145,859	6
LSHTC2-b	36,834/9,605	3,672	145,354	6
IPC	46,324/28,926	451	1,123,497	4

Dataset	$Size_{hier}$	$Size_{Flat}$	β	b	$\bigtriangledown$
LSHTC1-large	**2.8**	90.0	1.62	344	1.12
LSHTC2-a	**0.46**	5.4	1.35	55	1.14
LSHTC2-b	**1.1**	11.9	1.53	77	1.09
IPC	**3.6**	10.5	2.03	34	1.17

Table : Model size (in GB) for flat and hierarchical models, $\bigtriangledown$ refers to the quantity $\frac{K}{K-b(L-1)}$

Challenges in Large-scale Classification

- We addressed two challenges
  - Which approach to use, flat or hierarchical? Which hierarchy?
    $\rightarrow$ Presented theoretical explanation of when flat or hierarchical is to be preferred
    $\rightarrow$ Taxonomy adaptation through node pruning to select *better* taxonomies
  - Scale of datasets: is hierarchical really faster?
    $\rightarrow$ Yes (better time and space complexities)

- Third challenge: Dealing with rare categories

Rare category detection challenge

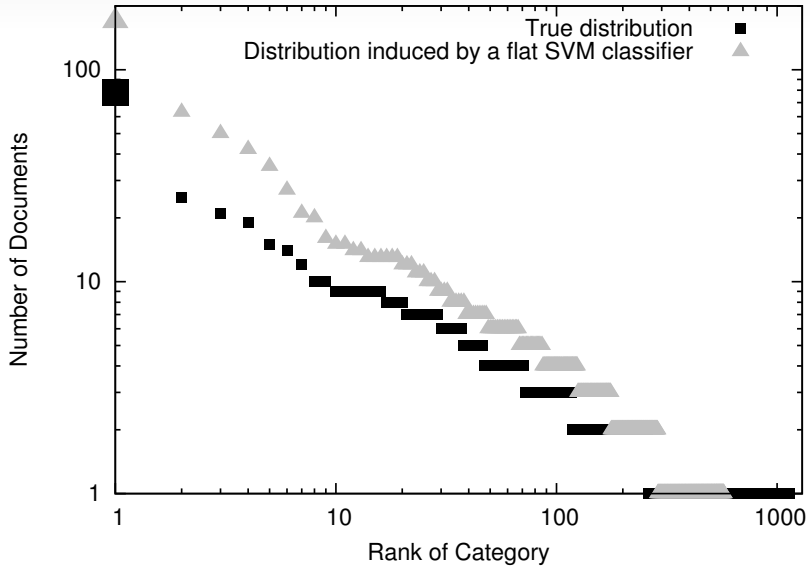


Figure : Comparison of true and induced test-set distributions

- Left part: induced distribution higher
  $\Longrightarrow$ high false positive rate for large categories
- Right part: tail of induced dist. too short
  $\Longrightarrow$ high false negative rate for small categories
- Out of 1139 classes, only 574 are discovered in the test set
  $\Longrightarrow$ low values of Macro and Micro F1

Soft-thresholding for prediction with rare categories

1. Can one quantify the difference in the distribution induced by a classifier and the true distribution on the test set?
2. Can this be related to an upper bound on the test-set accuracy of any classifier C, using the training set only?

Theorem ([Babbar et al., 2014b])

Let $S = \{(\mathbf{x}^{(j)}, y^{(j)})\}_{j=1}^{M}$ be the test set generated i.i.d. from $\mathcal{D}$. Let M_ℓ^C be the number of examples in S assigned to category y_ℓ by the classifier C which is trained on S_{train}. Then the following bound on the accuracy of C over S, denoted by $Acc(C)$, holds with high probability :

$$Acc(C) \leq \frac{1}{|S|} \sum_{\ell=1}^{|\mathcal{Y}|} \min\{(\hat{p}_{y_\ell} \times |S|), M_\ell^C\} \triangleq B(Acc(C))$$

where $\hat{p}_{y_\ell}$ denotes the estimate on the prior probability of the category y_ℓ in the training set.

Algorithm to achieve higher bound value

Input: Training data S_{train} and Test data S_{test}
  Learn Multiclass SVM (Crammer-Singer)
  for each test instance $\mathbf{x} \in S_{test}$ **do**
    Predict posterior probabilities $(\hat{p}_{y_l}|\mathbf{x}), \forall 1 \leq l \leq |\mathcal{Y}|$
    if $pred(\mathbf{x})$ is true **then**
      Create *instantaneous training set* t (odd) times
      To distinguish $\{y_{r1}, y_{r2}\}$, learn t binary classifiers
      Re-predict instance $\mathbf{x}$ with each binary classifier
      Output from $\{y_{r1}, y_{r2}\}$ the one with majority votes
    else
      Output category $\arg\max_{y_l \in \mathcal{Y}}(\hat{p}_{y_l}|\mathbf{x})$
    end if
  end for
  return Labels $\forall \mathbf{x} \in S_{test}$

$pred(\mathbf{x})$ is true iff $(\hat{p}_{y_{r1}}|\mathbf{x}) - (\hat{p}_{y_{r2}}|\mathbf{x}) \leq \Delta \quad \&\& \quad N_{y_{r1}}/N_{y_{r2}} \geq R$

Empirical Evaluation

Dataset	Training/Test instances	Categories $\lvert\mathcal{Y}\rvert$	Features d
DMOZ-2010-s	4,463/1858	1,139	51,033
DMOZ-2010-l	128,710/34,880	12,294	381,580
DMOZ-2012	383,408/103,435	11,947	348,548

Table : Datasets and their statistics

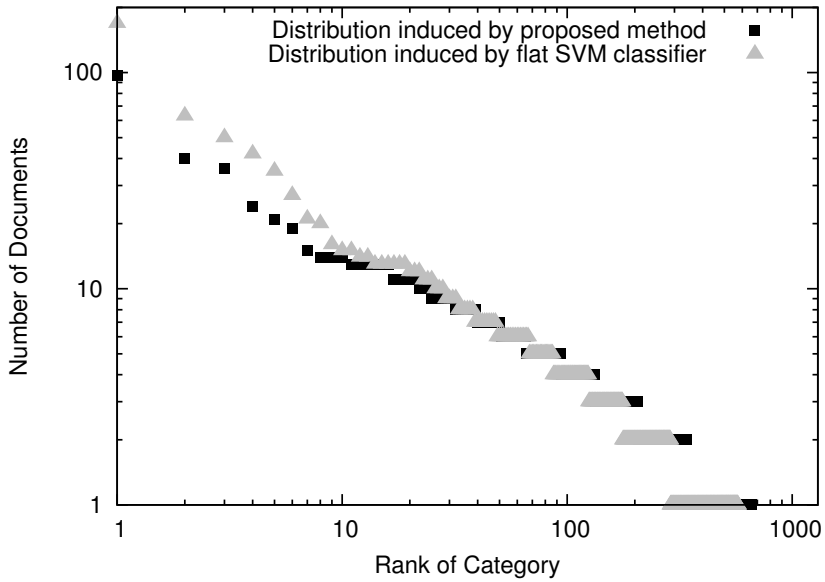
- Datasets are in the form of libSVM format with single label for each training instance
- Feature set size equals the size of the size of the vocabulary

Empirical Results

Dataset	Proposed Algorithm	HR-SVM	CS-SVM
DMOZ-2010-s			
Micro-F1	**47.36**†	45.31	45.15
Macro-F1	**32.91**$^{\downarrow}$	28.94	29.40
B(Acc(C))	**0.71**	0.63	0.64
Categories detected	**658**	570	574
Training Time	**1.1x**	1.7x	1x
DMOZ-2010-l			
Micro-F1	**46.67**†	46.02	45.82
Macro-F1	**34.65**$^{\downarrow}$	33.12	32.63
B(Acc(C))	**0.77**	0.73	0.72
Categories detected	**8523**	8102	8039
Training Time	**1.1x**	1.6x	1x
DMOZ-2012			
Micro-F1	**57.78**†	57.17	56.44
Macro-F1	**34.15**$^{\downarrow}$	33.05	31.59
B(Acc(C))	**0.76**	0.72	0.70
Categories detected	**8220**	7965	7882
Training Time	**1.1x**	1.6x	1x

Table : The significance-test results (using micro sign test (s-test) and macro t-test) are denoted for a p-value less than 1%.

Comparison of induced distributions



- $B(Acc(C)) = 0.71 > 0.64$ (for Flat Classifier)
- Out of 1139, # detected classes $= 658 > 574$ (for Flat Classifier)
- Re-prediction required for approx. 10% of test instances, hence does not impact it adversely

Conclusion

- Flat versus Top-down classification
  - Generalization error bounds to theoretically explain the performance of various methods
  - Hierarchy pruning strategy for improvement in classification accuracy
  - Proof that hierarchical approaches *really* faster than flat ones
- Classification with rare categories
  - Soft-thresholding based algorithm for classification with rare categories
- Datasets available from LSHTC challenges (http://lshtc.iit.demokritos.gr/) and BioASQ (http://bioasq.org/)

Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih-Reza Amini. On flat versus hierarchical classification in large-scale taxonomies. In **Advances in Neural Information Processing Systems**, pages 1824–1832, 2013.

Rohit Babbar, Cornelia Metzig, Ioannis Partalas, Eric Gaussier, and Massih-Reza Amini. On power law distributions in large-scale taxonomies. **ACM SIGKDD Explorations Newsletter**, 16(1):47–56, 2014a.

Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih-reza Amini. Re-ranking approach to classification in large-scale power-law distributed category systems. In **Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval**, pages 1059–1062. ACM, 2014b.

Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In **Neural Information Processing Systems**, pages 163–171, 2010.

Paul N. Bennett and Nam Nguyen. Refined experts: improving classification in large taxonomies. In **Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval**, pages 11–18, 2009.

Lijuan Cai and Thomas Hofmann. Hierarchical document categorization with support vector machines. In **Proceedings of the thirteenth ACM international conference on Information and knowledge management**, pages 78–87, 2004.

Ofer Dekel. Distribution-calibrated hierarchical classification. In **Advances in Neural Information Processing Systems 22**, pages 450–458. 2009.

Tianshi Gao and Daphne Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In **IEEE International Conference on Computer Vision (ICCV)**, pages 2072–2079, 2011.

Siddharth Gopal, Yiming Yang, Bing Bai, and Alexandru Niculescu-Mizil. Bayesian models for large-scale hierarchical classification. In **Neural Information Processing Systems**, 2012.

Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. Support vector machines classification with a very large-scale taxonomy. **SIGKDD**, 2005.

Hassan Malik. Improving hierarchical svms by hierarchy flattening and lazy classification. In **1st Pascal Workshop on Large Scale Hierarchical Classification**, 2009.

Florent Perronnin, Zeynep Akata, Zaïd Harchaoui, and Cordelia Schmid. Towards good practice in large-scale learning for image classification. In **Computer Vision and Pattern Recognition**, pages 3482–3489, 2012.

Yiming Yang and Xin Liu. A re-examination of text categorization methods. In **Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval**, pages 42–49. ACM, 1999.

Thank you!