Multi-class to Binary reduction of Large-scale classification Problems

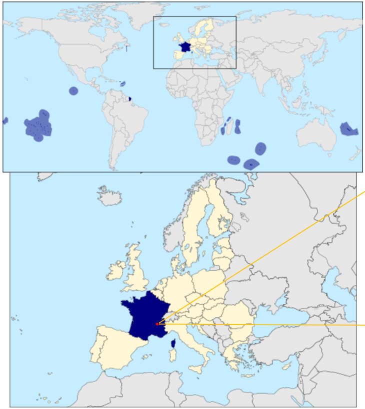
Massih-Reza Amini

Joint work with Bikash Joshi, Ioannis Partalas and Franck Iutzeler

University Grenoble Alps

October the 5^{th}, 2016

Grenoble, Capital of the Alps



Grenoble, France

Applied Mathematics and Computer Science Building

University Grenoble Alps

Computer Science and Applied Mathematics Laboratories

Classical learning framework

We consider an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and an output space $\mathcal{Y}$.

Hypothesis : Pairs of examples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ are *identically* and *independently* distributed (`i.i.d`) with respect to a fixed but unknown distribution $\mathcal{D}$.

Sampling : We observe a sequence of m pairs of examples $(\mathbf{x}_i, y_i)$ generated `i.i.d` with respect to $\mathcal{D}$.

Goal : Find a function $g : \mathcal{X} \to \mathcal{Y}$, which belongs to a class of functions $\mathcal{G}$, which predicts the output y of a new observation $\mathbf{x}$ such that :

$$\mathbb{P}(g(\mathbf{x}) \neq y) \text{ is the lowest possible.}$$

New challenges with Emerging Applications

We consider an input space $\mathcal{X} \subseteq \mathbb{R}^d$ $(d >> 1)$ and an output space $\mathcal{Y}, |\mathcal{Y}| >> 1$.

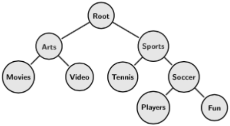
Pairs of examples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ are *identically* and *indepe* … ted but ur …

Samp … examp …

Goal … a class … new observ …

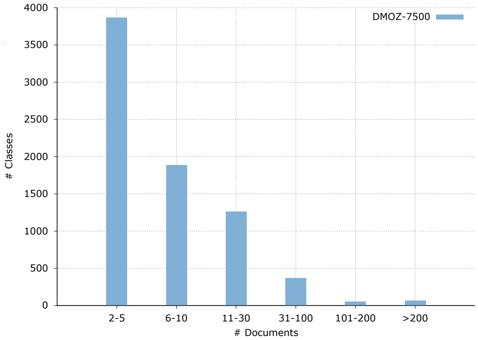
5,292,731 sites - 99,941 editors - over 1,020,828 categories

- 5×10^6 sites
- 10^6 categories
- 10^5 editors
- imbalanced nature of hierarchies
- Arbitrariness in taxonomy creation - personal biases



Large-scale classification : power law distribution of classes

Collection	K	d
DMOZ	7500	594158



Multiclass classification approaches

- Uncombined approaches, i.e. MSVM or MLP. The number of parameters, M, is at least $O(K \times d)$.

- Combined approaches based on binary classification :
  - One-Vs-one - $M \geq O(K^2 \times d)$
  - One-Vs-Rest - $M \geq O(K \times d)$

- For $K >> 1$ and $d >> 1$ traditional approaches do not pass the scale.

Outline

- Motivation
- Learning objective and reduction strategy
- Experimental results
- Conclusion

Outline

- Motivation
- Learning objective and reduction strategy
- Experimental results
- Conclusion

Learning objective

- Large-scale multiclass classification,
  - Hypothesis : Observations $\mathbf{x}^y = (x, y) \in \mathcal{X} \times \mathcal{Y}$ are i.i.d with respect to a distribution $\mathcal{D}$,
  - For a class of $\mathcal{H} = \{h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$, a ranking instanstaneous loss $h \in \mathcal{H}$ over an example $\mathbf{x}^y$ by :

$$e(h, \mathbf{x}^y) = \frac{1}{K-1} \sum_{y' \in \mathcal{Y} \setminus \{y\}} \mathbb{1}_{h(\mathbf{x}^y) \leq h(\mathbf{x}^{y'})},$$

  - The aim is to find a function $h \in \mathcal{H}$ that minimizes the generalization error $L(h)$:

$$L(h) = \mathbb{E}_{\mathbf{x}^y \sim \mathcal{D}} \left[e(h, \mathbf{x}^y) \right].$$

  - Empirical error of a function $h \in \mathcal{H}$ over a training set $\mathcal{S} = \left(\mathbf{x}_i^{y_i}\right)_{i=1}^{m}$ is

$$\hat{L}_m(h, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^{m} e(h, \mathbf{x}_i^{y_i})$$

Reduction strategy

- Consider the empirical loss

$$\begin{aligned}\hat{L}_m(h,\mathcal{S}) &= \frac{1}{m(K-1)}\sum_{i=1}^{m}\sum_{y'\in\mathcal{Y}\setminus\{y_i\}}\mathbb{1}_{h(\mathbf{x}_i^{y_i})\leq h(\mathbf{x}_i^{y'})}\\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\tilde{y}_i g(\mathbf{Z_i})\leq \mathbf{0}}}_{L_n^T(g,T(S))}\end{aligned}$$

where $n = m(K-1)$, Z_i is a pair of couples costituted by a couple of example and its class and the couple corresponding to the example and another class, $\tilde{y}_i = 1$ if the first couple in Z_i is the true couple and -1 otherwise, and $g(\mathbf{x}^y, \mathbf{x}^{y'}) = h(\mathbf{x}^y) - h(\mathbf{x}^{y'})$.

Reduction strategy for the class of linear functions

Input: Labeled training set $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m$;
A binary classifier $\mathcal{A}$;
Initialize
$T(S) \leftarrow \emptyset$;
for $i = 1..m$ **do**
  for $k = 1..K$ **do**
    if $y_i > k$ **then**
      $T(S) \leftarrow \{(\Phi(\mathbf{x}_i^{y_i}) - \Phi(\mathbf{x}_i^k), +1)\}$
    end
    if $y_i < k$ **then**
      $T(S) \leftarrow \{(\Phi(\mathbf{x}_i^k) - \Phi(\mathbf{x}_i^{y_i}), -1)\}$
    end
  end
end
Learn $\mathcal{A}$ on $T(S)$

Reduction strategy for the class of linear functions

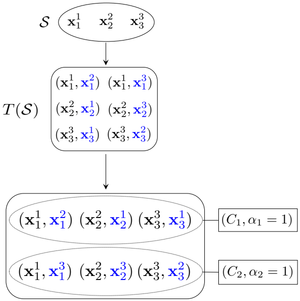
Input: Labeled training set $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m$;
A binary classifier $\mathcal{A}$;
Initialize
$T(S) \leftarrow \emptyset$;
for $i = 1..m$ **do**
  for $k = 1..K$ **do**
    if $y_i > k$ **then**
      $T(S) \leftarrow \{(\Phi(\mathbf{x}_i^{y_i}) - \Phi(\mathbf{x}_i^k), +1)\}$
    end
    if $y_i < k$ **then**
      $T(S) \leftarrow \{(\Phi(\mathbf{x}_i^k) - \Phi(\mathbf{x}_i^{y_i}), -1)\}$
    end
  end
end
Learn $\mathcal{A}$ on $T(S)$

Problems :

- How to define $\Phi(\mathbf{x}^y)$,
- Consistency of the ERM principle with interdependant data.

Consistency of the ERM principle with interdependant data

- Different statistical tools for extending concentration inequalities to the case of interdependent data,
- tools based on colorable graphs proposed by (Janson, 2004) [1].



1. S. Janson. Large deviations for sums of partly dependent random variables. Random Structures and Algorithms, 24(3) :234–248, 2004.

Theorem (Bikash et al. 2015)

Let $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ *be a training set constituted of* m *examples generated i.i.d. with respect to a probability distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$ *and* $T(\mathcal{S}) = ((\boldsymbol{Z}_i, \tilde{y}_i))_{i=1}^n \in (\mathcal{Z} \times \{-1, 1\})^n$ *the transformed set obtained with application* T. *Let* $\kappa : \mathcal{Z} \to \mathbb{R}$ *by a PSD kernel, and* $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{H}$ *the associated mapping function. For all* $1 > \delta > 0$, *and all* $g_w \in \mathcal{G}_B = \{\mathbf{x} \mapsto \langle \boldsymbol{w}, \Phi(\mathbf{x}) \rangle \mid ||\boldsymbol{w}|| \leq B\}$ *with probability at least* $(1 - \delta)$ *over* $T(\mathcal{S})$ *we have then :*

$$L^T(g_w) \leq \epsilon_n^T(g_w, T(\mathcal{S})) + \frac{2B\mathfrak{G}(T(\mathcal{S}))}{m\sqrt{K-1}} + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}} \tag{1}$$

where $\epsilon_n^T(g_w, T(\mathcal{S})) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\tilde{y}_i g_w(\boldsymbol{Z}_i))$ *with a surrogate Hinge loss* $\mathcal{L} : t \mapsto \min(1, \max(1 - t, 0))$, $L^T(g_w) = \mathbb{E}_{T(\mathcal{S})}[L_n^T(g_w, T(\mathcal{S}))]$ *et* $\mathfrak{G}(T(\mathcal{S})) = \sqrt{\sum_{i=1}^n d_\kappa(\boldsymbol{Z}_i)}$ *with*

$$d_\kappa(\mathbf{x}^y, \mathbf{x}^{y'}) = \kappa(\mathbf{x}^y, \mathbf{x}^y) + \kappa(\mathbf{x}^{y'}, \mathbf{x}^{y'}) - 2\kappa(\mathbf{x}^y, \mathbf{x}^{y'})$$

Key Features of Algorithm

- Data dependent bound :
  If the feature representation of (x,y) pairs is independent of original dimension, then : $\mathfrak{G}(T(S)) \leq \sqrt{n \times Constant} \approx \sqrt{m \times (K-1) \times Constant}$ and the convergence rate is of order $O(\frac{1}{\sqrt{m}})$.

- Non-trivial joint feature representation (example-class pair)

- Same for any number of class
- Same parameter vector for all classes

Outline

- Motivation
- Learning objective and reduction strategy
- Experimental results
- Conclusion

Feature representation $\Phi(\mathbf{x}^y)$

Features	
1. $\sum_{t \in y \cap x} \ln(1 + y_t)$	2. $\sum_{t \in y \cap x} \ln(1 + \frac{l_{\mathcal{S}}}{\mathcal{S}_t})$
3. $\sum_{t \in y \cap x} I_t$	4. $\sum_{t \in y \cap x} \ln(1 + \frac{y_t}{\lvert y \rvert})$
5. $\sum_{t \in y \cap x} \ln(1 + \frac{y_t}{\lvert y \rvert}.I_t)$	6. $\sum_{t \in y \cap x} \ln(1 + \frac{y_t}{\lvert y \rvert}.\frac{l_{\mathcal{S}}}{\mathcal{S}_t})$
7. $\sum_{t \in y \cap x} 1$	8. $\sum_{t \in y \cap x} \frac{y_t}{\lvert y \rvert}.I_t$
9. $d_1(\mathbf{x}^y)$	10. $d_2(\mathbf{x}^y)$

- x_t : number of occurrences of terme t in document x,
- $\mathcal{V}$: Number of distinct terms in $\mathcal{S}$,
- $y_t = \sum_{x \in y} x_t$, $\lvert y \rvert = \sum_{t \in \mathcal{V}} y_t$, $\mathcal{S}_t = \sum_{x \in \mathcal{S}} x_t$, $l_{\mathcal{S}} = \sum_{t \in \mathcal{V}} \mathcal{S}_t$.
- I_t : idf of the terme t,

Experimental results on text classification

Collection	K	d	m	Test size
DMOZ	7500	594158	394756	104263
WIKIPEDIA	7500	346299	456886	81262

$$K \times d = O(10^9)$$

- Random samples of 100, 500, 1000, 3000, 5000 and 7500

Experimental Setup

Implementation and comparison :

- SVM with linear kernel as binary classification algorithm
- Value of C chosen by cross-validation
- Comparison with OVA, OVO, M-SVM, LogT

Performance Evaluation :

- Accuracy : Correctly classified examples in test dataset
- Macro F-Measure : Harmonic mean of precision and recall

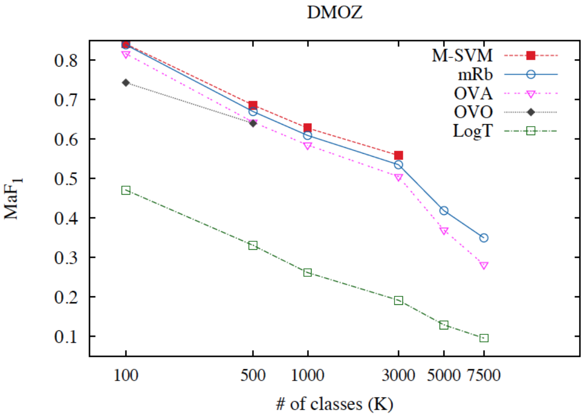
Experimental Results

Result for 7500 class :

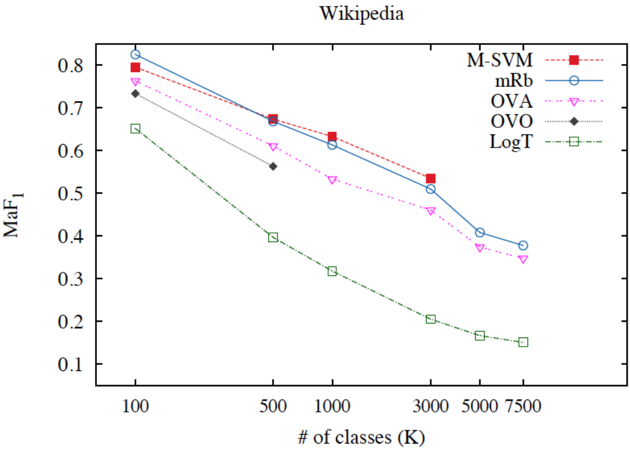
	DMOZ-7500			Wikipedia-7500		
	Acc.	MaF_1	N_c	Acc.	MaF_1	N_c
mRb	.479$^{\downarrow}$	**.352**	.495	.437$^{\downarrow}$	**.378**	.551
OVA	**.549**	.282$^{\downarrow}$	.379	**.484**	.348$^{\downarrow}$	.489
LogT	.311$^{\downarrow}$	.096$^{\downarrow}$	.194	.231$^{\downarrow}$	.151$^{\downarrow}$	.287

- OVO and M-SVM did not pass the scale for 7500 classes
- N_c : Proportion of classes for which at leaset one TP document found
- mRb covers 6-9.5% classes than OVA (500 - 700 classes)

of Classes Vs. Macro F-Measure



of Classes Vs. Macro F-Measure



Conclusion

- A new method of large-scale multiclass classification based on reduction of multiclass classification to binary classification.

- Efficiency of deduced algorithm comparable or better than the state of the art multiclass classification approaches.