New Developments in COSA: Clustering Objects on Subsets of Attributes

Jacqueline J. Meulman[1] and Jerome H. Friedman[2]

[1] Data Theory Group, FSW, and Mathematical Instute, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands
[2] Department of Statistics, Stanford University, 290 Serra Mall, Stanford CA 94305, USA

Abstract. The motivation for clustering objects on subsets of attributes (COSA) was given by consideration of data where the number of attributes is much larger than the number of objects. Obvious application is in systems biology (genomics, proteomics, and metabolomics). When we have a large numbers of attributes, objects might cluster on some attributes, and be far apart on all others. Common data analysis approaches in systems biology are to cluster the attributes first, and only after having reduced the original many-attribute data set to a much smaller one, one tries to cluster the objects. The problem here, of course, is that we would like to select those attributes that discriminate most among the objects (so we have to do this while regarding all attributes multivariately), and it is usually not good enough to inspect each attribute univariately. Therefore, two tasks have to be carried out simultaneously: cluster the objects into homogeneous groups, while selecting different subsets of variables (one for each group of objects). The attribute subset for any discovered group may be completely, partially or nonoverlapping with those for other groups. The notorious local optima problem is dealt with by starting with the inverse exponential mean (rather than the arithmetic mean) of the separate attribute distances. By using a homotopy strategy, the algorithm creates a smooth transition of the inverse exponential distance to the mean of the ordinary Euclidean distances over attributes. New insight will be presented for the homotopy strategy, and the weights that are crucial in the COSA procedure but that were rather underexposed as diagnostics in the original paper.

References

FRIEDMAN, J.H., and MEULMAN, J.J. (2004a). Clustering objects on subsets of variables (with discussion). *Journal of the Royal Statistical Society, Series B, 66, 815–849.*

FRIEDMAN, J.H., and MEULMAN, J.J. (2004b). *The COSA program in an R-environment*, available at http://www-stat.stanford.edu/jhf/COSA.html.

Keywords

BIOINFORMATICS, FEATURE SELECTION, INVERSE EXPONENTIAL DISTANCE, SUBSET WEIGHTS, SUBSPACE CLUSTERING, TARGETED CLUSTERING