A Generalized Motif Bicluster Algorithm

Sebastian Kaiser[1,*], **Friedrich Leisch**[1]

1. Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, 80539 München, Germany.

* Contact author: Sebastian.Kaiser@stat.uni-muenchen.de

Keywords: Biclustering, Two Way Clustering, Ordinal Data

In many application domains different clusters in data may be defined by different sets of variables. E.g., in maketing one group of consumers could mainly be concerned about price and technical features of a product, while others care most about design and how "cool" the product is (almost regardless of the price). Standard clustering algorithms use all variables for all clusters and hence may fail to detect such structures in the data. Biclustering is the simultaneous clustering of columns and rows in a data set: each cluster is defined by a different subset of variables, these subsets can of course be overlapping. R package `biclust` (Kaiser & Leisch 2008, Kaiser et al 2008) contains a comprehensive collection of bicluster algorithms, preprocessing methods, and validation and visualization techniques for bicluster results.

The main focus of this presentation will be on recent additions to the package: There are new functions for bicluster validation and comparison. A new generalization of the well-known motif bicluster algorithm has been developed which is particularly suited for biclustering of marketing survey data. While the standard motif algorithm only searches for constant entries in the data matrix, our generalization is better suited for ordinal and metric data. The user can specify "neighborhood patterns" like intervals or density kernels of pre-specified size for metric data. In addition to finding more general patterns than constant groups only this also allows to calculate a posterior probabilities of cluster membership and can be seen as a first step towards fully model-based biclustering. All new methods will be demonstrated using real data from marketing applications.

References

Sebastian Kaiser and Friedrich Leisch (2008). A toolbox for bicluster analysis in R. In Paula Brito, editor, *Compstat 2008–Proceedings in Computational Statistics*, pages 201-208. Physica Verlag, Heidelberg, Germany.

Sebastian Kaiser, Rodrigo Santamaria, Roberto Theron, Luis Quintales and Friedrich Leisch (2008). biclust: BiCluster Algorithms. `http://cran.R-project.org/package=biclust`.

Sara C. Madeira and Arlindo L. Oliveira (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1),24–45.