ULB UNIVERSITÉ LIBRE DE BRUXELLES, UNIVERSITÉ D'EUROPE

Vrije Universiteit Brussel

MLG MACHINE LEARNING GROUP

CoMo Computational Modeling Lab

# Supervised Dimensionality Reduction Unsupervised and Transfer Learning Challenge

*Yann-Aël Le Borgne*

*Machine Learning Group - Computational Modeling Lab*

*Université Libre de Bruxelles - Vrije universiteit Brussel*

*Brussels - Belgium*

## Challenge summary

- 6 datasets, with very different number of features, classes, and levels of sparsity.
- 3 subsets for each dataset: *developement*, *validation* and *final*. Each subsets contained instances of different classes.

| Dataset | Domain | Features | Sparsity | #Dev. samples | #Labels | #Classes |
|---|---|---|---|---|---|---|
| **AVICENNA** | OHR | 120 | 0% | 150205 | 50000 | 5 |
| **HARRY** | Video | 5000 | 98.1% | 69652 | 20000 | 4 |
| **RITA** | Images | 7200 | 1.1% | 111808 | 24000 | 4 |
| **SYLVESTER** | Ecology | 100 | 0% | 572820 | 100000 | 2 |
| **TERRY** | Text | 47236 | 99.8% | 217034 | 40000 | 4 |
| **ULE** | OHR | 784 | 80.9% | 26808 | 10000 | 4 |

Characteristics of the different datasets

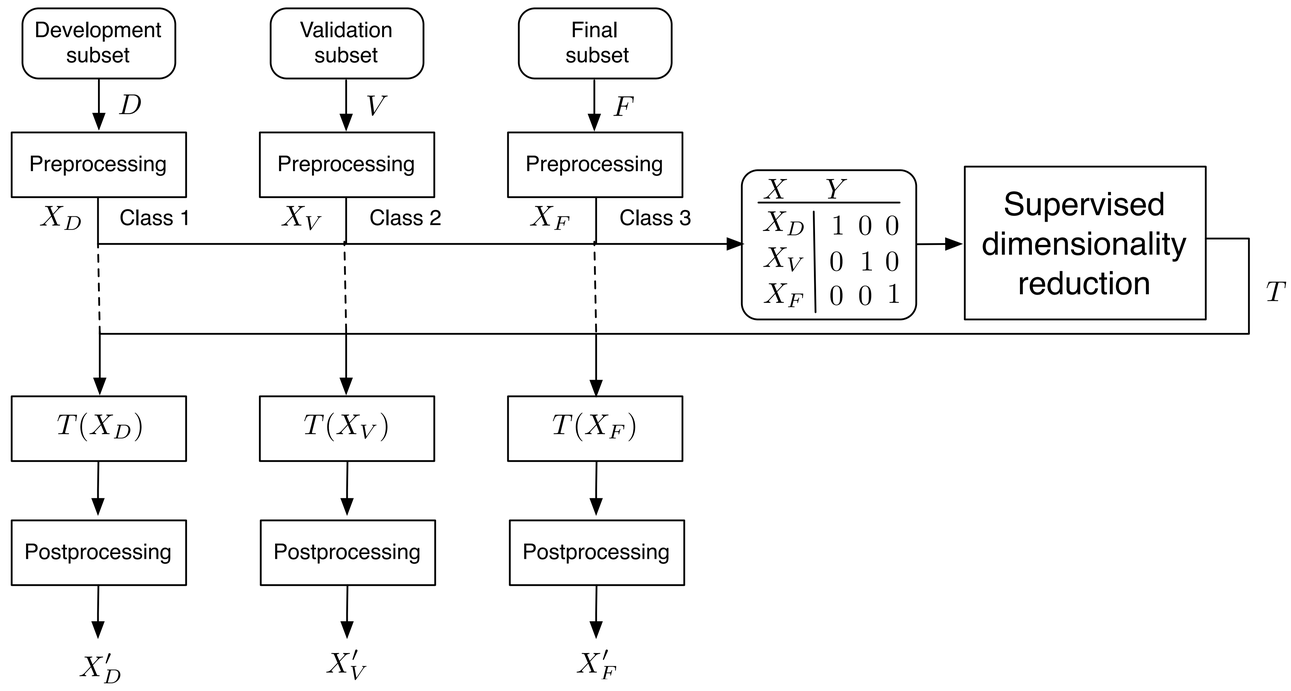
| Dataset | DevelSub | Validation | Final |
|---|---|---|---|
| **AVICENNA** | 0.2990 | 0.1034 | 0.1501 |
| **HARRY** | 0.1469 | 0.6264 | 0.6017 |
| **RITA** | 0.0799 | 0.2504 | 0.4133 |
| **SYLVESTER** | 0.2400 | 0.2167 | 0.3095 |
| **TERRY** | 0.5817 | 0.6969 | 0.7550 |
| **ULE** | 0.5237 | 0.7905 | 0.7169 |

AUC of each subset (using raw data)

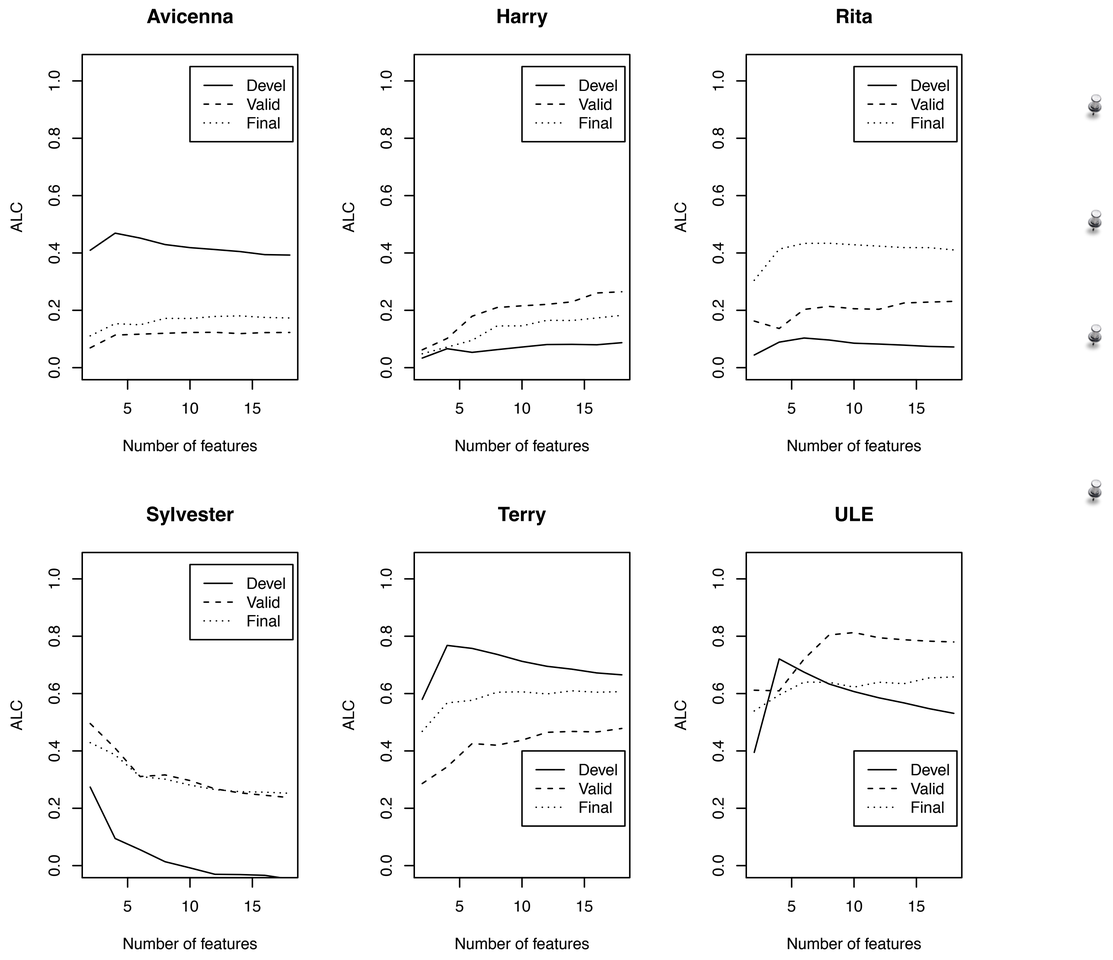
- The accuracy of each subset was assessed in terms of AUC (Area Under the Curve), after training with a Hebbian classifier.
- The goal was to transform the set of features from each subset in order to maximize the AUC. The competition had 2 phases:
  - First phase: Unsupervised learning, for which no labels were available.
  - Second phase: Transfer learning, for which the labels of some instances in the development set were given.
- AUCs of validation subsets could be obtained on the competition website. AUCs of final subsets were unknown and used for the final ranking.

## Supervised Dimensionality reduction

- Rationale: we know that each subset contains different sets of classes. A 'superclass' is assigned to each subset.
- Learning is carried out using those superclasses.
- Techniques used:
  - Preprocessing: normalization, whitening, PCA, hierarchical clustering, K-means.
  - Supervised dimensionality reduction: Partial least square.
  - Postprocessing: Discretization.




## Results




AUC on each dataset using Partial Least Squares

- Partial least squares on superclasses allowed to improve the AUC for the Avicenna and Sylvester datasets.
- Additional use of tansfer labels further improved accuracy for the Rita and ULE datasets.
- The choice for the number of components was difficult in practice since the AUC on development sets did not quite follow AUC on validation sets.
- Results for the competition were obtained with a combination of PCA and a discretization procedure which made the Hebbian classifier insensitive to the sign of the features.

| Dataset | Valid | Final |
|---|---|---|
| **AVICENNA (Raw)** | 0.1350 | 0.1798 |
| **HARRY (7 PCs)** | 0.7184 | 0.6400 |
| **RITA (7 PCs)** | 0.2725 | 0.4260 |
| **SYLVESTER (8 PCs)** | 0.6369 | 0.4775 |
| **TERRY (7 PCs)** | 0.7255 | 0.7560 |

Competition AUCs with PCA and discretization