Data integration based on similarity matrices

Georgi Nalbantov (Maastricht University)

November 4, 2011

Consider a classification task where the data are represented with n different views (feature sets) associated with the same output variable. The standard approach to this task is to train a classifier for each view and then to combine all the resulting n classifiers using a type of majority voting rule. Another option is to simply concatinate all views into one (huge) dataset and perform analysis on this dataset. This could lead however to curse-of-dimensionality problems, among others. We propose an alternative approach: to combine the n views into one view using similarity matrices and then to train a classifier on the data represented by that combined view only. One advantage is that unlike the cocatinating approach, the dimensionality of the input space is equal to the number of objects in the dataset, whatever the number of different views. Another advantage is that different data types can be conbined in a standardized way. We evaluate this approach on a case study and report promising results, including increased accuracy.

Note: co-authors of this talk are Robert Harms and Evgueni Smirnov (in CC).