Aggressive Double Sampling for Reducing Multi-class Classification to Binary Classification

Bikash Joshi (PhD Student)
AMA team, LIG

Supervised By:

Prof. Massih-Reza Amini and Dr. Franck Iutzeler

March 20, 2017

Outline

Outline

Multiclass Classification: Introduction

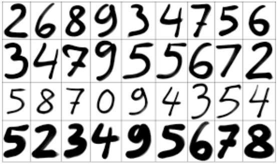


Figure : Digit Classification

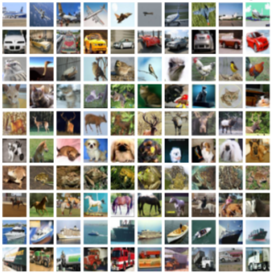


Figure : Image Classification

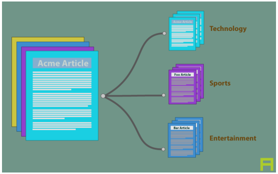


Figure : Text Classification

- Finite set of categories ($K > 2$)
- Popular applications: image and text classification.

Multiclass classification: Related Work

1. Combined approaches based on binary classification:
   - One-Vs-Rest
     - One binary problem for each class
     - K binary problems
     - O(K $\times$ d)
   - One-Vs-One
     - One binary problem for each pair of classes
     - O($K^2 \times$ d)
2. Uncombined Approaches
   - for example: multiclass SVM, MLP
   - One scoring function per class
3. Logarithmic Time Algorithms
   - For example: logTree, Recall-Tree
   - Each leaf node represents a class
   - O(logK)

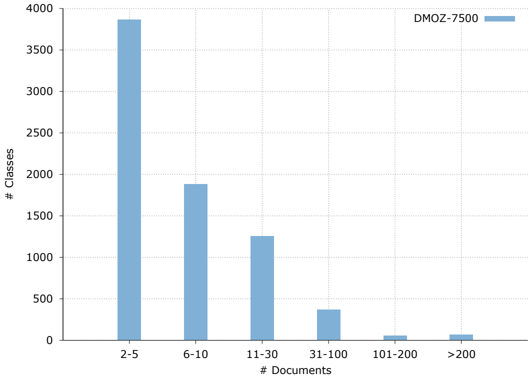
Multiclass classification : Challenges

- The number of classes, K, in new emerging multiclass problems, for example in text and image classification, may reach 10^5 to 10^6 categories.
- For example:

- 4×10^6 sites
- 10^6 categories
- 10^5 editors
- Imbalanced nature of hierarchies

Multiclass classification : Challenges

- Class imbalance problem
- Majority of classes have few representative examples
- Long tailed distribution



Text Classification:

Task: Automatic classification of an example text to one of fixed set of categories.

Feature Representation:

- Bag of Words:
  - From training corpus extract vocabulary.
  - Represent each terms as 0 or 1
  - Highly sparse
- Document-class joint feature representation:
  - Inspired by learning to rank
  - Similarity features between an example and class of examples
  - For example:

$$\sum_{t \in y \cap x} 1$$

Where,
x $\rightarrow$ One document
y $\rightarrow$ Class of documents

Outline

Motivation of our work

- Baselines: Model complexity increases with classes(K) and feature dimension (d).
- Algorithm that scales well for large scale data
- Does not suffer from class imbalance problem
- Less complex model
- Competitive with the state of the art approaches

Framework

- $\mathcal{X} \subseteq \mathbb{R}^d$: Input Space
- $\mathcal{Y} = 1,...,K$: Output Space
- $S = (x_i^{y_i})_{i=1}^m$: Training set of i.i.d. pairs
- $G = g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$: Class of predictors

Instantaneous Loss

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

- $\mathbb{1}_\pi$ is the indicator function (Value is 0 or 1)
- Average number of classes that get greater scoring by g than true class
- Ranking loss used in Multiclass-SVM [a]

[a] Weston et. al. (1998)

Framework

Empirical Loss

Empirical error of $g \in G$ over S is:

$$L_m(g,S) = \frac{1}{m(K-1)} \sum_{i=1}^{m} \sum_{y' \in Y \setminus y_i} \mathbb{1}_{g(x_i^{y_i}) \leq g(x_i^{y'})} \tag{2}$$

$$= \frac{1}{m(K-1)} \sum_{i=1}^{m} \sum_{y' \in Y \setminus y_i} \mathbb{1}_{\underbrace{h(x_i^{y_i}, x_i^{y'})}_{g(x_i^{y_i}) - g(x_i^{y'})} \leq 0} \tag{3}$$

- Resembles to binary-classification-loss based risk
- Selection of a hypothesis in G minimizing risk over S is equivalent to search a hypothesis in H minimizing risk over T(S) of size $m \times (K-1)$

Multiclass to binary reduction example

$$\mathcal{S} \quad \boxed{\mathbf{x}_1^{y_1} \quad \mathbf{x}_2^{y_2} \quad \mathbf{x}_3^{y_3} \quad \mathbf{x}_4^{y_4}}$$

$$\downarrow T$$

$$\begin{array}{lll}
(z_1 = (\mathbf{x}_1^{y_1}, \mathbf{x}_1^{y_2}), +1) & (z_2 = (\mathbf{x}_1^{y_1}, \mathbf{x}_1^{y_3}), +1) & (z_3 = (\mathbf{x}_1^{y_1}, \mathbf{x}_1^{y_4}), +1) \\
(z_4 = (\mathbf{x}_2^{y_1}, \mathbf{x}_2^{y_2}), -1) & (z_5 = (\mathbf{x}_2^{y_2}, \mathbf{x}_2^{y_3}), +1) & (z_6 = (\mathbf{x}_2^{y_2}, \mathbf{x}_2^{y_4}), +1) \\
(z_7 = (\mathbf{x}_3^{y_1}, \mathbf{x}_3^{y_3}), -1) & (z_8 = (\mathbf{x}_3^{y_2}, \mathbf{x}_3^{y_3}), -1) & (z_9 = (\mathbf{x}_3^{y_3}, \mathbf{x}_3^{y_4}), +1) \\
(z_{10} = (\mathbf{x}_4^{y_1}, \mathbf{x}_4^{y_4}), -1) & (z_{11} = (\mathbf{x}_4^{y_2}, \mathbf{x}_4^{y_4}), -1) & (z_{12} = (\mathbf{x}_4^{y_3}, \mathbf{x}_4^{y_4}), -1)
\end{array}$$

- We consider the following transformation

$$T(\mathcal{S}) = \left(\begin{cases} \left(\mathbf{z}_j = \left(\mathbf{x}_i^k, \mathbf{x}_i^{y_i}\right), \tilde{y}_j = -1\right) & \text{if } k < y_i \\ \left(\mathbf{z}_j = \left(\mathbf{x}_i^{y_i}, \mathbf{x}_i^k\right), \tilde{y}_j = +1\right) & \text{elsewhere} \end{cases} \right)_{j \doteq (i-1)(K-1)+k}$$

- $|T(S)| = \text{m} \times (\text{K - 1})$

Multiclass to binary reduction algorithm

[Bikash et al. 2015]

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 $T(S)$ to get learned weight vector w
Testing:For test example, x', estimate $\Phi(x'^y)$ for all x'^y pairs and predicted class is the one which maximizes $\langle w, \Phi x'^y \rangle$

Fig. 2. Multiclass reduced to binary classification (mRb)

Improvements and New challenges

Improvements:

- One parameter vector for all classes.
- Low-dimensional feature space.
- Overcome class imbalance.

New Challenges:

- Number of transformations huge for larger K
- Large computational overhead
- Large memory requirement

Outline

Aggressive double sampling

1. Drawing uniformly μ examples per class, in order to form practical set S_μ;
   - Reduce redundancy in examples
   - Emphasizing rare classes
2. For each example $\mathbf{x}^y$ in S_μ, drawing uniformly κ adversarial classes in $\mathcal{Y}\backslash\{y\}$.
   - Reduces time complexity
   - Low memory requirement

Double Sampled Multi to Binary Reduction

Input: Labeled training set $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m$

Initialize

$T_\kappa(\mathcal{S}_\mu) \leftarrow \emptyset$

$\mathcal{S}_\mu \leftarrow \emptyset$

for $k = 1..K$ **do**

  Draw randomly a set $\underline{\mathcal{S}}$ of μ examples of class k from $\mathcal{S}$ $\triangleright$ $\mu << |\mathcal{S}_k|$

  $\mathcal{S}_\mu \leftarrow \mathcal{S}_\mu \cup \underline{\mathcal{S}}$

end

forall the $\mathbf{x}^y \in \mathcal{S}_\mu$ **do**

  Draw uniformly a set $\underline{\mathcal{K}}$ of κ classes from $\mathcal{Y} \backslash \{y\}$ $\triangleright$ $\kappa << |K|$

  forall the $k \in \underline{\mathcal{K}}$ **do**

    if $k < y$ **then**

      $T_\kappa(\mathcal{S}_\mu) \leftarrow T_\kappa(\mathcal{S}_\mu) \cup \big(\mathbf{z} = \big(\phi(\mathbf{x}^k), \phi(\mathbf{x}^y)\big) \quad , \tilde{y} = -1\big)$

    end

    else

      $T_\kappa(\mathcal{S}_\mu) \leftarrow T_\kappa(\mathcal{S}_\mu) \cup \big(\mathbf{z} = \big(\phi(\mathbf{x}^y), \phi(\mathbf{x}^k)\big) \quad , \tilde{y} = +1\big)$

    end

  end

end

return $T_\kappa(\mathcal{S}_\mu)$

Outline

Experimental Setup

Datasets:

- Application: Text Classification
- DMOZ and Wikipedia datasets. (LSHTC challenge)
- Pre-processed with stop word removal and stemming.
- Random samples of 1000, 2000, 3000, 4000, 5000, 7500, 10000, 20000.

Comparison:

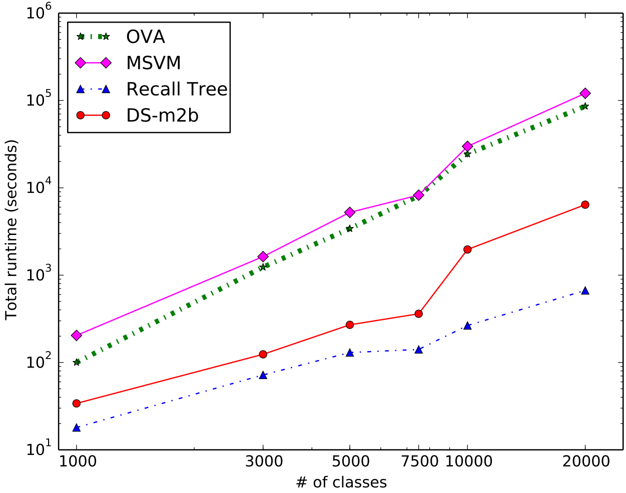
- DS-m2b: Proposed double sampled multiclass to binary algorithm
- OVA: One-Vs-All algorithm
- M-SVM: Crammar-Singer implementation of multiclass SVM
- Recall Tree: Hierarchical One-Vs-Some algorithm

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

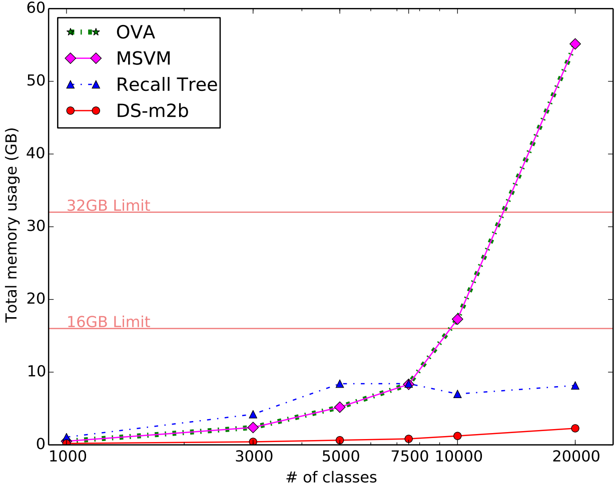
Features			
1.	$\sum_{t\in y\cap x} \ln(1+y_t)$	2.	$\sum_{t\in y\cap x} \ln(1+\frac{l_{\mathcal{S}}}{\mathcal{S}_t})$
3.	$\sum_{t\in y\cap x} I_t$	4.	$\sum_{t\in y\cap x} \ln(1+\frac{y_t}{\lvert y\rvert})$
5.	$\sum_{t\in y\cap x} \ln(1+\frac{y_t}{\lvert y\rvert}.I_t)$	6.	$\sum_{t\in y\cap x} \ln(1+\frac{y_t}{\lvert y\rvert}.\frac{l_{\mathcal{S}}}{\mathcal{S}_t})$
7.	$\sum_{t\in y\cap x} 1$	8.	$\sum_{t\in y\cap x} \frac{y_t}{\lvert y\rvert}.I_t$
9.	BM25	10.	$d(x^y, centroid(y))$

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

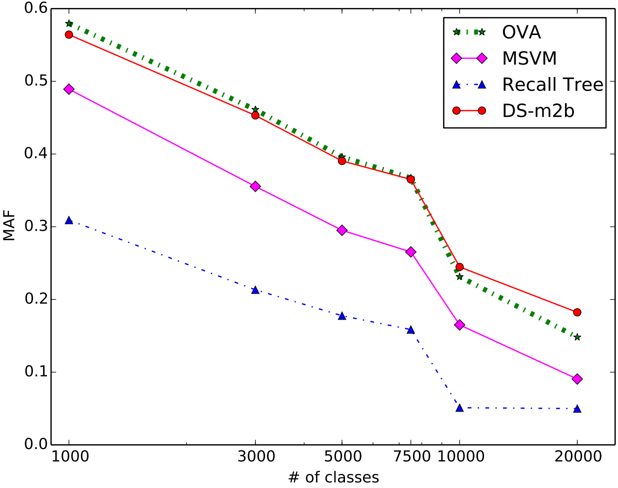
Results: Runtime Comparison



Results: Memory Comparison



Results: Prediction Performance Comparison



Outline

Conclusion:

- Multiclass to binary reduction to handle large-class scenario and overcome class imbalance problem.
- Use of double sampling to further improve computational complexity and memory usage.

Questions?