A class-conditional dissimilarity function for multi-instance learning

Kajsa Møllersen*, Jon-Yngve Hardeberg[1], Fred Godtliebsen[2]

*kajsa.mollersen@uit.no, Department of Community Medicine, UiT - The Arctic University of Norway; [1]Department of Computer Science, NTNU; [2]Department of Mathematics and Statistics , UiT - The Arctic University of Norway

In multi-instance (MI) learning, each object (bag) consists of multiple feature vectors (instances), and can be regarded as a set of points in a multidimensional space. A different viewpoint is that the instances are realisations of random vectors with corresponding probability distribution, and that a bag is the distribution, not the realisations. In MI classification, each bag in the training set has a class label, but the instances are unlabelled. By introducing the probability distribution space to bag-level classification problems, dissimilarities between probability distributions can be applied. The bag-to-bag Kullback-Leibler (KL) information is asymptotically the best classifier, but the typical sparseness of MI training sets is an obstacle. We introduce bag-to-class distribution dissimilarity to MI learning, emphasising the hierarchical nature of the random vectors that makes bags from the same class different.

We propose the class-conditional KL information

$$cKL(f_{bag}, f_{POS}|f_{NEG}) = \int \frac{f_{NEG}(\mathbf{x})}{f_{POS}(\mathbf{x})} f_{bag}(\mathbf{x}) \log \frac{f_{bag}(\mathbf{x})}{f_{POS}(\mathbf{x})} d\mathbf{x},$$

where f_{bag} is the probability density function (pdf) of a bag with unknown class label, and f_{POS} and f_{NEG} are the pdfs of the two classes in a binary classification problem. Simulations studies show that cKL performs better than the bag-to-class KL information, and the bag-to-bag KL information for sparse training sets. Images of breast tissue divided into segments, where the task is to classify it as malignant or benign, is a typical MI classification problem. Breast tissue images, obtained from

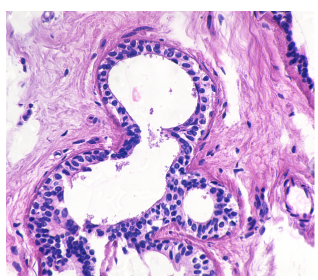

Figure 1: Breast tissue

miproblems.org and described in detail by Kandemir et al. are used as example. This is a sparse training set with only 58 images. Both cKL and the bag-to-class KL information exceeds the performance of Kandemir et al. in terms of area under the receiver-operating characteristic curve (AUC), obtained by 4-fold cross-validation. The AUCs are 0.97 for cKL, 0.93 for KL information, and 0.90 by Kandemir et al.'s proposed algorithm

M. Kandemir, C. Zhang, F. A. Hamprecht, Empowering multiple instance histopathology cancer diagnosis by cell graphs, in: P. Golland, N. Hata, C. Barillot, J. Hornegger, R. Howe (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2014, Springer International Publishing, 2014, pp. 228–235.