gcExplorer: graphical exploration of centroid–based cluster solutions

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

[1]Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien

[2]Department of Biotechnology
University of Natural Resources and Applied Life Sciences,Vienna

[3]Institut für Statistik
Ludwig-Maximilians-Universität München

R Workshop, Boku Wien, January 16th, 2009

Outline

Motivation

Visualizing Cluster Solutions

- Interpretation of cluster results.
- Understanding of the cluster structure.
- Relationships between segments of a partition.

R package `gcExplorer`

- Visualize cluster solutions.
- Explore clusters interactively.
- Investigate additional properties of clusters.

Cluster algorithms

Centroid-based cluster algorithms

Cluster algorithms like K–means and PAM or others where clusters can be represented by centroids (e.g., QT–Clust, Heyer et al., Genome Research, 1999).

Task

Minimize the average distance between each data point and its closest centroid

$$D(X_n, C_K) = \frac{1}{N} \sum_{n=1}^{N} d(x_n, c(x_n)) \rightarrow \min_{C_K}$$

Graphical representation of a partition

- Projection of the data into two dimensions.
- Methods: e.g., principal components analysis, multidimensional scaling, linear discriminant analysis.
- Note: points that are close to each other in the 2-dimensional projection may have arbitrary distance in the original space.
- Note: linear projection into 2-d may not scale well in the number of clusters.

Neighborhood graphs

(Leisch, 2006)

- Neighborhood graphs use mean relative distances as edge weights.
- Assume we are given a data set $X_N = \{x_1, \ldots, x_N\}$ and
- a set of centroids $C_K = \{c_1, \ldots, c_K\}$.
- The centroid closest to x is denoted by

$$c(x) = \operatorname*{argmin}_{c \in C_K} d(x, c).$$

- And the second closest centroid to x is denoted by

$$\tilde{c}(x) = \operatorname*{argmin}_{c \in C_K \setminus \{c(x)\}} d(x, c).$$

TRNs and silhouette plots

Topology–representing networks

(Martinetz and Schulten, 1994)

- Count the number of data points a pair of centroids is closest and second–closest.
- Centroid pairs with a positive count are connected.

Silhouette plots (Rousseeuw, 1987)

- Compare the distance from each point to the points in its own cluster to the distance to points in the second closest cluster.
- The larger the silhouette values the better a cluster is separated from the other clusters.

Neighborhood graphs

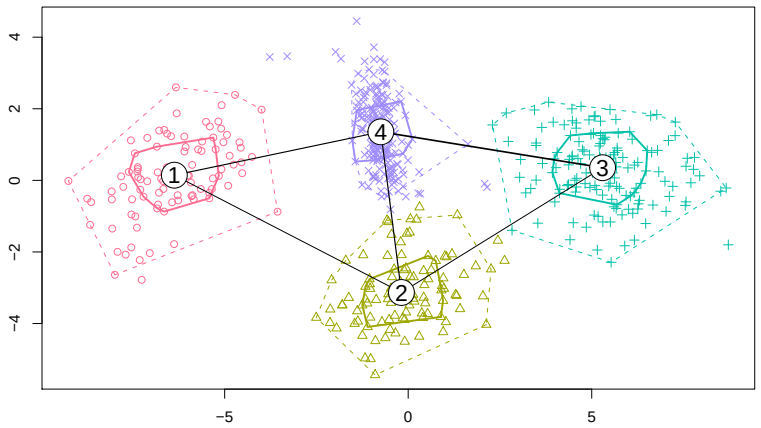
- The set of all points where c_i is the closest centroid and c_j is second–closest is given by

$$A_{ij} = \{x_n | c(x_n) = c_i, \tilde{c}(x_n) = c_j\}.$$

- Now we define edge weights

$$s_{ij} = \begin{cases} |A_{ij}|^{-1} \sum_{x \in A_{ij}} \frac{2d(x,c(x))}{d(x,c(x)) + d(x,\tilde{c}(x))}, & A_{ij} \neq \emptyset \\ 0, & A_{ij} = \emptyset \end{cases}$$

Example: artificial 2-dimensional data



gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

R package `flexclust`

- Flexible toolbox to investigate the influence of distance measures and cluster algorithms.
- Extensible implementations of the generalized k–Means and QT–Clust algorithm.
- Possibility to try out a variety of distance or similarity measures.
- Cluster algorithms are treated separately from distance measures.
- New distance measures can easily be incorporated into cluster procedures.
- Graphical representation of cluster objects using neighborhood graphs.

gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

R package `gcExplorer`

An interactive visualization toolbox for clusters

- New visualization techniques to display cluster results of high dimensional data.
- Nonlinear arrangements of the cluster centroids using Bioconductor packages `Rgraphviz` and `graph`
- Show similarities between clusters.
- Visualize properties of clusters like cluster size or cluster tightness.
- Classification to functional groups (e.g., Gene Ontology).

`gcExporer` is now available on CRAN **http://cran.r-project.org/package=gcExplorer**.
See the `README` file in the package for detailed installation instructions.

gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

Bioconductor packages `Rgraphviz` and `graph`

- Bioconductor project: **http://www.bioconductor.org**
- Infrastructure for creating, manipulating and visualizing graphs.
- Efficient representations of very large graphs.
- Interface to Graphviz (www.graphviz.org)
- Choice of several non–linear layout algorithms.
- Global and local properties (e.g. labels, shape, color, . . .) for both nodes and edges.

E. coli cultivation data

(Dürrschmid et al., 2008)

- A recombinant E. coli process.
- Stress response was measured during expression of the human recombinant protein SOD (human superoxide dismutase, hSOD) on the host metabolism.
- Non-induced state was compared to samples past induction.
- After preprocessing the data consisting of 527 genes at 6 time points was clustered using stochastic QT–Clust.
- The genes were separated into 16 clusters.

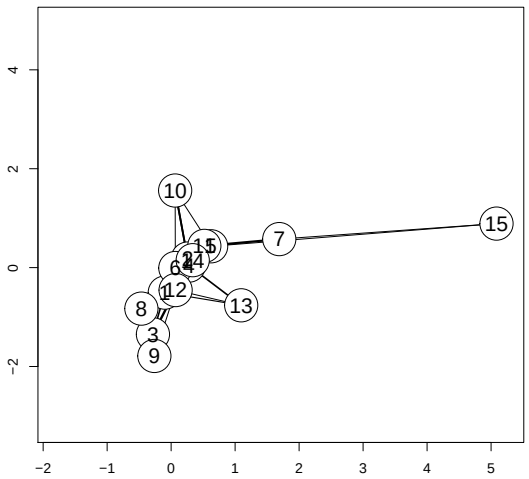
gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

How to use `gcExplorer`

```
R> library("gcExplorer")
R> data("hsod")
R> set.seed(1111)
R> cl1 <- qtclust(hsod, radius = 2,
+                 save.data=TRUE)
R> gcExplorer(cl1, filt = 0.1)
```

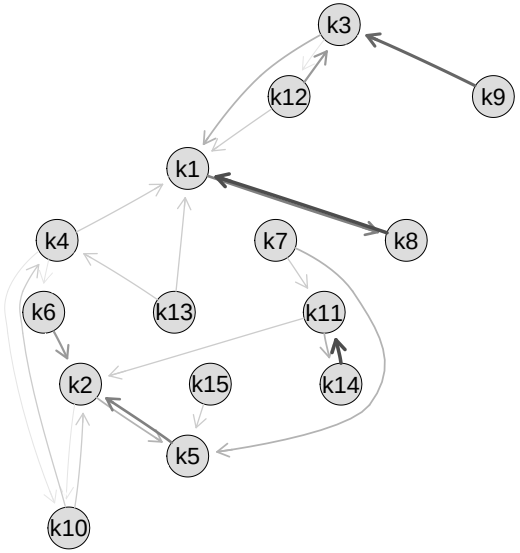
gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

Cluster solution using PCA



gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

Cluster solution using gcExplorer



Functionality of gcExplorer

Node functions

- Highlight clusters with specific properties, e.g. cluster size or cluster tightness.
- Draw arbitrary cluster plots in nodes.
- Highlight external information about gene functions.

Panel functions

- Allow arbitrary panel functions, e.g., matrix plots, boxplots or HTML tables.

Edge options

- Drawn edges if the similarity between clusters is above a certain threshold, e.g. 10%.
- Plot directed or undirected graphs.

gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

How to use gcExplorer

Interactive gcExplorer

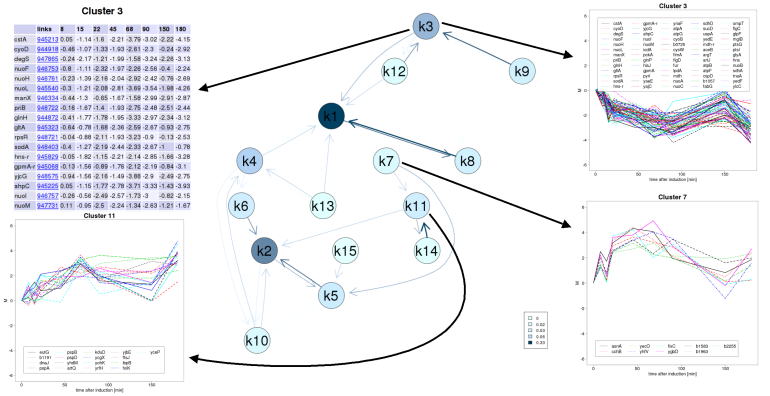
```
R> gcExplorer(cl1, dev = "many",
+       panel.function = gcProfile,
+       node.function = node.size,
+       legend.pos= "topleft")
```

Use of matrix plot as node function

```
R> gcExplorer(cl1, node.function = gmatplot,
+       doViewPort = TRUE, filt = 0.1)
```

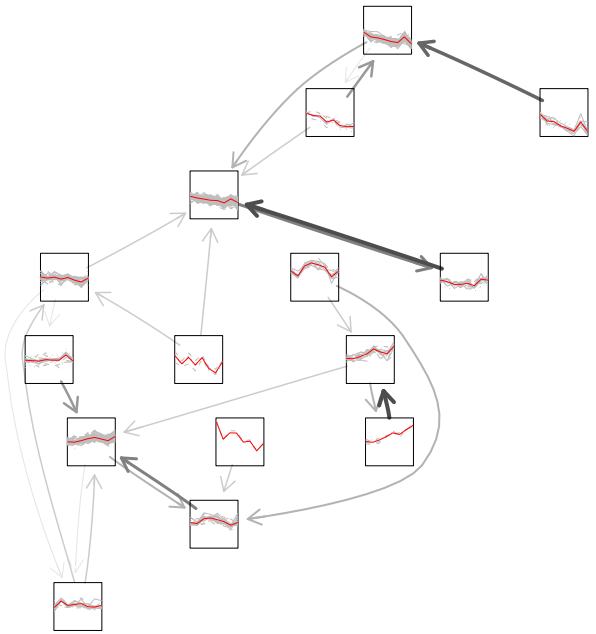
gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

Interactive cluster toolbox



gcExplorer: graphical exploration of centroid–based cluster solutions
Scharl, Leisch
Motivation
Cluster Algorithms
Neighborhood graphs
Software
Application
Summary

Matrix plot as node function



Neighborhood graph for general cluster functions

Cluster results from cluster functions like `kmeans` from package `stats` or `pam` from package `cluster` can be converted to objects of class `kcca` and visualized using the neighborhood graph:

Conversion

```
R> k1 <- kmeans(hsod, centers = 15)
R> k2 <- as.kcca(k1, data = hsod,
                 save.data = TRUE)
R> gcExplorer(k2)
```

gcExplorer: graphical exploration of centroid–based cluster solutions

Scharl, Leisch

Motivation

Cluster Algorithms

Neighborhood graphs

Software

Application

Summary

Summary

- **Neighborhood graphs** help to reveal structure in cluster solutions.
- **gcExplorer** is a flexible tool for the interactive exploration of clusters allowing arbitrary panel and node functions.

- Download and try **http://cran.r-project.org/package=gcExplorer**