R package gcExplorer: graphical and inferential exploration of cluster solutions

Theresa Scharl[1,2,*] **, Friedrich Leisch**[3]

1. Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
2. Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Muthgasse 18, A-1190 Vienna, Austria
3. Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany

* Contact author: theresa.scharl@ci.tuwien.ac.at

Keywords: cluster analysis, microarray data, visualization, neighborhood graphs, cluster validation

Cluster analysis is commonly applied to microarray data in order to find groups of co–expressed genes where cluster algorithms with the ability to visualize the resulting cluster objects (e.g., a dendrogram for hierarchical clustering) are usually preferred. The display of cluster solutions particularly for a large number of clusters is very important in exploratory data analysis. It gives practitioners an idea of the relationships between segments of a partition and allows to interpret the cluster results. Neighborhood graphs (Leisch, 2006) can be used for visual assessment of the cluster structure of centroid–based cluster solutions. In a neighborhood graph each node represents a cluster and two nodes are connected if there exist data points that have the two corresponding centroids as closest and second closest centroid.

In this work we present new visualization methods based on the neighborhood graph. For node representation different plot symbols visualizing single clusters are used allowing a quick overview of the data. On the one hand the corresponding data points themselves can be visualized using for example line diagrams for gene expression over time. On the other hand node symbols like pie charts can be used to visualize further properties of the clusters like association to functional groups under study. Finally the neighborhood graph can be used for the validation of a cluster solution, e.g., by testing the relationship between a clustering and a priori information about gene functions. All visualization methods and test procedures used are implemented in R package **gcExplorer** (Scharl and Leisch, 2009) which is now available on CRAN. The grid–based node symbols are implemented in R package **symbols** (`http://r-forge.r-project.org/projects/symbols/`).

References

Friedrich Leisch (2006). A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis*, 51(2): 526–544.

Theresa Scharl and Friedrich Leisch (2009). gcExplorer: Interactive Exploration of Gene Clusters. *Bioinformatics*, doi: 10.1093/bioinformatics/btp099.