Asymmetry in Point Set Dissimilarities

Veronika Cheplygina V.CHEPLYGINA@TUDELFT.NL
David M.J. Tax D.M.J.TAX@TUDELFT.NL
Marco Loog M.LOOG@TUDELFT.NL
Pattern Recognition Laboratory, Delft University of Technology, The Netherlands

Keywords: multiple instance learning, dissimilarity representation, non-metric distances

In supervised learning, an object is characterized by a vector of features which aim to distinguish between objects of different classes. In some problems, it may not be straightforward to define what the features should be. This is often the case for complex objects such as graphs, where compressing such objects into a single feature vector representation may increase class overlap.

In multiple instance learing (MIL), the complex objects are called *bags* of *instances*. A bag is a collection of feature vectors, or a point set in a m-dimensional space $B = \{\mathbf{x}_i | i = 1, ..., |B|\} \subset \mathbb{R}^m$. Only bags are labeled $Y(B) \in \{+1, -1\}$, although hidden instance labels $y(\mathbf{x}) \in \{+1, -1\}$ and a mapping $Y(B) = f(\{y(\mathbf{x})\})$ are often assumed. In particular, positive or so-called *concept* instances are assumed to be most important for determining $Y(B)$. This learning setting has originated in drug activity prediction, but has also been applied to classification of images, documents, audio recordings and so forth.

A way of learning with complex objects is to learn from distances or *dissimilarities*, i.e., for MIL by defining a distance measure between bags. Such dissimilarities can be used with the nearest neighbor rule (assigning the object to the class of its closest neighbors), or more generally, as a feature space, where each feature is a dissimilarity to a set of prototypes $\mathcal{R}$(Pękalska & Duin, 2005). In the dissimilarity space (DS) each point set B is represented as a vector $\mathbf{d}(B, \mathcal{R}) = [d(B, R_1), ...d(B, R_{|\mathcal{R}|})]$. In this space, any supervised learner can be used.

Distance measures on complex objects often display non-metric properties, such as asymmetry: $d(B, B') \neq d(B', B)$. Consider the point sets in Fig.1. The metric Hausdorff distance is defined as the overall maximum of the minimum instance distances $\{d(\mathbf{x}, \mathbf{x}') | \mathbf{x} \in B, \mathbf{x}' \in B'\}$ between the two sets. However, we could also measure the minimum, average, etc. leading to possibly non-metric distances. This deteriorates the performance of the nearest neighbor rule, but in the DS, such dissimilarities may be more informative than their metric counterparts.

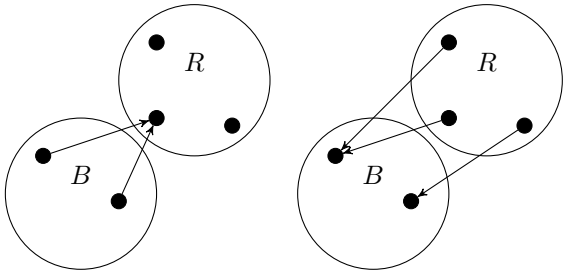

Figure 1. Minimum instance distances between two bags.

In our work(Cheplygina et al., 2012), we show that in some MIL problems, not both directions $d_{to}(B \rightarrow R)$ and $d_{from}(B \leftarrow R)$ are equally informative. In particular, when R is positive, it is often better to use $d_{from}(B \leftarrow R)$ because this ensures that the concept instances in R influence the dissimilarity value, leading to different values for positive and negative bags. On the other hand, with $d_{to}(B \rightarrow R)$ there is a risk that the concept instances in R are disregarded, therefore introducing unnecessary class overlap. In such cases it is not advisable to symmetrize the dissimilarity, but to use the asymmetric versions instead.

References

Cheplygina, V., Tax, D., & Loog, M. (2012). Class-dependent dissimilarity measures for multiple instance learning. *Structural, Syntactic, and Statistical Pattern Recognition*, 602–610.

Pękalska, E., & Duin, R. (2005). *The dissimilarity representation for pattern recognition: foundations and applications*, vol. 64. World Scientific Pub Co Inc.