On the Informativeness of Asymmetric Dissimilarities

Yenisel Plasencia Calaña [ab] Veronika Cheplygina [a] Robert P.W. Duin [a]

Edel García-Reyes [b] Mauricio Orozco-Alzate [c] David M.J. Tax [a]

Marco Loog [ad]

[a] *Pattern Recognition Laboratory, Delft University of Technology, The Netherlands*
[b] *Advanced Technologies Application Center, La Habana, Cuba*
[c] *Depto. Inf. y Comp., Universidad Nacional de Colombia - Sede Manizales*
[d] *Image Group, University of Copenhagen, Denmark*

1 Introduction

Nearest-neighbor (NN) classification has been widely used in many research areas, as it is a very intuitive technique. As long as we can defined a similarity or distance between two objects, we can apply NN, therefore making it suitable even for non-vectorial data such as graphs. An alternative to NN is the dissimilarity space [2], where distances are used as features, i.e. an object is represented as a vector of its distances to prototypes or landmarks. This representation can be used with any classifier, and has been shown to be potentially more effective than NN classification on the same dissimilarities.

Defining distance measures on complex objects is not a trivial task. Due to human judgments, suboptimal matching procedures or simply by construction, distance measures on non-vectorial data may often be asymmetric. A common solution for NN approaches is to symmetrize the measure by averaging the two distances [2]. However, in the dissimilarity space, symmetric measures are not required. We explore whether asymmetry is an artifact that needs to be removed, or an important source of information. This abstract highlights one example of informative asymmetric measures, covered in [1].

2 Asymmetry

One example where asymmetric dissimilarities can occur is in multiple instance learning (MIL), where we are given labeled sets (*bags*) of feature vectors (*instances*). MIL is used in molecule activity prediction, text and image classification. For example, an image can be represented by all the patches in the image, and a molecule can be represented by all the shapes it can fold into.

Consider the bags in Fig.1. The directed Hausdorff distance is defined as the maximum minimum instance distance, $d_h(B,R) = \max_{\mathbf{x}\in B} \min_{\mathbf{x}'\in R} d(\mathbf{x},\mathbf{x}')$. To achieve metricity, it is symmetrized as $d_H(B,R) = \max(d_h(B,R), d_h(R,B))$. However, as we explain shortly, the directed versions $d_h(B,R)$ and $d_h(R,B)$ may be more informative for MIL problems.

The full paper has been published in Similarity-Based Pattern Recognition, pages 75-89, 2013

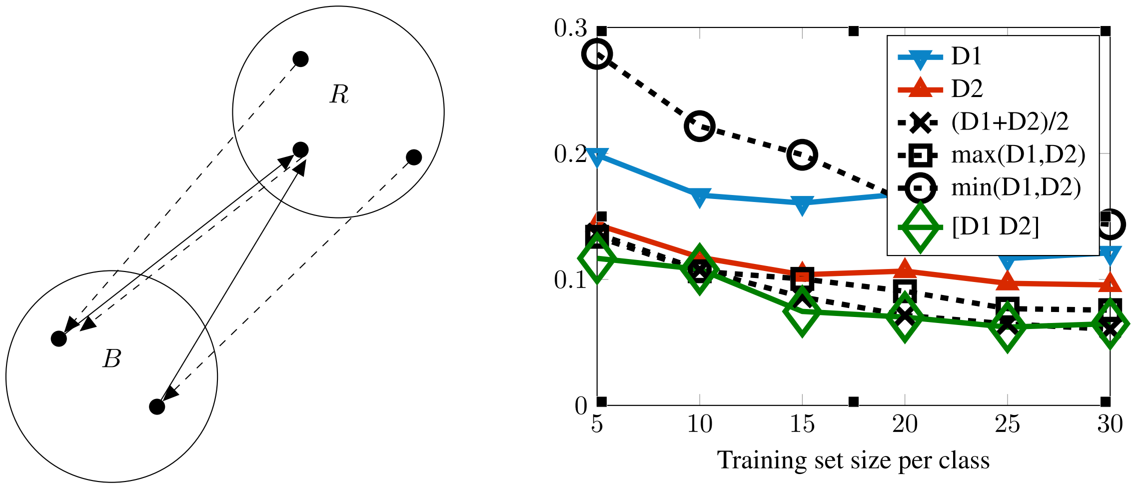

Figure 1: *Left*: Minimum instance distances between a bag B and a reference bag R, solid lines are from B to R, dashed lines are from R to B. *Right*: SVM Classification error plotted against the training set size.

In MIL, only the bag labels $y(B)$ are given, although a relationship between the bag label and the instance labels is often assumed. In particular, *concept* instances are assumed to be most important for $y(B)$. For example, in images labeled "tiger", concept instances are parts of the image that correspond to the tiger. Due to the asymmetry of the dissimilarities in the left of Fig. 1, not all the instances influence $d(B, R)$. However, if the topmost instance in Fig. 1 is a concept instance, and concept instances are indeed very important, $d(R, B)$ will potentially be more informative.

In the dissimilarity space, there are several choices for using the asymmetry information:

- Directed dissimilarities of the objects to the prototypes (D_1), or of the prototypes to the objects (D_2).
- Symmetrizing the two directions by $\frac{1}{2}(D_1 + D_2)$, $\max(D_1, D_2)$ or $\min(D_1, D_2)$.
- Concatenating D_1 and D_2 into an extended asymmetric dissimilarity space (EADS), doubling the dimensionality ($[D_1, D_2]$). EADS has the potential to preserve more information than the other techniques, because the classifier is able to decide which of the directed dissimilarities is more informative.

3 Results and Discussion

Some typical results for a MIL dataset are shown in the right plot of Fig. 1. Here, and in many other cases EADS outperforms the other representations under consideration. For this data, the direction from the prototypes to the bags (D_2) is more informative because the prototype concept instances are included. The opposite direction, and hence also averaging of the directed distances, are harmful for performance, whereas EADS still produces good results.

References

[1] Yenisel Plasencia Calaña, Veronika Cheplygina, Robert P. W. Duin, Edel García-Reyes, Mauricio Orozco-Alzate, David M.J. Tax, and Marco Loog. On the informativeness of asymmetric dissimilarities. In *Similarity-Based Pattern Recognition*, volume 7953, pages 75–89. Springer, 2013.

[2] Elżbieta Pękalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific Publishing Co., Inc., 2005.