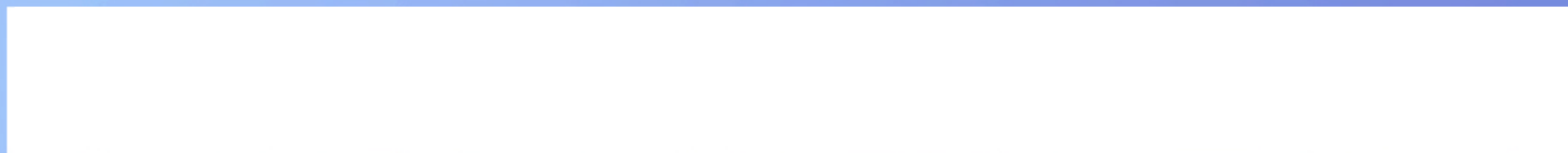


# **Incremental Algorithms for Hierarchical Classification**

Authors: Cesa-Bianchi, N.  
Gentile, C.  
Zaniboni, L.

Journal of Machine Learning Research 7 (2006) 31-54

**Presented by: Jörg Meyer**



# Overview



- Introduction
- H-Loss
- H-RLS
- Analysis
- Experiments

**Introduction**

**H-Loss**

**H-RLS**

**Analysis**

**Experiments**

# Introduction



- Hierarchical online classifier
- Data is produced frequently / in large amount
- Classification scenario:
  - Data
  - Hierarchy
  - Linear-threshold classifier for each node
  - Evaluation

# Introduction



- Notation:
  - Instance  $x \in \mathbb{R}^N$
  - Label / multilabel  $v = (v_1, v_2, \dots, v_N) \in \{0, 1\}^N$
  - Example  $(x, v)$
  - Taxonomy  $G$  (forest of trees)
  - Multilabel respects taxonomy

# Introduction



- Notation:
  - $\text{Anc}(i)$
  - $\text{Par}(i)$
  - $\text{Root}(G)$
- Multi-/partial-path labelling

# H-Loss



- H-RLS algorithm basics
- Loss functions:
  - Zero-one loss  $l_{0/1}$
  - Symmetric difference loss  $l_{\Delta}$
  - H-Loss  $l_H$



- H-LOSS:

$$l_H(\hat{y}, v) = \sum_{i=1}^N \{\hat{y}_i \neq v_i \wedge \hat{y}_j = v_j, j \in \text{ANC}(i)\}$$

- $l_{0/1} \leq l_H \leq l_\Delta$

# H-RLS



- H-RLS = Hierarchical – Regularized Least Squares
- Online algorithm
- $N$  linear-threshold classifier
- Label all root nodes
- Label all children of nodes labelled with 1



**Algorithm H-RLS.****Initialization:** Weight vectors  $w_{i,1} = (0, \dots, 0)$ ,  $i = 1, \dots, N$ .For  $t = 1, 2, \dots$  do

1. Observe instance  $x_t \in \{x \in \mathbb{R}^d : \|x\| = 1\}$ ;
2. For each  $i = 1, \dots, N$  compute predictions  $\hat{y}_{i,t} \in \{0, 1\}$  as follows:

$$\hat{y}_{i,t} = \begin{cases} \{w_{i,t}^\top x_t \geq 0\} & \text{if } i \text{ is a root node,} \\ \{w_{i,t}^\top x_t \geq 0\} & \text{if } i \text{ is not a root node and } \hat{y}_{j,t} = 1 \text{ for } j = \text{PAR}(i), \\ 0 & \text{if } i \text{ is not a root node and } \hat{y}_{j,t} = 0 \text{ for } j = \text{PAR}(i), \end{cases}$$

where

$$w_{i,t} = (I + S_{i,Q(i,t-1)} S_{i,Q(i,t-1)}^\top + x_t x_t^\top)^{-1} \times \\ \times S_{i,Q(i,t-1)} (v_{i,i_1}, v_{i,i_2}, \dots, v_{i,i_{Q(i,t-1)}})^\top \\ S_{i,Q(i,t-1)} = [x_{i_1} \ x_{i_2} \ \dots \ x_{i_{Q(i,t-1)}}] \quad i = 1, \dots, N.$$

3. Observe multilabel  $v_t$  and update weights.



- Standard perceptron weight update:

$$w_{ij}^{\text{neu}} = w_{ij}^{\text{alt}} + \Delta w_{ij}$$

$$\Delta w_{ij} = \alpha(t_j - o_j) \cdot x_i$$

- Old weight is basis value for new weight
- H-RLS weight update:

$$w_{i,t} = \left( I + S_{i,Q(i,t-1)} S_{i,Q(i,t-1)}^\top + x_t x_t^\top \right)^{-1} S_{i,Q(i,t-1)} (v_{i,i_1}, \dots, v_{i,i_{Q(i,t-1)}})^\top$$

- Indirect influence of old weights



- Weight update:  $w_{i,t} = \left( I + S_{i,Q(i,t-1)} S_{i,Q(i,t-1)}^\top + x_t x_t^\top \right)^{-1} S_{i,Q(i,t-1)} (v_{i,i_1}, \dots, v_{i,i_{Q(i,t-1)}})^\top$

- With:  $Q(i,t) = |\{1 \leq s \leq t : v_{\text{PAR}(i),s} = 1\}|$

$$S_{i,Q(i,t-1)} = [x_{i_1} x_{i_2} \dots x_{i_{Q(i,t-1)}}]$$

$$(v_{i,i_1}, v_{i,i_2}, \dots, v_{i,i_{Q(i,t-1)}})$$



- Evaluate performance of the algorithm
- Find error-bound
- Label generation

– Probability distribution

$$f_G(\mathbf{v} | x) = \prod_{i=1}^N \mathbb{P}(V_i = v_i | V_j = v_j, j = \text{PAR}(i), x)$$

– Respect taxonomy

$$\mathbb{P}(V_i = 1 | V_j = 0, x) = 0$$

– Probability for node (non root)

$$\mathbb{P}(V_i = 1 | V_j = 1, x) = \frac{1 + u_i^\top x}{2}.$$



- Reference classifier
  - Built on true parameters  $u_i$
  - Same form as H-RLS

$$y_i = \begin{cases} \{u_i^\top x \geq 0\} & \text{if } i \text{ is a root node,} \\ \{u_i^\top x \geq 0\} & \text{if } i \text{ is not a root and } y_j = 1 \text{ for } j = \text{PAR}(i), \\ 0 & \text{if } i \text{ is not a root and } y_j = 0 \text{ for } j = \text{PAR}(i). \end{cases}$$

- Create multilabel distribution, as shown before

# Analysis



- Cumulative regret  $\sum_{t=1}^T (\mathbb{E} \ell(\hat{y}_t, V_t) - \mathbb{E} \ell(y_t, V_t))$

- Will hold theoretical regret bound

$$\sum_{t=1}^T (\mathbb{E} \ell_H(\hat{y}_t, V_t) - \mathbb{E} \ell_H(y_t, V_t)) \leq 16(1 + 1/e) \sum_{i=1}^N \frac{C_i}{\Delta_i^2} \mathbb{E} \left[ \sum_{j=1}^d \log(1 + \lambda_{i,j}) \right],$$

where

$$\Delta_{i,t} = u_i^\top x_t, \quad \Delta_i^2 = \min_{t=1, \dots, T} \Delta_{i,t}^2, \quad C_i = |\text{SUB}(i)|,$$

- $w_i$  is an asymptotically unbiased estimator for  $u_i$



- Theoretical regret bound

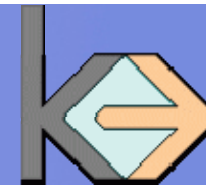
$$\sum_{t=1}^T (\mathbb{E} \ell_H(\hat{y}_t, V_t) - \mathbb{E} \ell_H(y_t, V_t)) \leq 16(1 + 1/e) \sum_{i=1}^N \frac{C_i}{\Delta_i^2} \mathbb{E} \left[ \sum_{j=1}^d \log(1 + \lambda_{i,j}) \right],$$

where

$$\Delta_{i,t} = u_i^\top x_t, \quad \Delta_i^2 = \min_{t=1, \dots, T} \Delta_{i,t}^2, \quad C_i = |\text{SUB}(i)|,$$

- Depends on hierarchy structure
- The deeper the node  $i$ , the less the contribution

- Cost sensitive H-Loss  $\ell_H(\hat{y}, v) = \sum_{i=1}^N c_i \{ \hat{y}_i \neq v_i \wedge \hat{y}_j = v_j, j \in \text{ANC}(i) \},$



Newswire stories from Reuters Corpus Volume 1  
(first 100.000 stories)

- Taxonomy: document topics, 101 nodes
- 5 experiments, adjacent pair = training & test set

Subtree of „Quality of Health Care“ (55.503 documents)

- Taxonomy: remove cycles, 94 nodes
- 5 experiments, 40.000 training & 15.503 test

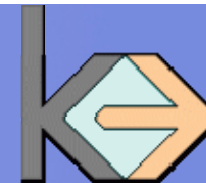


# Experiments



- Linear grow of space complexity  $\Rightarrow$  SH-RLS
- Algorithms
  - Hierarchical: H-PERC, H-SVM
  - Flat: PERC, SVM, S-RLS

# Experiments



RCV1			
Algorithm	zero-one loss	uniform H-loss	$\Delta$ -loss
PERC	0.702( $\pm 0.045$ )	1.196( $\pm 0.127$ )	1.695( $\pm 0.182$ )
H-PERC	0.655( $\pm 0.040$ )	1.224( $\pm 0.114$ )	1.861( $\pm 0.172$ )
S-RLS	0.559( $\pm 0.005$ )	0.981( $\pm 0.020$ )	1.413( $\pm 0.033$ )
SH-RLS	<b>0.456(<math>\pm 0.010</math>)</b>	<b>0.743(<math>\pm 0.026</math>)</b>	<b>1.086(<math>\pm 0.036</math>)</b>
SVM	0.482( $\pm 0.009$ )	0.790( $\pm 0.023$ )	1.173( $\pm 0.051$ )
H-SVM	0.440( $\pm 0.008$ )	0.712( $\pm 0.021$ )	1.050( $\pm 0.027$ )

OHSUMED			
Algorithm	zero-one loss	uniform H-loss	$\Delta$ -loss
PERC	0.899( $\pm 0.024$ )	1.938( $\pm 0.219$ )	2.639( $\pm 0.226$ )
H-PERC	0.846( $\pm 0.024$ )	1.560( $\pm 0.155$ )	2.528( $\pm 0.251$ )
S-RLS	0.873( $\pm 0.004$ )	1.814( $\pm 0.024$ )	2.627( $\pm 0.027$ )
SH-RLS	<b>0.769(<math>\pm 0.004</math>)</b>	<b>1.200(<math>\pm 0.007</math>)</b>	<b>1.957(<math>\pm 0.011</math>)</b>
SVM	0.784( $\pm 0.003$ )	1.206( $\pm 0.003$ )	1.872( $\pm 0.005$ )
H-SVM	0.759( $\pm 0.002$ )	1.170( $\pm 0.005$ )	1.910( $\pm 0.007$ )

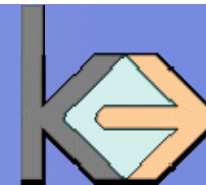
Introduction

H-Loss

H-RLS

Analysis

Experiments



Thank you for your attention

**Introduction**

**H-Loss**

**H-RLS**

**Analysis**

**Experiments**