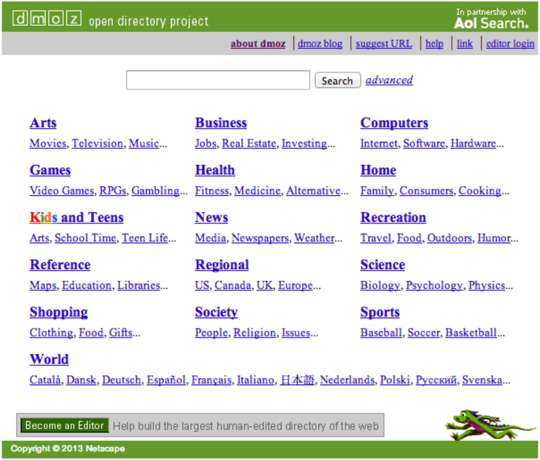
On Flat versus Hierarchical Classification in Large-Scale Taxonomies

R. Babbar, I. Partalas, É. Gaussier, M.-R. Amini

Gargantua (CNRS Mastodons) - November the 26^{th}, 2013

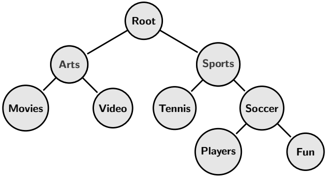
Large-scale Hierarchical Classification in Practice

- Directory Mozilla



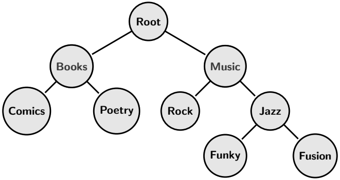
5,292,731 sites - 99,941 editors - over 1,020,828 categories

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



Approaches for Large Scale Hierarchical Classification (LSHC)

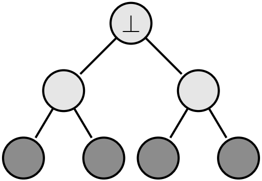
- Hierarchical
  - Top-down - solve individual classification problems at every node
  - Big-bang - solve the problem at once for entire tree
- Flat - ignore the taxonomy structure *altogether*
- Flattening Approaches in LSHTC
  - Somewhat arbitrary as they flatten entire layers
  - Not quite clear which layers to flatten when taxonomy are much deeper with 10-15 levels



Key Challenges in LSHC

- **How reliable is the given hierarchical structure ?**
  - Arbitrariness in taxonomy creation based on personal biases and choices
  - Other sources of *noise* include imbalanced nature of hierarchies
- **Which Approach - Flat or Hierarchical ?**
  - Lack of clarity on exploiting the hierarchical structure of categories
  - Speed versus Accuracy trade-off

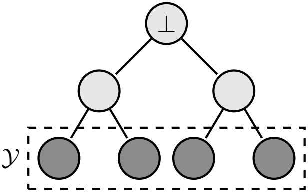
Hierarchical Rademacher-based Generalization Bound



- hierarchy of classes $\mathcal{H} = (V, E)$ is defined in the form of a rooted tree, with a root $\perp$ and a parent relationship π
- Nodes at the leaf level, $\mathcal{Y} = \{y \in V : \nexists v \in V, (y, v) \in E\} \subset V$, constitute the set of target classes
- $\forall v \in V \setminus \{\perp\}$, we define the set of its sisters $\mathfrak{S}(v) = \{v' \in V \setminus \{\perp\}; v \neq v' \wedge \pi(v) = \pi(v')\}$ and its daughters $\mathfrak{D}(v) = \{v' \in V \setminus \{\perp\}; \pi(v') = v\}$
- $\forall y \in \mathcal{Y}, \mathfrak{P}(y) = \{v_1^y, \dots, v_{k_y}^y; v_1^y = \pi(y) \wedge \forall l \in \{1, \dots, k_y - 1\}, v_{l+1}^y = \pi(v_l^y) \wedge \pi(v_{k_y}^y) = \perp\}$

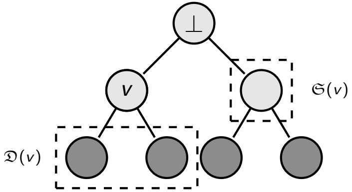
LIG

Hierarchical Rademacher-based Generalization Bound



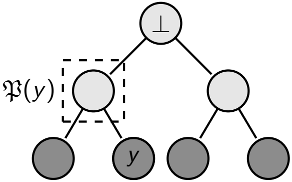
- hierarchy of classes $\mathcal{H}=(V,E)$ is defined in the form of a rooted tree, with a root $\perp$ and a parent relationship π
- Nodes at the leaf level, $\mathcal{Y}=\{y\in V:\nexists v\in V,(y,v)\in E\}\subset V$, constitute the set of target classes
- $\forall v\in V\setminus\{\perp\}$, we define the set of its sisters $\mathfrak{S}(v)=\{v'\in V\setminus\{\perp\};v\neq v'\wedge\pi(v)=\pi(v')\}$ and its daughters $\mathfrak{D}(v)=\{v'\in V\setminus\{\perp\};\pi(v')=v\}$
- $\forall y\in\mathcal{Y},\mathfrak{P}(y)=\{v_1^y,\dots,v_{k_y}^y;v_1^y=\pi(y)\wedge\forall l\in\{1,\dots,k_y-1\},v_{l+1}^y=\pi(v_l^y)\wedge\pi(v_{k_y}^y)=\perp\}$

Hierarchical Rademacher-based Generalization Bound



- hierarchy of classes $\mathcal{H} = (V, E)$ is defined in the form of a rooted tree, with a root $\perp$ and a parent relationship π
- Nodes at the leaf level, $\mathcal{Y} = \{y \in V : \nexists v \in V, (y, v) \in E\} \subset V$, constitute the set of target classes
- $\forall v \in V \setminus \{\perp\}$, we define the set of its sisters $\mathfrak{S}(v) = \{v' \in V \setminus \{\perp\}; v \neq v' \wedge \pi(v) = \pi(v')\}$ and its daughters $\mathfrak{D}(v) = \{v' \in V \setminus \{\perp\}; \pi(v') = v\}$
- $\forall y \in \mathcal{Y}, \mathfrak{P}(y) = \{v_1^y, \dots, v_{k_y}^y; v_1^y = \pi(y) \wedge \forall l \in \{1, \dots, k_y - 1\}, v_{l+1}^y = \pi(v_l^y) \wedge \pi(v_{k_y}^y) = \perp\}$

Hierarchical Rademacher-based Generalization Bound



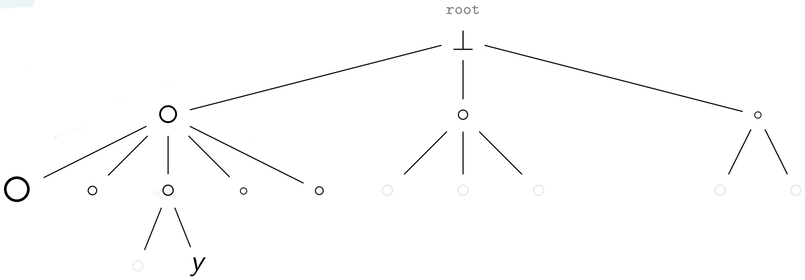
- hierarchy of classes $\mathcal{H} = (V, E)$ is defined in the form of a rooted tree, with a root $\perp$ and a parent relationship π
- Nodes at the leaf level, $\mathcal{Y} = \{y \in V : \nexists v \in V, (y, v) \in E\} \subset V$, constitute the set of target classes
- $\forall v \in V \setminus \{\perp\}$, we define the set of its sisters $\mathfrak{S}(v) = \{v' \in V \setminus \{\perp\}; v \neq v' \wedge \pi(v) = \pi(v')\}$ and its daughters $\mathfrak{D}(v) = \{v' \in V \setminus \{\perp\}; \pi(v') = v\}$
- $\forall y \in \mathcal{Y}, \mathfrak{P}(y) = \{v_1^y, \ldots, v_{k_y}^y; v_1^y = \pi(y) \wedge \forall l \in \{1, \ldots, k_y - 1\}, v_{l+1}^y = \pi(v_l^y) \wedge \pi(v_{k_y}^y) = \perp\}$

Hierarchical Rademacher-based Generalization Bound

- We consider a top-down hierarchical classification strategy ;
- Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a PDS kernel and let $\Phi : \mathcal{X} \rightarrow \mathbb{H}$ be the associated feature mapping function, we suppose that there exists $R > 0$ such that $K(\mathbf{x}, \mathbf{x}) \leq R^2$ for all $\mathbf{x} \in \mathcal{X}$;
- We consider the class of functions $f \in \mathcal{F}_B = \{f : (\mathbf{x}, v) \in \mathcal{X} \times V \mapsto \langle \Phi(\mathbf{x}), \mathbf{w}_v \rangle \mid \mathbf{W} = (w_1 \ldots, w_{|V|}), ||\mathbf{W}||_{\mathbb{H}} \leq B\}$;
- An exemple $(\mathbf{x}, y)$ is misclassified iff by $f \in \mathcal{F}_B$

$$\min_{v \in \mathfrak{P}(y)} \left(f(\mathbf{x}, v) - \max_{v' \in \mathfrak{S}(v)} f(\mathbf{x}, v')\right) \leq 0$$

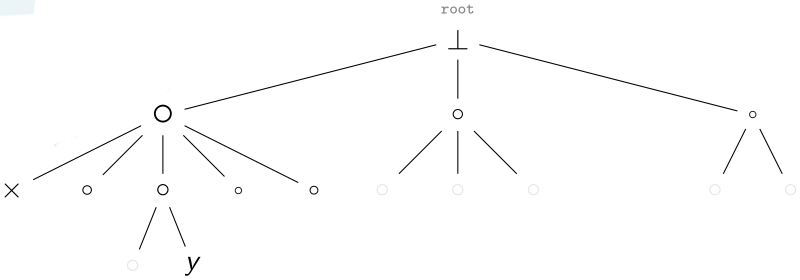
Hierarchical Rademacher-based Generalization Bound



- An exemple $(\mathbf{x}, y)$ is misclassified iff by $f \in \mathcal{F}_B$

$$\min_{v \in \mathfrak{P}(y)} \left(f(\mathbf{x}, v) - \max_{v' \in \mathfrak{S}(v)} f(\mathbf{x}, v')\right) \leq 0$$

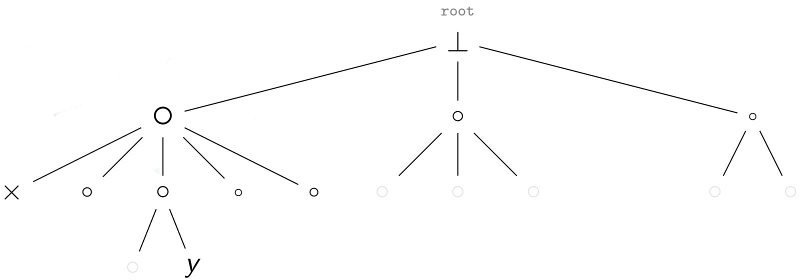
Hierarchical Rademacher-based Generalization Bound



- An exemple $(\mathbf{x}, y)$ is misclassified iff by $f \in \mathcal{F}_B$

$$\min_{v \in \mathfrak{P}(y)} \underbrace{\left(f(\mathbf{x}, v) - \max_{v' \in \mathfrak{S}(v)} f(\mathbf{x}, v')\right)}_{\text{multi-class margin}} \leq 0$$

Hierarchical Rademacher-based Generalization Bound



- Top-Down hierarchical techniques suffer from error propagation, but imbalancement harms less as it does for flat approaches ⇒ a generalization bound to study these effects.

Hierarchical Rademacher-based Generalization Bound

Theorem

Let $\mathcal{S} = ((\mathbf{x}^{(i)}, y^{(i)}))_{i=1}^m$ *an i.i.d. training set drawn 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* $R > 0$ *such that* $K(\mathbf{x}, \mathbf{x}) \leq R^2$ *for all* $\mathbf{x} \in \mathcal{X}$*. Then, with probability at least* $(1 - \delta)$ *the following bound holds for all*
$f \in \mathcal{F}_B = \{f : (\mathbf{x}, v) \in \mathcal{X} \times V \mapsto \langle \Phi(\mathbf{x}), \mathbf{w}_v \rangle \mid \mathbf{W} = (w_1 \ldots, w_{|V|}), ||\mathbf{W}||_{\mathbb{H}} \leq B\}$:

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

where $\mathcal{G}_{\mathcal{F}_B} = \{g_f : (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mapsto \min_{v \in \mathfrak{P}(y)}(f(\mathbf{x}, v) - \max_{v' \in \mathfrak{S}(v)} f(\mathbf{x}, v')) \mid f \in \mathcal{F}_B\}$ *and* $|\mathfrak{D}(v)|$ *denotes the number of daughters of node* v*.*

Extension of an existing result for flat multi-class classification

Theorem (Guermeur, 2007)

Let $\mathcal{S} = ((\mathbf{x}^{(i)}, y^{(i)}))_{i=1}^m$ *an i.i.d. training set drawn 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* $R > 0$ *such that* $K(\mathbf{x}, \mathbf{x}) \leq R^2$ *for all* $\mathbf{x} \in \mathcal{X}$*. Then, with probability at least* $(1 - \delta)$ *the following bound holds for all* $f \in \mathcal{F}_B = \{f : (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mapsto \langle \Phi(\mathbf{x}), \mathbf{w}_y \rangle \mid \mathbf{W} = (w_1 \ldots, w_{|\mathcal{Y}|}), ||\mathbf{W}||_{\mathbb{H}} \leq B\}$*:*

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

where

$\mathcal{G}_{\mathcal{F}_B} = \{g_f : (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mapsto (f(\mathbf{x}, y) - \max_{y' \in \mathcal{Y} \setminus \{y\}} f(\mathbf{x}, y')) \mid f \in \mathcal{F}_B\}$.

Trade-offs in Flat versus Top-down techniques

- Empirical Error vs Error due to Complexity
  - Empirical Error is higher in top-down method due to series of decisions to be made in cascade
  - Complexity Term dominated by $|\mathfrak{D}(v)|(|\mathfrak{D}(v)|-1)$ is lower in top-down methods
- Degree of imbalance in training data
  - *Imbalanced data* (DMOZ) flat method suffers but top-down method can counter it better and also has lower error due to complexity term, and hence preferable
  - *Balanced data* (IPC with sample complexity bounds satisfied for most classes), flat method should be preferred
- Motivates Hierarchy Pruning to achieve the trade-off between error terms

Empirical study

Dataset	# Tr.	# Test	# Classes	# Feat.	CR	Error ratio
LSHTC2-1	25,310	6,441	1,789	145,859	0.008	1.24
LSHTC2-2	50,558	13,057	4,787	271,557	0.003	1.32
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

- Complexity Ratio (CR) defined as $\sum_{v \in V \setminus \mathcal{Y}} |\mathfrak{D}(v)|(|\mathfrak{D}(v)|-1)/|\mathcal{Y}|(|\mathcal{Y}|-1)$ is in favour of Top-down methods
- Empirical error ratio favours Flat approaches

Asymptotic Approximation Error Bounds

Relationship between the generalization error of a trained Multiclass Logistic Regression classifier and its asymptotic version.

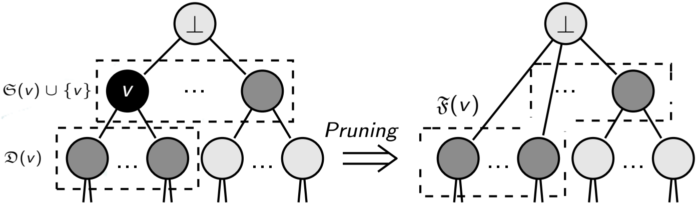
Theorem

For a multi-class classification problem in d dimensional feature space with a training set of size m, $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{m}$, $\mathbf{x}^{(i)} \in \mathcal{X}$, $y^{(i)} \in \mathcal{Y}$, sampled i.i.d. from a probability distribution $\mathcal{D}$, let h_m and h_∞ denote the multiclass logistic regression classifiers learned from a training set of finite size m and its asymptotic version respectively, and let $\mathcal{E}(h_m)$ and $\mathcal{E}(h_\infty)$ be their generalization errors. Then, with probability at least $(1-\delta)$ we have:

$$\mathcal{E}(h_m) \leq \mathcal{E}(h_\infty) + G_{\mathcal{Y}}\left(d\sqrt{\frac{R|\mathcal{Y}|\sigma_0}{\delta m}}\right) \tag{3}$$

where $\sqrt{R}$ is a bound on the function $\exp(\beta_0^y + \sum_{j=1}^{d} \beta_j^y x_j)$, $\forall \mathbf{x} \in \mathcal{X}$ and $\forall y \in \mathcal{Y}$, and σ_0 is a constant and $G_{\mathcal{Y}}(\tau)$ is a measure of confusion and increasing function of τ.

Hierarchy Pruning via Meta-learning



- The bounds (1) and (2) are not directly exploitable but indicate crucial (meta)features which control the generalization error
- We train a meta-classifier on a sub-hierarchy with meta-instances
- Meta-features include values of KL-divergence, category sizes, feature-set sizes etc. before and after pruning.
- For meta-classifier, applied AdaBoost with Random forest as base-classifier with different number of trees and depths

Experimental Setup

Datasets used : **LSHTC2-1** and **LSHTC2-2** used for training Meta-classifier

Dataset	# Tr.	# Test	# Classes	# Feat.	CR	Error ratio
LSHTC2-1	25,310	6,441	1,789	145,859	0.008	1.24
LSHTC2-2	50,558	13,057	4,787	271,557	0.003	1.32
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 : Datasets used, the complexity ratio of hierarchical over the flat case ($\sum_{v \in V \setminus \mathcal{Y}} |\mathfrak{D}(v)|(|\mathfrak{D}(v)|-1)/|\mathcal{Y}|(|\mathcal{Y}|-1)$), the ratio of empirical error for hierarchical over flat models is shown in last two columns

- Complexity Ratio is in favour of Top-down methods
- Empirical error ratio favours Flat approaches

Error results

	LSHTC2-3			LSHTC2-4			IPC		
	MNB	MLR	SVM	MNB	MLR	SVM	MNB	MLR	SVM
FL	.729$^{\downarrow\downarrow}$	.528$^{\downarrow\downarrow}$	.535$^{\downarrow\downarrow}$	.848$^{\downarrow\downarrow}$	.497$^{\downarrow\downarrow}$	.501$^{\downarrow\downarrow}$	.671$^{\downarrow\downarrow}$	.546	**.446**
RN	.612$^{\downarrow\downarrow}$	.493$^{\downarrow\downarrow}$	.517$^{\downarrow\downarrow}$	.704$^{\downarrow\downarrow}$	.478$^{\downarrow\downarrow}$	.484$^{\downarrow\downarrow}$	.642$^{\downarrow\downarrow}$	.547$^{\downarrow}$	.458$^{\downarrow\downarrow}$
FH	.619$^{\downarrow\downarrow}$	.484$^{\downarrow\downarrow}$	.498$^{\downarrow\downarrow}$	.682$^{\downarrow}$	.473$^{\downarrow\downarrow}$	.476$^{\downarrow}$	.643$^{\downarrow\downarrow}$	.552$^{\downarrow}$	.465$^{\downarrow\downarrow}$
PR	**.613**	**.480**	**.493**	**.677**	**.469**	**.472**	**.639**	**.544**	.450

- Top-down method better than Flat approach on LSHTC datasets with a large fraction of *rare categories* but not on IPC dataset
- Pruning via meta-learning improves classification accuracy

- Conclusion
  - Generalization error bounds for multi-class hierarchical classifiers to theoretically explain the performance of flat and hierarchical methods
  - Proposed a hierarchy pruning strategy for improvement in classification accuracy
- Future Work
  - Use the theoretical framework for building taxonomies
  - Explore other frameworks for hierarchy pruning