Label Stability in Multiple Instance Learning

Veronika Cheplygina[*†] Lauge Sørensen[‡] David M. J. Tax[†] Marleen de Bruijne[*‡]
Marco Loog[†‡]

[*]Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands, [†]Pattern Recognition Laboratory, Delft University of Technology, The Netherlands,
[‡]The Image Group, Department of Computer Science, University of Copenhagen, Denmark

Introduction

- We want to predict local patch labels in images, but only global image labels are available for training
- (Some) multiple instance learning (MIL) classifiers can do this, and are gaining popularity in computer-aided diagnosis
- Are the patch labels reliable?

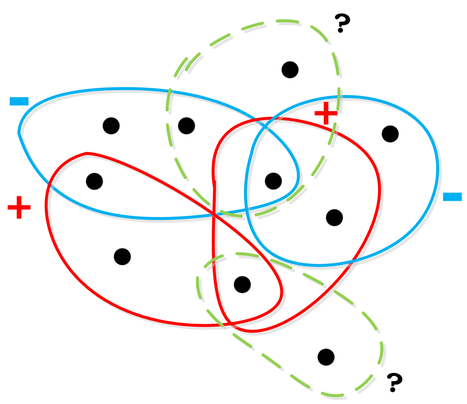

Figure 1: MIL problem with positive (red) and negative (blue) bags

- Patch = feature vector or *instance* $\mathbf{x}$
- Image = *bag* of instances $B_i = \{\mathbf{x}_{ij}\}_{j=1}^{n_i}$
- Only bag labels $y_i \in \{+1, -1\}$ are given
- Positive bag $\leftrightarrow$ at least one positive instance?
- Classifiers optimize performance on training bag labels

Instance Stability

Instance labels are often not evaluated, or only qualitatively with fixed training set. What if the training set changes?

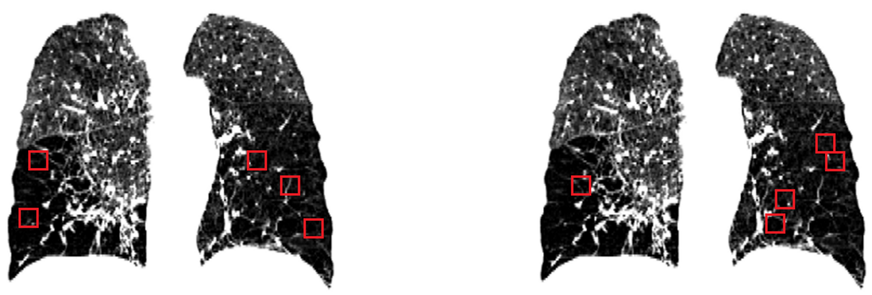

Figure 2: Unstable instance labels

In the absence of ground truth, we propose to evaluate *instance stability*:

$$S_+(\mathbf{z}, \mathbf{z}') = n_{11}/(n_{01} + n_{10} + n_{11}) \quad (1)$$

where $\mathbf{z}$ and $\mathbf{z}'$ is are vectors of instance labels, $n_{11} = |\{i|z_i = 1 \wedge z'_i = 1\}|$ and the other n's are defined analogously.

Experiments

- Six MIL datasets [1] including CAD of diabetes in retinal images, breast cancer in histopathology images and COPD in CT lung images [2]
- Eight MIL classifiers
- Subsample training set 10$\times$ to train classifiers
- Evaluate bag AUC and instance stability on test set

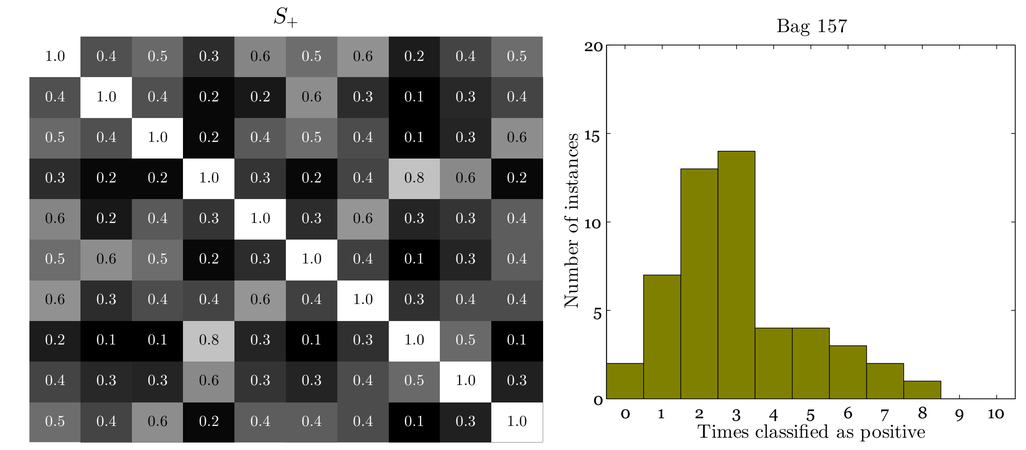

Figure 3: Left: Pairwise stability measures for 10 MILES classifiers for COPD data. Right: Instance classifications for a true positive bag that is classified as positive 10 times

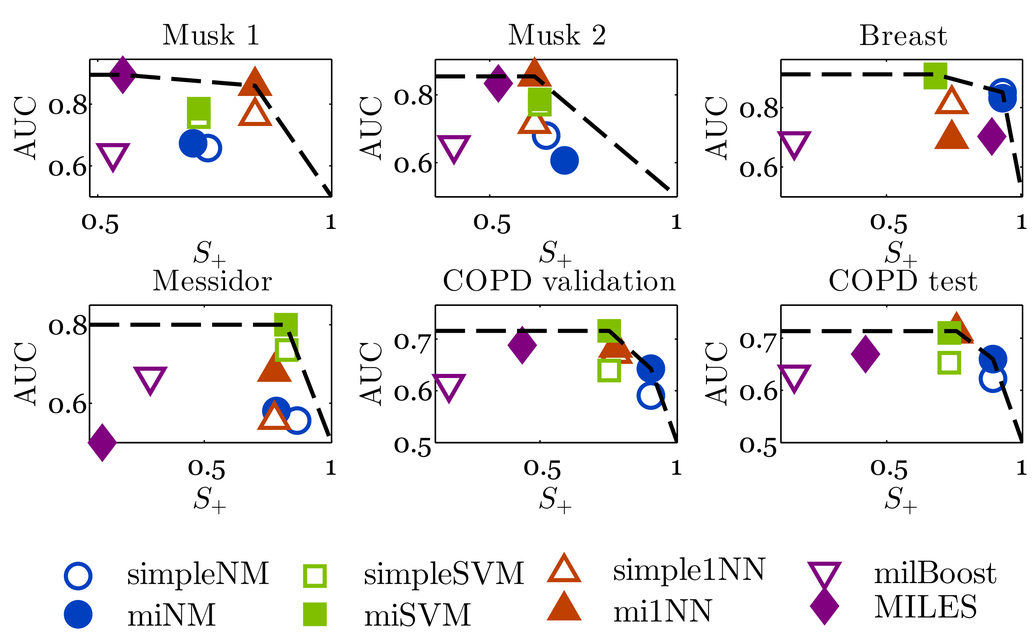

Figure 4: Trade-off of bag AUC and instance stability in six datasets. "Best" classifier on bags not always reliable for instances!

Conclusions

- MIL gaining popularity in computer-aided diagnosis
- Instance labels often not evaluated
- Trade-off bag performance and instance stability
- Use stability as additional evaluation measure for classifiers

[1] V. Cheplygina et al. Multiple instance learning with bag dissimilarities, *Pattern Recognition*, 48(1): 264–275, 2015.
`http://www.miproblems.org`

[2] L. Sørensen et al. Texture-based analysis of COPD: a data-driven approach. *IEEE Transactions on Medical Imaging*, 31(1): 70–78, 2012.

Data and toolbox

Delft University of Technology