Characterizing Multiple Instance Datasets

Veronika Cheplygina[*†] David M. J. Tax[†]

[*]Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands, [†]Pattern Recognition Laboratory, Delft University of Technology, The Netherlands

Introduction

- How to choose benchmark datasets to evaluate your new and improved classifier?
- Often metadata like sample size or dimensionality used to select a diverse set
- "Similar" datasets may have different behavior and vice versa
- Case study with multiple instance learning (MIL) datasets

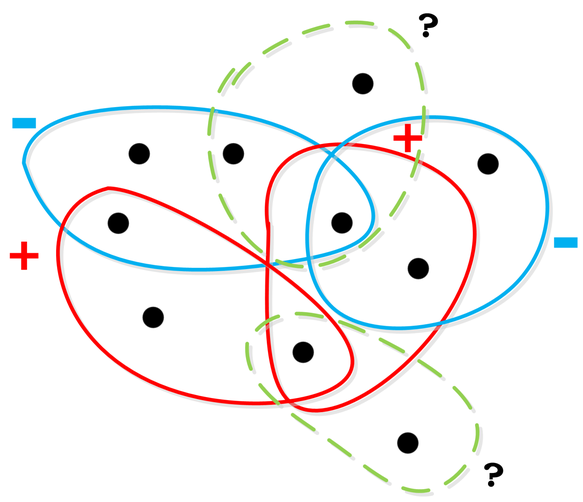

Figure 1: MIL problem with positive (red) and negative (blue) bags

- Sample = *bag* of *instances* $B_i = \{\mathbf{x}_{ij}\}_{j=1}^{n_i}$
- Only bag labels $y_i \in \{+1, -1\}$ given
- Positive bag $\leftrightarrow$ at least 1 positive (concept) instance?
- Drug activity, images, text...

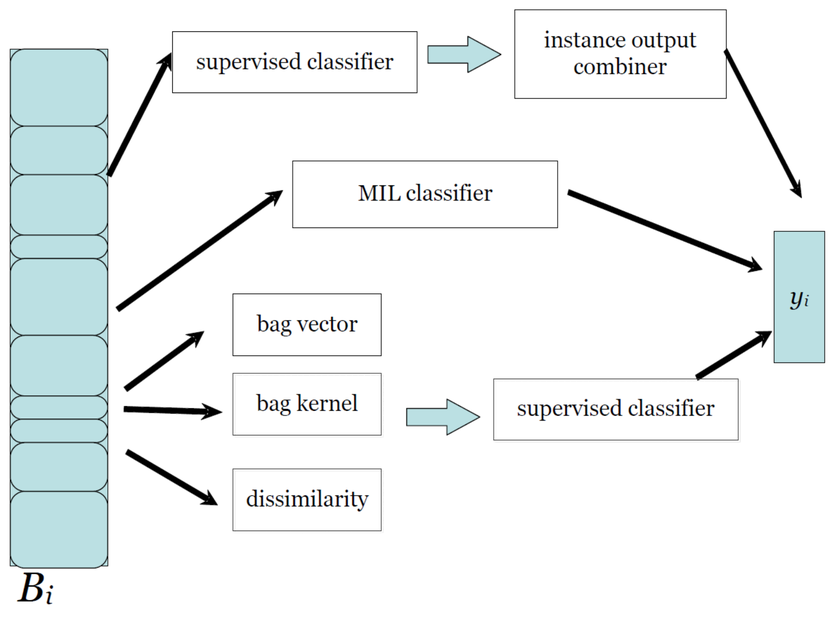

Figure 2: Strategies of MIL classifiers

Proposed Approach

- Characterize 40 datasets [1] by 22 uncorrelated classifiers
- Dataset distance defined by 6 meta-features (D_{meta}), AUC (D_{auc}) or ROC curves (D_{roc})
- Embed distances by multi-dimensional scaling

Dataset i, classifier k: ROC curve $\mathbf{ROC}_k^{(i)}$ with $\mathbf{AUC}_k^{(i)} = \mathcal{A}(\mathbf{ROC}_k^{(i)})$

$$D_{auc}(i,j) = \|\mathbf{AUC}^{(i)} - \mathbf{AUC}^{(j)}\| \quad (1)$$

where $\mathbf{AUC}^{(i)} = [\mathcal{A}(\mathbf{ROC}_1^{(i)}), \ldots, \mathcal{A}(\mathbf{ROC}_L^{(i)})]^T$.

$$D_{roc}(X_i, X_j) = \sqrt{\sum_k \mathcal{A}(\mathbf{ROC}_k^{(i)} - \mathbf{ROC}_k^{(j)})^2} \quad (2)$$

Experiments

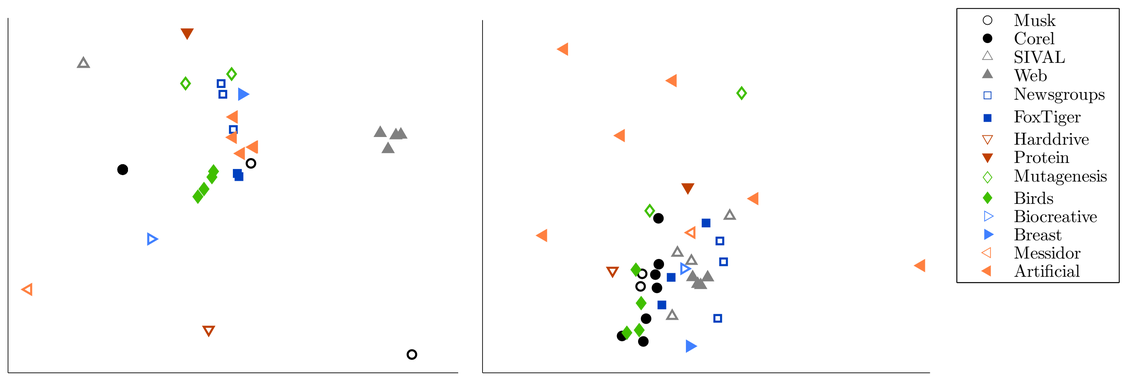

Figure 3: Left: MDS embedding of the Euclidean distances between the meta-representations of the datasets. Right: MDS embedding of D_{roc}.

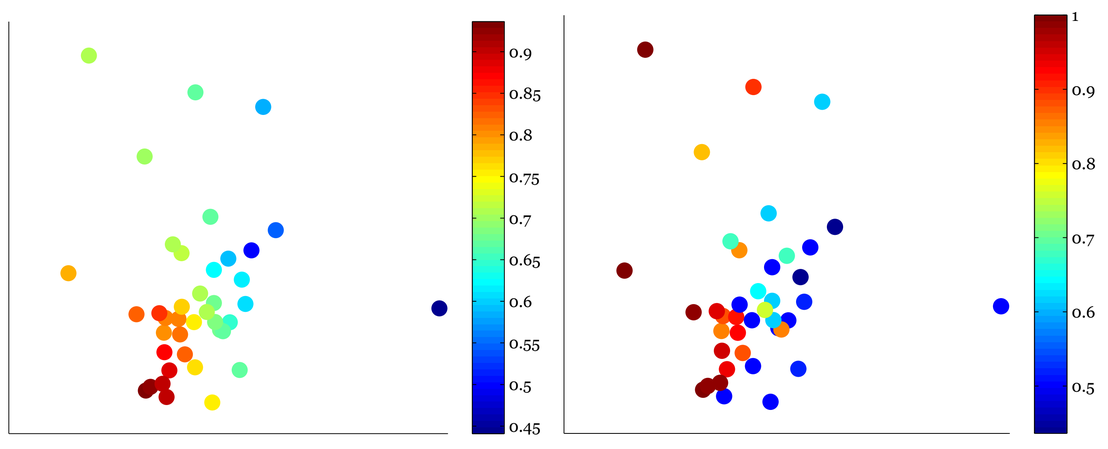

Figure 4: D_{roc} embedding with (left) average performance of all classifiers and (right) performance of concept-based EMDD classifier

Conclusions

- Characterization by classifier performances more relevant to end goal of classifier evaluation
- Datasets with similar characteristics behave differently
- D_{auc} and D_{roc} similar in practice
- Artificial datasets are outliers
- Can we generate "realistic" artificial data?
- Use ranks of classifiers instead of performances?

[1] Veronika Cheplygina, David M. J. Tax, and Marco Loog. Bag Dissimilarities for Multiple Instance Learning, *Pattern Recognition*, 48(1): 264–275, 2015. `http://www.miproblems.org`

Data and tools

Delft University of Technology