

Package ‘flexclust’

July 22, 2025

Version 1.5.0

Date 2025-02-27

Encoding UTF-8

Title Flexible Cluster Algorithms

Depends R (>= 2.14.0)

Imports graphics, grid, lattice, methods, modeltools, parallel, stats,
stats4, class

Suggests ellipse, clue, cluster, seriation, skmeans

Description The main function `kcca` implements a general framework for k-centroids cluster analysis supporting arbitrary distance measures and centroid computation. Further cluster methods include hard competitive learning, neural gas, and QT clustering. There are numerous visualization methods for cluster results (neighborhood graphs, convex cluster hulls, barcharts of centroids, ...), and bootstrap methods for the analysis of cluster stability.

License GPL-2

LazyLoad yes

NeedsCompilation yes

Author Friedrich Leisch [aut] (ORCID: <<https://orcid.org/0000-0001-7278-1983>>, maintainer up to 2024),
Evgenia Dimitriadou [ctb],
Lena Ortega Menjivar [ctb] (ORCID:
<<https://orcid.org/0000-0001-5785-4021>>),
Dominik Ernst [ctb] (ORCID: <<https://orcid.org/0009-0005-5579-2636>>),
Bettina Grün [ctb, cre] (ORCID:
<<https://orcid.org/0000-0001-7265-4773>>)

Maintainer Bettina Grün <Bettina.Gruen@R-project.org>

Repository CRAN

Date/Publication 2025-02-28 06:40:02 UTC

Contents

| | |
|------------------------|----|
| achieve | 3 |
| auto | 3 |
| barplot-methods | 5 |
| bclust | 8 |
| birth | 10 |
| bootFlexclust | 10 |
| bundestag | 12 |
| bwplot-methods | 14 |
| cclust | 15 |
| clusterSim | 17 |
| conversion | 18 |
| dentitio | 20 |
| dist2 | 21 |
| distances | 22 |
| flexclustControl-class | 23 |
| flxColors | 25 |
| histogram-methods | 26 |
| image-methods | 27 |
| info | 28 |
| kcca | 29 |
| kcca2df | 32 |
| milk | 32 |
| Nclus | 33 |
| nutrient | 34 |
| pairs-methods | 34 |
| parameters | 35 |
| plot-methods | 35 |
| predict-methods | 37 |
| priceFeature | 37 |
| projAxes | 38 |
| propBarchart | 40 |
| qtclust | 42 |
| randIndex | 43 |
| randomTour | 45 |
| relabel | 47 |
| shadow | 48 |
| shadowStars | 50 |
| slsaplot | 52 |
| slswFlexclust | 53 |
| stepFlexclust | 54 |
| stripes | 56 |
| vacmot | 58 |
| volunteers | 59 |

achieve

Achievement Test Scores for New Haven Schools

Description

Measurements at the beginning of the 4th grade (when the national average is 4.0) and of the 6th grade in 25 schools in New Haven.

Usage

```
data(achieve)
```

Format

A data frame with 25 observations on the following 4 variables.

read4 4th grade reading.

arith4 4th grade arithmetic.

read6 6th grade reading.

arith6 6th grade arithmetic.

References

John A. Hartigan: Clustering Algorithms. Wiley, New York, 1975.

auto

Automobile Customer Survey Data

Description

A German manufacturer of premium cars asked customers approximately 3 months after a car purchase which characteristics of the car were most important for the decision to buy the car. The survey was done in 1983 and the data set contains all responses without missing values.

Usage

```
data(auto)
```

Format

A data frame with 793 observations on the following 46 variables.

model A factor with levels A, B, C, or D; model bought by the customer.

gear A factor with levels 4 gears, 5 econo, 5 sport, or automatic.

leasing A logical vector, was leasing used to finance the car?

usage A factor with levels private, both, business.

previous_model A factor describing which type of car was owned directly before the purchase.

other_consider A factor with levels same manuf, other manuf, both, or none.

test_drive A logical vector, did you do a test drive?

info_adv A logical vector, was advertising an important source of information?

info_exp A logical vector, was experience an important source of information?

info_rec A logical vector, were recommendations an important source of information?

ch_clarity A logical vector.

ch_economy A logical vector.

ch_driving_properties A logical vector.

ch_service A logical vector.

ch_interior A logical vector.

ch_quality A logical vector.

ch_technology A logical vector.

ch_model_continuity A logical vector.

ch_comfort A logical vector.

ch_reliability A logical vector.

ch_handling A logical vector.

ch_reputation A logical vector.

ch_concept A logical vector.

ch_character A logical vector.

ch_power A logical vector.

ch_resale_value A logical vector.

ch_styling A logical vector.

ch_safety A logical vector.

ch_sporty A logical vector.

ch_consumption A logical vector.

ch_space A logical vector.

satisfaction A numeric vector describing overall satisfaction (1=very good, 10=very bad).

good1 Conception, styling, dimensions.

good2 Auto body.

good3 Driving and coupled axles.

good4 Engine.
 good5 Electronics.
 good6 Financing and customer service.
 good7 Other.
 sporty What do you think about the balance of sportiness and comfort? (good, more sport, more comfort).
 drive_char Driving characteristics (gentle < speedy < powerful < extreme).
 tempo Which average speed do you prefer on German Autobahn in km/h? (< 130 < 130-150 < 150-180 < > 180)
 consumption An ordered factor with levels low < ok < high < too high.
 gender A factor with levels male and female
 occupation A factor with levels self-employed, freelance, and employee.
 household Size of household, an ordered factor with levels 1-2 < >=3.

Source

The original German data are in the public domain and available from LMU Munich (doi:10.5282/ubm/data.14). The variable names and help page were translated to English and converted into Rd format by Friedrich Leisch.

References

Open Data LMU (1983): Umfrage unter Kunden einer Automobilfirma, doi:10.5282/ubm/data.14

Examples

```
data(auto)
summary(auto)
```

barplot-methods

Barplot/chart Methods in Package 'flexclust'

Description

Barplot of cluster centers or other cluster statistics.

Usage

```
## S4 method for signature 'kcca'
barplot(height, bycluster = TRUE, oneplot = TRUE,
  data = NULL, FUN = colMeans, main = deparse(substitute(height)),
  which = 1:height@k, names.arg = NULL,
  oma = par("oma"), col = NULL, mcol = "darkred", srt = 45, ...)

## S4 method for signature 'kcca'
```

```

barchart(x, data, xlab="", strip.labels=NULL,
         strip.prefix="Cluster ", col=NULL, mcol="darkred", mlcol=mcol,
         which=NULL, legend=FALSE, shade=FALSE, diff=NULL, byvar=FALSE,
         clusters=1:x@k, ...)
## S4 method for signature 'hclust'
barchart(x, data, xlab="", strip.labels=NULL,
         strip.prefix="Cluster ", col=NULL, mcol="darkred", mlcol=mcol,
         which=NULL, shade=FALSE, diff=NULL, byvar=FALSE, k=2, ...)
## S4 method for signature 'bclust'
barchart(x, data, xlab="", strip.labels=NULL,
         strip.prefix="Cluster ", col=NULL, mcol="darkred", mlcol=mcol,
         which=NULL, legend=FALSE, shade=FALSE, diff=NULL, byvar=FALSE,
         k=x@k, clusters=1:k, ...)

```

Arguments

| | |
|----------------------|--|
| height, x | An object of class "kcca". |
| bycluster | If TRUE, then each barplot shows one cluster. If FALSE, then each barplot compares all cluster for one input variable. |
| oneplot | If TRUE, all barplots are plotted together on one page, else each plot is on a separate page. |
| data | If not NULL, cluster membership is predicted for the new data and used for the plots. By default the values from the training data are used. Ignored by the barchart method. |
| FUN | The function to be applied to each cluster for calculating the bar heights. Only used, if data is not NULL. |
| which | For barplot index number of clusters for the plot, for barchart index numbers or names of variables to plot. |
| names.arg | A vector of names to be plotted below each bar. |
| main, oma, xlab, ... | Graphical parameters. |
| col | Vector of colors for the clusters. |
| mcol, mlcol | If not NULL, the value of FUN for the complete data set is plotted over each bar as a point with color mcol and a line segment starting in zero with color mlcol. |
| srt | Number between 0 and 90, rotation of the x-axis labels. |
| strip.labels | Vector of strings for the strips of the Trellis display. |
| strip.prefix | Prefix string for the strips of the Trellis display. |
| legend | If TRUE, the barchart is always plotted on the current graphics device and a legend is added to the bottom of the plot. |
| shade | If TRUE, only bars with large absolute or relative deviation deviation from the sample mean of the respective variables are plotted in color. |
| diff | A numerical vector of length two with absolute and relative deviations for shading, default is $max/4$ absolute deviation and 50% relative deviation. |

| | |
|----------|--|
| byvar | If TRUE, a panel is plotted for each variable. By default a panel is plotted for each cluster. |
| clusters | Integer vector of clusters to plot. |
| k | Integer specifying the desired number of clusters. |

Note

The flexclust barchart method uses a horizontal arrangements of the bars, and sorts them from top to bottom. Default barcharts in lattice are the other way round (bottom to top). See the examples below how this affects, e.g., manual labels for the y axis.

The barplot method is legacy code and only maintained to keep up with changes in R, all active development is done on barchart.

Author(s)

Friedrich Leisch

References

Sara Dolnicar and Friedrich Leisch. Using graphical statistics to better understand market segmentation solutions. *International Journal of Market Research*, 56(2), 97-120, 2014.

Examples

```

cl <- cclust(iris[,-5], k=3)
barplot(cl)
barplot(cl, bycluster=FALSE)

## plot the maximum instead of mean value per cluster:
barplot(cl, bycluster=FALSE, data=iris[,-5],
        FUN=function(x) apply(x,2,max))

## use lattice for plotting:
barchart(cl)
## automatic abbreviation of labels
barchart(cl, scales=list(abbreviate=TRUE))
## origin of bars at zero
barchart(cl, scales=list(abbreviate=TRUE), origin=0)

## Use manual labels. Note that the flexclust barchart orders bars
## from top to bottom (the default does it the other way round), hence
## we have to rev() the labels:
LAB <- c("SL", "SW", "PL", "PW")
barchart(cl, scales=list(y=list(labels=rev(LAB))), origin=0)

## deviation of each cluster center from the population means
barchart(cl, origin=rev(cl@xcent), mlcol=NULL)

## use shading to highlight large deviations from population mean
barchart(cl, shade=TRUE)

```

```
## use smaller deviation limit than default and add a legend
barchart(cl, shade=TRUE, diff=0.2, legend=TRUE)
```

bclust

Bagged Clustering

Description

Cluster the data in `x` using the bagged clustering algorithm. A partitioning cluster algorithm such as [cclust](#) is run repeatedly on bootstrap samples from the original data. The resulting cluster centers are then combined using the hierarchical cluster algorithm [hclust](#).

Usage

```
bclust(x, k = 2, base.iter = 10, base.k = 20, minsize = 0,
      dist.method = "euclidian", hclust.method = "average",
      FUN = "cclust", verbose = TRUE, final.cclust = FALSE,
      resample = TRUE, weights = NULL, maxcluster = base.k, ...)
## S4 method for signature 'bclust,missing'
plot(x, y, maxcluster = x@maxcluster, main = "", ...)
## S4 method for signature 'bclust,missing'
clusters(object, newdata, k, ...)
## S4 method for signature 'bclust'
parameters(object, k)
```

Arguments

| | |
|----------------------------|---|
| <code>x</code> | Matrix of inputs (or object of class "bclust" for plot). |
| <code>k</code> | Number of clusters. |
| <code>base.iter</code> | Number of runs of the base cluster algorithm. |
| <code>base.k</code> | Number of centers used in each repetition of the base method. |
| <code>minsize</code> | Minimum number of points in a base cluster. |
| <code>dist.method</code> | Distance method used for the hierarchical clustering, see dist for available distances. |
| <code>hclust.method</code> | Linkage method used for the hierarchical clustering, see hclust for available methods. |
| <code>FUN</code> | Partitioning cluster method used as base algorithm. |
| <code>verbose</code> | Output status messages. |
| <code>final.cclust</code> | If TRUE, a final <code>cclust</code> step is performed using the output of the bagged clustering as initialization. |
| <code>resample</code> | Logical, if TRUE the base method is run on bootstrap samples of <code>x</code> , else directly on <code>x</code> . |
| <code>weights</code> | Vector of length <code>nrow(x)</code> , weights for the resampling. By default all observations have equal weight. |

| | |
|------------|---|
| maxcluster | Maximum number of clusters memberships are to be computed for. |
| object | Object of class "bclust". |
| main | Main title of the plot. |
| ... | Optional arguments top be passed to the base method in bclust, ignored in plot. |
| y | Missing. |
| newdata | An optional data matrix with the same number of columns as the cluster centers. If omitted, the fitted values are used. |

Details

First, `base.iter` bootstrap samples of the original data in `x` are created by drawing with replacement. The base cluster method is run on each of these samples with `base.k` centers. The `base.method` must be the name of a partitioning cluster function returning an object with the same slots as the return value of `cclust`.

This results in a collection of `iter.base * base.centers` centers, which are subsequently clustered using the hierarchical method `hclust`. Base centers with less than `minsize` points in there respective partitions are removed before the hierarchical clustering. The resulting dendrogram is then cut to produce `k` clusters.

Value

`bclust` returns objects of class "bclust" including the slots

| | |
|-------------------------|---|
| <code>hclust</code> | Return value of the hierarchical clustering of the collection of base centers (Object of class "hclust"). |
| <code>cluster</code> | Vector with indices of the clusters the inputs are assigned to. |
| <code>centers</code> | Matrix of centers of the final clusters. Only useful, if the hierarchical clustering method produces convex clusters. |
| <code>allcenters</code> | Matrix of all <code>iter.base * base.centers</code> centers found in the base runs. |

Author(s)

Friedrich Leisch

References

Friedrich Leisch. Bagged clustering. Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", August 1999. doi:[10.57938/9b129f95b53b-44cea1295b7a1168d832](https://doi.org/10.57938/9b129f95b53b-44cea1295b7a1168d832)

Sara Dolnicar and Friedrich Leisch. Winter tourist segments in Austria: Identifying stable vacation styles using bagged clustering techniques. *Journal of Travel Research*, 41(3):281-292, 2003.

See Also

[hclust](#), [cclust](#)

Examples

```
data(iris)
bc1 <- bclust(iris[,1:4], 3, base.k=5)
plot(bc1)

table(clusters(bc1, k=3))
parameters(bc1, k=3)
```

birth

Birth and Death Rates

Description

Birth and death rates for 70 countries.

Usage

```
data(birth)
```

Format

A data frame with 70 observations on the following 2 variables.

birth Birth rate (in percent).

death Death rate (in percent).

References

John A. Hartigan: Clustering Algorithms. Wiley, New York, 1975.

bootFlexclust

Bootstrap Flexclust Algorithms

Description

Runs clustering algorithms repeatedly for different numbers of clusters on bootstrap replica of the original data and returns corresponding cluster assignments, centroids and (adjusted) Rand indices comparing pairs of partitions.

Usage

```
bootFlexclust(x, k, nboot=100, correct=TRUE, seed=NULL,
             multicore=TRUE, verbose=FALSE, ...)
```

```
## S4 method for signature 'bootFlexclust'
summary(object)
## S4 method for signature 'bootFlexclust,missing'
plot(x, y, ...)
## S4 method for signature 'bootFlexclust'
boxplot(x, ...)
## S4 method for signature 'bootFlexclust'
densityplot(x, data, ...)
```

Arguments

| | |
|-----------|--|
| x, k, ... | Passed to stepFlexclust . |
| nboot | Number of bootstrap pairs of partitions. |
| correct | Logical, correct the Rand index for agreement by chance also called adjusted Rand index)? |
| seed | If not NULL, a call to <code>set.seed()</code> is made before any clustering is done. |
| multicore | If TRUE, use package parallel for parallel processing. In addition, it may be a workstation cluster object as returned by makeCluster , see examples below. |
| verbose | If TRUE, show progress information during computations. Will not work with <code>multicore=TRUE</code> . |
| y, data | Not used. |
| object | An object of class "bootFlexclust". |

Details

Availability of **multicore** is checked when **flexclust** is loaded. This information is stored and can be obtained using `getOption("flexclust")$have_multicore`. Set to FALSE for debugging and more sensible error messages in case something goes wrong.

Author(s)

Friedrich Leisch

See Also

[stepFlexclust](#)

Examples

```
## Not run:

## data uniform on unit square
x <- matrix(runif(400), ncol=2)
```

```

cl <- FALSE

## to run bootstrap replications on a workstation cluster do the following:
library("parallel")
cl <- makeCluster(2, type = "PSOCK")
clusterCall(cl, function() require("flexclust"))

## 50 bootstrap replicates for speed in example,
## use more for real applications
bcl <- bootFlexclust(x, k=2:7, nboot=50, FUN=cclust, multicore=cl)

bcl
summary(bcl)

## splitting the square into four quadrants should be the most stable
## solution (increase nboot if not)
plot(bcl)
densityplot(bcl, from=0)

## End(Not run)

```

bundestag

German Parliament Election Data

Description

Results of the elections 2002, 2005 or 2009 for the German Bundestag, the first chamber of the German parliament.

Usage

```

data(btw2002)
data(btw2005)
data(btw2009)
bundestag(year, second=TRUE, percent=TRUE, nzero=TRUE, state=FALSE)

```

Arguments

| | |
|---------|---|
| year | Numeric or character, year of the election. |
| second | Logical, return second or first votes? |
| percent | Logical, return percentages or absolute numbers? |
| nzero | Logical, convert NAs to 0? |
| state | Logical or character. If TRUE then only column state from the corresponding data frame is returned, and all other arguments are ignored. If character, then it is used as pattern to grep for the corresponding state(s), see examples. |

Format

btw200x are data frames with 299 rows (corresponding to constituencies) and 17 columns. All columns except state are numeric.

state Factor, the 16 German federal states.

eligible Number of citizens eligible to vote.

votes Number of eligible citizens who did vote.

invalid1, invalid2 Number of invalid first and second votes (see details below).

valid1, valid2 Number of valid first and second votes.

SPD1, SPD2 Number of first and second votes for the Social Democrats.

UNION1, UNION2 Number of first and second votes for CDU/CSU, the conservative Christian Democrats.

GRUENE1, GRUENE2 Number of first and second votes for the Green Party.

FDP1, FDP2 Number of first and second votes for the Liberal Party.

LINKE1, LINKE2 Number of first and second votes for the Left Party (PDS in 2002).

Missing values indicate that a party did not candidate in the corresponding constituency.

Details

btw200x are the original data sets. bundestag() is a helper function which extracts first or second votes, calculates percentages (number of votes for a party divided by number of valid votes), replaces missing values by zero, and converts the result from a data frame to a matrix. By default it returns the percentage of second votes for each party, which determines the number of seats each party gets in parliament.

German Federal Elections

Half of the Members of the German Bundestag are elected directly from Germany's 299 constituencies, the other half on the parties' state lists. Accordingly, each voter has two votes in the elections to the German Bundestag. The first vote, allowing voters to elect their local representatives to the Bundestag, decides which candidates are sent to Parliament from the constituencies.

The second vote is cast for a party list. And it is this second vote that determines the relative strengths of the parties represented in the Bundestag. At least 598 Members of the German Bundestag are elected in this way. In addition to this, there are certain circumstances in which some candidates win what are known as "overhang mandates" when the seats are being distributed.

References

Homepage of the Bundestag: <https://www.bundestag.de>

Examples

```
p02 <- bundestag(2002)
pairs(p02)
p05 <- bundestag(2005)
pairs(p05)
p09 <- bundestag(2009)
```

```

pairs(p09)

state <- bundestag(2002, state=TRUE)
table(state)

start.with.b <- bundestag(2002, state="^B")
table(start.with.b)

pairs(p09, col=2-(state=="Bayern"))

```

 bwplot-methods

Box-Whisker Plot Methods in Package 'flexclust'

Description

Separate boxplot of variables in each cluster in comparison with boxplot for complete sample.

Usage

```

## S4 method for signature 'kcca'
bwplot(x, data, xlab="",
        strip.labels=NULL, strip.prefix="Cluster ",
        col=NULL, shade=!is.null(shadefun), shadefun=NULL, byvar=FALSE, ...)
## S4 method for signature 'bclust'
bwplot(x, k=x@k, xlab="", strip.labels=NULL,
        strip.prefix="Cluster ", clusters=1:k, ...)

```

Arguments

| | |
|--------------|---|
| x | An object of class "kcca" or "bclust". |
| data | If not NULL, cluster membership is predicted for the new data and used for the plots. By default the values from the training data are used. |
| xlab, ... | Graphical parameters. |
| col | Vector of colors for the clusters. |
| strip.labels | Vector of strings for the strips of the Trellis display. |
| strip.prefix | Prefix string for the strips of the Trellis display. |
| shade | If TRUE, only boxes with larger deviation from the median or quartiles of the total population of the respective variables are filled with color. |
| shadefun | A function or name of a function to compute which boxes are shaded, e.g. "medianInside" (default) or "boxOverlap". |
| byvar | If TRUE, a panel is plotted for each variable. By default a panel is plotted for each group. |
| k | Number of clusters. |
| clusters | Integer vector of clusters to plot. |

Examples

```

set.seed(1)
cl <- cclust(iris[,-5], k=3, save.data=TRUE)
bwplot(cl)
bwplot(cl, byvar=TRUE)

## fill only boxes with color which do not contain the overall median
## (grey dot of background box)
bwplot(cl, shade=TRUE)

## fill only boxes with color which do not overlap with the box of the
## complete sample (grey background box)
bwplot(cl, shadefun="boxOverlap")

```

cclust

Convex Clustering

Description

Perform k-means clustering, hard competitive learning or neural gas on a data matrix.

Usage

```

cclust(x, k, dist = "euclidean", method = "kmeans",
       weights=NULL, control=NULL, group=NULL, simple=FALSE,
       save.data=FALSE)

```

Arguments

| | |
|-----------|---|
| x | A numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns). |
| k | Either the number of clusters, or a vector of cluster assignments, or a matrix of initial (distinct) cluster centroids. If a number, a random set of (distinct) rows in x is chosen as the initial centroids. |
| dist | Distance measure, one of "euclidean" (mean square distance) or "manhattan" (absolute distance). |
| method | Clustering algorithm: one of "kmeans", "hardcl" or "neuralgas", see details below. |
| weights | An optional vector of weights for the observations (rows of the x) to be used in the fitting process. Works only in combination with hard competitive learning. |
| control | An object of class "cclustControl". |
| group | Currently ignored. |
| simple | Return an object of class "kccsimple"? |
| save.data | Save a copy of x in the return object? |

Details

This function uses the same computational engine as the earlier function of the same name from package ‘cclust’. The main difference is that it returns an S4 object of class “kcca”, hence all available methods for “kcca” objects can be used. By default [kcca](#) and `cclust` use exactly the same algorithm, but `cclust` will usually be much faster because it uses compiled code.

If `dist` is “euclidean”, the distance between the cluster center and the data points is the Euclidian distance (ordinary kmeans algorithm), and cluster means are used as centroids. If “manhattan”, the distance between the cluster center and the data points is the sum of the absolute values of the distances, and the column-wise cluster medians are used as centroids.

If `method` is “kmeans”, the classic kmeans algorithm as given by MacQueen (1967) is used, which works by repeatedly moving all cluster centers to the mean of their respective Voronoi sets. If “hardcl”, on-line updates are used (AKA hard competitive learning), which work by randomly drawing an observation from `x` and moving the closest center towards that point (e.g., Ripley 1996). If “neuralgas” then the neural gas algorithm by Martinetz et al (1993) is used. It is similar to hard competitive learning, but in addition to the closest centroid also the second closest centroid is moved in each iteration.

Value

An object of class “kcca”.

Author(s)

Evgenia Dimitriadou and Friedrich Leisch

References

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds L. M. Le Cam & J. Neyman, 1, pp. 281–297. Berkeley, CA: University of California Press.

Martinetz T., Berkovich S., and Schulten K (1993). ‘Neural-Gas’ Network for Vector Quantization and its Application to Time-Series Prediction. *IEEE Transactions on Neural Networks*, 4 (4), pp. 558–569.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge.

See Also

[cclustControl-class](#), [kcca](#)

Examples

```
## a 2-dimensional example
x <- rbind(matrix(rnorm(100, sd=0.3), ncol=2),
           matrix(rnorm(100, mean=1, sd=0.3), ncol=2))
cl <- cclust(x,2)
plot(x, col=predict(cl))
points(cl@centers, pch="x", cex=2, col=3)
```



```
## a 3-dimensional example
x <- rbind(matrix(rnorm(150, sd=0.3), ncol=3),
           matrix(rnorm(150, mean=2, sd=0.3), ncol=3),
           matrix(rnorm(150, mean=4, sd=0.3), ncol=3))
cl <- cclust(x, 6, method="neuralgas", save.data=TRUE)
pairs(x, col=predict(cl))
plot(cl)
```

clusterSim

Cluster Similarity Matrix

Description

Returns a matrix of cluster similarities. Currently two methods for computing similarities of clusters are implemented, see details below.

Usage

```
## S4 method for signature 'kcca'
clusterSim(object, data=NULL, method=c("shadow", "centers"),
           symmetric=FALSE, ...)
## S4 method for signature 'kccasimple'
clusterSim(object, data=NULL, method=c("shadow", "centers"),
           symmetric=FALSE, ...)
```

Arguments

| | |
|-----------|--|
| object | Fitted object. |
| data | Data to use for computation of the shadow values. If the cluster object <code>x</code> was created with <code>save.data=TRUE</code> , then these are used by default. Ignored if <code>method="centers"</code> . |
| method | Type of similarities, see details below. |
| symmetric | Compute symmetric or asymmetric shadow values? Ignored if <code>method="centers"</code> . |
| ... | Currently not used. |

Details

If `method="shadow"` (the default), then the similarity of two clusters is proportional to the number of points in a cluster, where the centroid of the other cluster is second-closest. See Leisch (2006, 2008) for detailed formulas.

If `method="centers"`, then first the pairwise distances between all centroids are computed and rescaled to $[0,1]$. The similarity between two clusters is then simply 1 minus the rescaled distance.

Author(s)

Friedrich Leisch

References

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

Friedrich Leisch. Visualizing cluster analysis and finite mixture models. In Chun houh Chen, Wolfgang Haerdle, and Antony Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics. Springer Verlag, 2008.

Examples

```
example(Nclus)

clusterSim(cl)
clusterSim(cl, symmetric=TRUE)

## should have similar structure but will be numerically different:
clusterSim(cl, symmetric=TRUE, data=Nclus[sample(1:550, 200),])

## different concept of cluster similarity
clusterSim(cl, method="centers")
```

conversion

Conversion Between S3 Partition Objects and KCCA

Description

These functions can be used to convert the results from cluster functions like [kmeans](#) or [pam](#) to objects of class "kcca" and vice versa.

Usage

```
as.kcca(object, ...)

## S3 method for class 'hclust'
as.kcca(object, data, k, family=NULL, save.data=FALSE, ...)
## S3 method for class 'kmeans'
as.kcca(object, data, save.data=FALSE, ...)
## S3 method for class 'partition'
as.kcca(object, data=NULL, save.data=FALSE, ...)
## S3 method for class 'skmeans'
as.kcca(object, data, save.data=FALSE, ...)
## S4 method for signature 'kccasimple,kmeans'
coerce(from, to="kmeans", strict=TRUE)

Cutree(tree, k=NULL, h=NULL)
```

Arguments

| | |
|------------------|---|
| object | Fitted object. |
| data | Data which were used to obtain the clustering. For "partition" objects created by functions from package cluster this is optional, if object contains the data. |
| save.data | Save a copy of the data in the return object? |
| k | Number of clusters. |
| family | Object of class "kccaFamily", can be omitted for some known distances. |
| ... | Currently not used. |
| from, to, strict | Usual arguments for coerce |
| tree | A tree as produced by hclust . |
| h | Numeric scalar or vector with heights where the tree should be cut. |

Details

The standard [cutree](#) function orders clusters such that observation one is in cluster one, the first observation (as ordered in the data set) not in cluster one is in cluster two, etc. [Cutree](#) orders clusters as shown in the dendrogram from left to right such that similar clusters have similar numbers. The latter is used when converting to [kcca](#).

For hierarchical clustering the cluster memberships of the converted object can be different from the result of [Cutree](#), because one [KCCA](#)-iteration has to be performed in order to obtain a valid [kcca](#) object. In this case a warning is issued.

Author(s)

Friedrich Leisch

Examples

```
data(Nclus)

c11 <- kmeans(Nclus, 4)
c11
c11a <- as.kcca(c11, Nclus)
c11a
c11b <- as(c11a, "kmeans")

library("cluster")
c12 <- pam(Nclus, 4)
c12
c12a <- as.kcca(c12)
c12a
## the same
c12b <- as.kcca(c12, Nclus)
c12b
```

```
## hierarchical clustering
hc <- hclust(dist(USArrests))
plot(hc)
rect.hclust(hc, k=3)
c3 <- Cutree(hc, k=3)
k3 <- as.kcca(hc, USArrests, k=3)
barchart(k3)
table(c3, clusters(k3))
```

dentitio

Dentition of Mammals

Description

Mammal's teeth divided into the 4 groups: incisors, canines, premolars and molars.

Usage

```
data(dentitio)
```

Format

A data frame with 66 observations on the following 8 variables.

top.inc Top incisors.

bot.inc Bottom incisors.

top.can Top canines.

bot.can Bottom canines.

top.pre Top premolars.

bot.pre Bottom premolars.

top.mol Top molars.

bot.mol Bottom molars.

References

John A. Hartigan: Clustering Algorithms. Wiley, New York, 1975.

`dist2`*Compute Pairwise Distances Between Two Data sets*

Description

This function computes and returns the distance matrix computed by using the specified distance measure to compute the pairwise distances between the rows of two data matrices.

Usage

```
dist2(x, y, method = "euclidean", p=2)
```

Arguments

| | |
|---------------------|--|
| <code>x</code> | A data matrix. |
| <code>y</code> | A vector or second data matrix. |
| <code>method</code> | the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given. |
| <code>p</code> | The power of the Minkowski distance. |

Details

This is a two-data-set equivalent of the standard function `dist`. It returns a matrix of all pairwise distances between rows in `x` and `y`. The current implementation is efficient only if `y` has not too many rows (the code is vectorized in `x` but not in `y`).

Note

The definition of Canberra distance was wrong for negative data prior to version 1.3-5.

Author(s)

Friedrich Leisch

See Also

`dist`

Examples

```
x <- matrix(rnorm(20), ncol=4)
rownames(x) = paste("X", 1:nrow(x), sep=".")
y <- matrix(rnorm(12), ncol=4)
rownames(y) = paste("Y", 1:nrow(y), sep=".")

dist2(x, y)
dist2(x, y, "man")
```

```
data(milk)
dist2(milk[1:5,], milk[4:6,])
```

distances

Distance and Centroid Computation

Description

Helper functions to create [kccaFamily](#) objects.

Usage

```
distAngle(x, centers)
distCanberra(x, centers)
distCor(x, centers)
distEuclidean(x, centers)
distJaccard(x, centers)
distManhattan(x, centers)
distMax(x, centers)
distMinkowski(x, centers, p=2)

centAngle(x)
centMean(x)
centMedian(x)

centOptim(x, dist)
centOptim01(x, dist)
```

Arguments

| | |
|---------|--------------------------------------|
| x | A data matrix. |
| centers | A matrix of centroids. |
| p | The power of the Minkowski distance. |
| dist | A distance function. |

Author(s)

Friedrich Leisch

flexclustControl-class

Classes "flexclustControl" and "cclustControl"

Description

Hyperparameters for cluster algorithms.

Objects from the Class

Objects can be created by calls of the form `new("flexclustControl", ...)`. In addition, named lists can be coerced to `flexclustControl` objects, names are completed if unique (see examples).

Slots

Objects of class "flexclustControl" have the following slots:

`iter.max`: Maximum number of iterations.

`tolerance`: The algorithm is stopped when the (relative) change of the optimization criterion is smaller than `tolerance`.

`verbose`: If a positive integer, then progress is reported every `verbose` iterations. If 0, no output is generated during model fitting.

`classify`: Character string, one of "auto", "weighted", "hard" or "simann".

`initcent`: Character string, name of function for initial centroids, currently "randomcent" (the default) and "kmeanspp" are available.

`gamma`: Gamma value for weighted hard competitive learning.

`simann`: Parameters for simulated annealing optimization (only used when `classify="simann"`).

`ntry`: Number of trials per iteration for QT clustering.

`min.size`: Clusters smaller than this value are treated as outliers.

Objects of class "cclustControl" inherit from "flexclustControl" and have the following additional slots:

`method`: Learning rate for hard competitive learning, one of "polynomial" or "exponential".

`pol.rate`: Positive number for polynomial learning rate of form $1/iter^{par}$.

`exp.rate`: Vector of length 2 with parameters for exponential learning rate of form $par1 * (par2/par1)^{(iter/iter.max)}$.

`ng.rate`: Vector of length 4 with parameters for neural gas, see details below.

Learning Rate of Neural Gas

The neural gas algorithm uses updates of form

$$c_{new} = c_{old} + e * \exp(-m/l) * (x - c_{old})$$

for every centroid, where m is the order (minus 1) of the centroid with respect to distance to data point x (0=closest, 1=second, ...). The parameters e and l are given by

$$e = par1 * (par2/par1)^{(iter/iter.max)},$$

$$l = par3 * (par4/par3)^{(iter/iter.max)}.$$

See Martinetz et al (1993) for details of the algorithm, and the examples section on how to obtain default values.

Author(s)

Friedrich Leisch

References

Martinetz T., Berkovich S., and Schulten K. (1993). "Neural-Gas Network for Vector Quantization and its Application to Time-Series Prediction." IEEE Transactions on Neural Networks, 4 (4), pp. 558–569.

Arthur D. and Vassilvitskii S. (2007). "k-means++: the advantages of careful seeding". Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms. pp. 1027-1035.

See Also

[kcca](#), [cclust](#)

Examples

```
## have a look at the defaults
new("flexclustControl")

## corce a list
mycont <- list(iter=500, tol=0.001, class="w")
as(mycont, "flexclustControl")

## some additional slots
as(mycont, "cclustControl")

## default values for ng.rate
new("cclustControl")@ng.rate
```

`flxColors`*Flexclust Color Palettes*

Description

Create and access palettes for the plot methods.

Usage

```
flxColors(n=1:8, color=c("full","medium", "light","dark"), grey=FALSE)
flxPalette(n, ...)
```

Arguments

| | |
|--------------------|---|
| <code>n</code> | Index number of color to return (1 to 8) for <code>flxColor</code> , number of colors to return for <code>flxPalette()</code> . |
| <code>color</code> | Type of color, see details. |
| <code>grey</code> | Return grey value corresponding to palette. |
| <code>...</code> | Passed on to <code>flxColors()</code> . |

Details

This function creates color palettes in HCL space for up to 8 colors. All palettes have constant chroma and luminance, only the hue of the colors change within a palette.

Palettes "full" and "dark" have the same luminance, and palettes "medium" and "light" have the same luminance.

Author(s)

Friedrich Leisch

See Also

[hcl](#)

Examples

```
opar <- par(c("mfrow", "mar", "xaxt"))
par(mfrow=c(2, 2), mar=c(0, 0, 2, 0), yaxt="n")

x <- rep(1, 8)

barplot(x, col = flxColors(color="full"), main="full")
barplot(x, col = flxColors(color="dark"), main="dark")
barplot(x, col = flxColors(color="medium"), main="medium")
barplot(x, col = flxColors(color="light"), main="light")

par(opar)
```

histogram-methods *Methods for Function histogram in Package 'flexclust'*

Description

Plot a histogram of the similarity of each observation to each cluster.

Usage

```
## S4 method for signature 'kccasimple,missing'  
histogram(x, data, xlab="", ...)  
## S4 method for signature 'kccasimple,data.frame'  
histogram(x, data, xlab="", ...)  
## S4 method for signature 'kccasimple,matrix'  
histogram(x, data, xlab="Similarity",  
          power=1, ...)
```

Arguments

| | |
|-------|---|
| x | An object of class "kccasimple". |
| data | If not missing, the distance and thus similarity between observations and cluster centers is determined for the new data and used for the plots. By default the values from the training data are used. |
| xlab | Label for the x-axis. |
| power | Numeric indicating how similarities are transformed, for more details see Dolnicar et al. (2018). |
| ... | Additional arguments passed to histogram . |

Author(s)

Friedrich Leisch

References

Dolnicar S., Gruen B., and Leisch F. (2018) Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful. Springer Singapore.

Description

Image plot of cluster segments overlaid by neighbourhood graph.

Usage

```
## S4 method for signature 'kcca'  
image(x, which = 1:2, npoints = 100,  
      xlab = "", ylab = "", fastcol = TRUE, col=NULL,  
      clwd=0, graph=TRUE, ...)
```

Arguments

| | |
|-----------------|--|
| x | An object of class "kcca". |
| which | Index number of dimensions of input space to plot. |
| npoints | Number of grid points for image. |
| fastcol | If TRUE, a greedy algorithm is used for the background colors of the segments, which may result in neighbouring segments having the same color. If FALSE, neighbouring segments always have different colors at a speed penalty. |
| col | Vector of background colors for the segments. |
| clwd | Line width of contour lines at cluster boundaries, use larger values for npoints than the default to get smooth lines. (Warning: We really need a smarter way to draw cluster boundaries!) |
| graph | Logical, add a neighborhood graph to the plot? |
| xlab, ylab, ... | Graphical parameters. |

Details

This works only for "kcca" objects, no method is available for "kccasimple" objects.

Author(s)

Friedrich Leisch

See Also

[kcca](#)

info

Get Information on Fitted Flexclust Objects

Description

Returns descriptive information about fitted flexclust objects like cluster sizes or sum of within-cluster distances.

Usage

```
## S4 method for signature 'flexclust,character'  
info(object, which, drop=TRUE, ...)
```

Arguments

| | |
|--------|---|
| object | Fitted object. |
| which | Which information to get. Use which="help" to list available information. |
| drop | Logical. If TRUE the result is coerced to the lowest possible dimension. |
| ... | Passed to methods. |

Details

Function `info` can be used to access slots of fitted flexclust objects in a portable way, and in addition computes some meta-information like sum of within-cluster distances.

Function `infoCheck` returns a logical value that is TRUE if the requested information can be computed from the object.

Author(s)

Friedrich Leisch

See Also

[info](#)

Examples

```
data("Nclus")  
plot(Nclus)  
  
c11 <- cclust(Nclus, k=4)  
summary(c11)  
  
## these two are the same  
info(c11)  
info(c11, "help")
```

```

## cluster sizes
i1 <- info(cl1, "size")
i1

## average within cluster distances
i2 <- info(cl1, "av_dist")
i2

## the sum of all within-cluster distances
i3 <- info(cl1, "distsum")
i3

## sum(i1*i2) must of course be the same as i3
stopifnot(all.equal(sum(i1*i2), i3))

## This should return TRUE
modeltools::infoCheck(cl1, "size")
## and this FALSE
modeltools::infoCheck(cl1, "Homer Simpson")
## both combined
i4 <- modeltools::infoCheck(cl1, c("size", "Homer Simpson"))
i4

stopifnot(all.equal(i4, c(TRUE, FALSE)))

```

kcca

K-Centroids Cluster Analysis

Description

Perform k-centroids clustering on a data matrix.

Usage

```

kcca(x, k, family=kccaFamily("kmeans"), weights=NULL,
     group=NULL, control=NULL, simple=FALSE, save.data=FALSE)
kccaFamily(which=NULL, dist=NULL, cent=NULL, name=which, preproc = NULL,
           genDist=NULL, trim=0, groupFun = "minSumClusters")

## S4 method for signature 'kccasimple'
summary(object)

```

Arguments

x A numeric matrix of data, or an object that can be coerced to such a matrix using [data.matrix](#).

| | |
|-----------|---|
| k | Either the number of clusters, or a vector of cluster assignments, or a matrix of initial (distinct) cluster centroids. If a number, a random set of (distinct) rows in x is chosen as the initial centroids. |
| family | Object of class "kccaFamily". |
| weights | An optional vector of weights to be used in the clustering process, cannot be combined with all families. |
| group | An optional grouping vector for the data, see details below. |
| control | An object of class "flexclustControl". |
| simple | Return an object of class "kccasimple"? |
| save.data | Save a copy of x in the return object? |
| which | One of "kmeans", "kmedians", "angle", "jaccard", or "ejaccard". |
| name | Optional long name for family, used only for show methods. |
| dist | A function for distance computation, ignored if which is specified. |
| cent | A function for centroid computation, ignored if which is specified. |
| preproc | Function for data preprocessing. Defaults to <code>function(x) x</code> . |
| genDist | Function for updating the family object based on x . |
| trim | A number in between 0 and 0.5, if non-zero then trimmed means are used for the kmeans family, ignored by all other families. |
| groupFun | Function or name of function to obtain clusters for grouped data, see details below. |
| object | Object of class "kcca". |

Details

See the paper *A Toolbox for K-Centroids Cluster Analysis* referenced below for details.

Value

Function `kcca` returns objects of class "kcca" or "kccasimple" depending on the value of argument `simple`. The simpler objects contain fewer slots and hence are faster to compute, but contain no auxiliary information used by the plotting methods. Most plot methods for "kccasimple" objects do nothing and return a warning. If only centroids, cluster membership or prediction for new data are of interest, then the simple objects are sufficient.

Predefined Families

Function `kccaFamily()` currently has the following predefined families (distance / centroid):

kmeans: Euclidean distance / mean

kmedians: Manhattan distance / median

angle: angle between observation and centroid / standardized mean

jaccard: Jaccard distance / numeric optimization

ejaccard: Jaccard distance / mean

See Leisch (2006) for details on all combinations.

Group Constraints

If group is not NULL, then observations from the same group are restricted to belong to the same cluster (must-link constraint) or different clusters (cannot-link constraint) during the fitting process. If groupFun = "minSumClusters", then all group members are assigned to the cluster where the center has minimal average distance to the group members. If groupFun = "majorityClusters", then all group members are assigned to the cluster the majority would belong to without a constraint.

groupFun = "differentClusters" implements a cannot-link constraint, i.e., members of one group are not allowed to belong to the same cluster. The optimal allocation for each group is found by solving a linear sum assignment problem using `solve_LSAP`. Obviously the group sizes must be smaller than the number of clusters in this case.

Ties are broken at random in all cases. Note that at the moment not all methods for fitted "kcca" objects respect the grouping information, most importantly the plot method when a data argument is specified.

Author(s)

Friedrich Leisch

References

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

Friedrich Leisch and Bettina Gruen. Extending standard cluster algorithms to allow for group constraints. In Alfredo Rizzi and Maurizio Vichi, editors, *Compstat 2006-Proceedings in Computational Statistics*, pages 885-892. Physica Verlag, Heidelberg, Germany, 2006.

See Also

[stepFlexclust](#), [cclust](#), [distances](#)

Examples

```
data("Nclus")
plot(Nclus)

## try kmeans
c11 <- kcca(Nclus, k=4)
c11

image(c11)
points(Nclus)

## A barplot of the centroids
barplot(c11)

## now use k-medians and kmeans++ initialization, cluster centroids
## should be similar...
```

```

c12 <- kcca(Nclus, k=4, family=kccaFamily("kmedians"),
           control=list(initcent="kmeanspp"))
c12

## ... but the boundaries of the partitions have a different shape
image(c12)
points(Nclus)

```

| | |
|---------|---|
| kcca2df | <i>Convert Cluster Result to Data Frame</i> |
|---------|---|

Description

Convert object of class "kcca" to a data frame in long format.

Usage

```
kcca2df(object, data)
```

Arguments

| | |
|--------|---------------------------------------|
| object | Object of class "kcca". |
| data | Optional data if not saved in object. |

Value

A data.frame with columns value, variable and group.

Examples

```

c.iris <- cclust(iris[,-5], 3, save.data=TRUE)
df.c.iris <- kcca2df(c.iris)
summary(df.c.iris)
densityplot(~value|variable+group, data=df.c.iris)

```

| | |
|------|------------------------|
| milk | <i>Milk of Mammals</i> |
|------|------------------------|

Description

The data set contains the ingredients of mammal's milk of 25 animals.

Usage

```
data(milk)
```


Format

A data frame with 25 observations on the following 5 variables (all in percent).

water Water.

protein Protein.

fat Fat.

lactose Lactose.

ash Ash.

References

John A. Hartigan: Clustering Algorithms. Wiley, New York, 1975.

Nclus

Artificial Example with 4 Gaussians

Description

A simple artificial regression example with 4 clusters, all of them having a Gaussian distribution.

Usage

```
data(Nclus)
```

Details

The Nclus data set can be re-created by loading package **flexmix** and running `ExNclus(100)` using `set.seed(2602)`. It has been saved as a data set for simplicity of examples only.

Examples

```
data(Nclus)
cl <- cclust(Nclus, k=4, simple=FALSE, save.data=TRUE)
plot(cl)
```

| | |
|----------|---|
| nutrient | <i>Nutrients in Meat, Fish and Fowl</i> |
|----------|---|

Description

The data set contains the measurements of nutrients in several types of meat, fish and fowl.

Usage

```
data(nutrient)
```

Format

A data frame with 27 observations on the following 5 variables.

energy Food energy (calories).

protein Protein (grams).

fat Fat (grams).

calcium calcium (milli grams).

iron Iron (milli grams).

References

John A. Hartigan: Clustering Algorithms. Wiley, New York, 1975.

| | |
|---------------|--|
| pairs-methods | <i>Methods for Function pairs in Package 'flexclust'</i> |
|---------------|--|

Description

Plot a matrix of neighbourhood graphs.

Usage

```
## S4 method for signature 'kcca'
pairs(x, which=NULL, project=NULL, oma=NULL, ...)
```

Arguments

| | |
|---------|--|
| x | An object of class "kcca" |
| which | Index numbers of dimensions of (projected) input space to plot, default is to plot all dimensions. |
| project | Projection object for which a predict method exists, e.g., the result of prcomp . |
| oma | Outer margin. |
| ... | Passed to the plot method. |

Details

This works only for "kcca" objects, no method is available for "kccasimple" objects.

Author(s)

Friedrich Leisch

parameters *Get Centroids from KCCA Object*

Description

Returns the matrix of centroids of a fitted object of class "kcca".

Usage

```
## S4 method for signature 'kccasimple'
parameters(object, ...)
```

Arguments

| | |
|--------|---------------------|
| object | Fitted object. |
| ... | Currently not used. |

Author(s)

Friedrich Leisch

plot-methods *Methods for Function plot in Package 'flexclust'*

Description

Plot the neighbourhood graph of a cluster solution together with projected data points.

Usage

```
## S4 method for signature 'kcca,missing'
plot(x, y, which=1:2, project=NULL,
     data=NULL, points=TRUE, hull=TRUE, hull.args=NULL,
     number = TRUE, simlines=TRUE,
     lwd=1, maxlwd=8*lwd, cex=1.5, numcol=FALSE, nodes=16,
     add=FALSE, xlab="", ylab="", xlim = NULL,
     ylim = NULL, pch=NULL, col=NULL, ...)
```

Arguments

| | |
|---------------|--|
| x | An object of class "kcca" |
| y | Not used |
| which | Index numbers of dimensions of (projected) input space to plot. |
| project | Projection object for which a predict method exists, e.g., the result of <code>prcomp</code> . |
| data | Data to include in plot. If the cluster object x was created with <code>save.data=TRUE</code> , then these are used by default. |
| points | Logical, shall data points be plotted (if available)? |
| hull | If TRUE, then hulls of the data are plotted (if available). Can either be a logical value, one of the strings "convex" (the default) or "ellipse", or a function for plotting the hulls. |
| hull.args | A list of arguments for the hull function. |
| number | Logical, plot number labels in nodes of graph? |
| numcol, cex | Color and size of number labels in nodes of graph. If numcol is logical, it switches between black and the color of the clusters, else it is taken as a vector of colors. |
| nodes | Plotting symbol to use for nodes if no numbers are drawn. |
| simlines | Logical, plot edges of graph? |
| lwd, maxlwd | Numerical, thickness of lines. |
| add | Logical, add to existing plot? |
| xlab, ylab | Axis labels. |
| xlim, ylim | Axis range. |
| pch, col, ... | Plotting symbols and colors for data points. |

Details

This works only for "kcca" objects, no method is available for "kccasimple" objects.

Author(s)

Friedrich Leisch

References

Friedrich Leisch. Visualizing cluster analysis and finite mixture models. In Chun houh Chen, Wolfgang Haerdle, and Antony Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics. Springer Verlag, 2008.

| | |
|-----------------|-----------------------------------|
| predict-methods | <i>Predict Cluster Membership</i> |
|-----------------|-----------------------------------|

Description

Return either the cluster membership of training data or predict for new data.

Usage

```
## S4 method for signature 'kccasimple'
predict(object, newdata, ...)
## S4 method for signature 'flexclust,ANY'
clusters(object, newdata, ...)
```

Arguments

| | |
|---------|---|
| object | Object of class inheriting from "flexclust". |
| newdata | An optional data matrix with the same number of columns as the cluster centers. If omitted, the fitted values are used. |
| ... | Currently not used. |

Details

clusters can be used on any object of class "flexclust" and returns the cluster memberships of the training data.

predict can be used only on objects of class "kcca" (which inherit from "flexclust"). If no newdata argument is specified, the function is identical to clusters, if newdata is specified, then cluster memberships for the new data are predicted. clusters(object, newdata, ...) is an alias for predict(object, newdata, ...).

Author(s)

Friedrich Leisch

| | |
|--------------|--|
| priceFeature | <i>Artificial 2d Market Segment Data</i> |
|--------------|--|

Description

Simple artificial 2-dimensional data to demonstrate clustering for market segmentation. One dimension is the hypothetical feature sophistication (or performance or quality, etc) of a product, the second dimension the price customers are willing to pay for the product.

Usage

```
priceFeature(n, which=c("2clust", "3clust", "3clustold", "5clust",
                        "ellipse", "triangle", "circle", "square",
                        "largesmall"))
```

Arguments

| | |
|-------|--------------------|
| n | Sample size. |
| which | Shape of data set. |

References

Sara Dolnicar and Friedrich Leisch. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, 21:83-101, 2010.

Examples

```
plot(priceFeature(200, "2clust"))
plot(priceFeature(200, "3clust"))
plot(priceFeature(200, "3clustold"))
plot(priceFeature(200, "5clust"))
plot(priceFeature(200, "ell"))
plot(priceFeature(200, "tri"))
plot(priceFeature(200, "circ"))
plot(priceFeature(200, "square"))
plot(priceFeature(200, "largesmall"))
```

 projAxes

Add Arrows for Projected Axes to a Plot

Description

Adds arrows for original coordinate axes to a projection plot.

Usage

```
projAxes(object, which=1:2, center=NULL,
         col="red", radius=NULL,
         minradius=0.1, textargs=list(col=col),
         col.names=getColnames(object),
         which.names="", group = NULL, groupFun = colMeans,
         plot=TRUE, ...)

placeLabels(object)
## S4 method for signature 'projAxes'
placeLabels(object)
```

Arguments

| | |
|-------------|--|
| object | Return value of a projection method like prcomp . |
| which | Index number of dimensions of (projected) input space that have been plotted. |
| center | Center of the coordinate system to use in projected space. Default is the center of the plotting region. |
| col | Color of arrows. |
| radius | Relative size of the arrows. |
| minradius | Minimum radius of arrows to include (relative to arrow size). |
| textargs | List of arguments for text . |
| col.names | Variable names of the original data. |
| which.names | A regular expression which variable names to include in the plot. |
| group | An optional grouping variable for the original coordinates. Coordinates with group NA are omitted. |
| groupFun | Function used to aggregate the projected coordinates if group is specified. |
| plot | Logical, if TRUE the axes are added to the current plot. |
| ... | Passed to arrows . |

Value

projAxes invisibly returns an object of class "projAxes", which can be added to an existing plot by its plot method.

Author(s)

Friedrich Leisch

Examples

```
data(milk)
milk.pca <- prcomp(milk, scale=TRUE)

## create a biplot step by step
plot(predict(milk.pca), type="n")
text(predict(milk.pca), rownames(milk), col="green", cex=0.8)
projAxes(milk.pca)

## the same, but arrows are blue, centered at origin and all arrows are
## plotted
plot(predict(milk.pca), type="n")
text(predict(milk.pca), rownames(milk), col="green", cex=0.8)
projAxes(milk.pca, col="blue", center=0, minradius=0)

## use points instead of text, plot PC2 and PC3, manual radius
## specification, store result
plot(predict(milk.pca)[,c(2,3)])
arr <- projAxes(milk.pca, which=c(2,3), radius=1.2, plot=FALSE)
```

```

plot(arr)

## Not run:

## manually try to find new places for the labels: each arrow is marked
## active in turn, use the left mouse button to find a better location
## for the label. Use the right mouse button to go on to the next
## variable.

arr1 <- placeLabels(arr)

## now do the plot again:
plot(predict(milk.pca)[,c(2,3)])
plot(arr1)

## End(Not run)

```

propBarchart

Barcharts and Boxplots for Columns of a Data Matrix Split by Groups

Description

Split a binary or numeric matrix by a grouping variable, run a series of tests on all variables, adjust for multiple testing and graphically represent results.

Usage

```

propBarchart(x, g, alpha=0.05, correct="holm", test="prop.test",
             sort=FALSE, strip.prefix="", strip.labels=NULL,
             which=NULL, byvar=FALSE, ...)

## S4 method for signature 'propBarchart'
summary(object, ...)

groupBWplot(x, g, alpha=0.05, correct="holm", xlab="", col=NULL,
            shade=!is.null(shadefun), shadefun=NULL,
            strip.prefix="", strip.labels=NULL, which=NULL, byvar=FALSE,
            ...)

```

Arguments

| | |
|---------|--|
| x | A binary data matrix. |
| g | A factor specifying the groups. |
| alpha | Significance level for test of differences in proportions. |
| correct | Correction method for multiple testing, passed to p.adjust . |
| test | Test to use for detecting significant differences in proportions. |

| | |
|--------------|--|
| sort | Logical, sort variables by total sample mean? |
| strip.prefix | Character string prepended to strips of the barchart (the remainder of the strip are group levels and group sizes). Ignored if <code>strip.labels</code> is specified. |
| strip.labels | Character vector of labels to use for strips of barchart . |
| which | Index numbers or names of variables to plot. |
| byvar | If TRUE, a panel is plotted for each variable. By default a panel is plotted for each group. |
| ... | Passed on to barchart or bwplot . |
| object | Return value of <code>propBarchart</code> . |
| xlab | A title for the x-axis: see title . The default is "". |
| col | Vector of colors for the panels. |
| shade | If TRUE, only variables with significant differences in median are filled with color. |
| shadefun | A function or name of a function to compute which boxes are shaded, e.g. "kruskalTest" (default), "medianInside" or "boxOverlap". |

Details

Function `propBarchart` splits a binary data matrix into subgroups, computes the percentage of ones in each column and compares the proportions in the groups using [prop.test](#). The p-values for all variables are adjusted for multiple testing and a barchart of group percentages is drawn highlighting variables with significant differences in proportion. The `summary` method can be used to create a corresponding table for publications.

Function `groupBWplot` takes a general numeric matrix, also splits into subgroups and uses boxes instead of bars. By default [kruskal.test](#) is used to compute significant differences in location, in addition the heuristics from [bwplot](#), [kcca-method](#) can be used. Boxes of the complete sample are used as reference in the background.

Author(s)

Friedrich Leisch

See Also

[barplot-methods](#), [bwplot](#), [kcca-method](#)

Examples

```
## create a binary matrix from the iris data plus a random noise column
x <- apply(iris[,-5], 2, function(z) z>median(z))
x <- cbind(x, Noise=sample(0:1, 150, replace=TRUE))

## There are significant differences in all 4 original variables, Noise
## has most likely no significant difference (of course the difference
## will be significant in alpha percent of all random samples).
p <- propBarchart(x, iris$Species)
p
```

```
summary(p)
propBarchart(x, iris$Species, byvar=TRUE)

x <- iris[,-5]
x <- cbind(x, Noise=rnorm(150, mean=3))
groupBWplot(x, iris$Species)
groupBWplot(x, iris$Species, shade=TRUE)
groupBWplot(x, iris$Species, shadefun="medianInside")
groupBWplot(x, iris$Species, shade=TRUE, byvar=TRUE)
```

qtclust

Stochastic QT Clustering

Description

Perform stochastic QT clustering on a data matrix.

Usage

```
qtclust(x, radius, family = kccaFamily("kmeans"), control = NULL,
        save.data=FALSE, kcca=FALSE)
```

Arguments

| | |
|-----------|---|
| x | A numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns). |
| radius | Maximum radius of clusters. |
| family | Object of class "kccaFamily" specifying the distance measure to be used. |
| control | An object of class "flexclustControl" specifying the minimum number of observations per cluster (<code>min.size</code>), and trials per iteration (<code>ntry</code> , see details below). |
| . | . |
| save.data | Save a copy of x in the return object? |
| kcca | Run <i>kcca</i> after the QT cluster algorithm has converged? |

Details

This function implements a variation of the QT clustering algorithm by Heyer et al. (1999), see Scharl and Leisch (2006). The main difference is that in each iteration not all possible cluster start points are considered, but only a random sample of size `control@entry`. We also consider only points as initial centers where at least one other point is within a circle with radius `radius`. In most cases the resulting solutions are almost the same at a considerable speed increase, in some cases even better solutions are obtained than with the original algorithm. If `control@entry` is set to the size of the data set, an algorithm similar to the original algorithm as proposed by Heyer et al. (1999) is obtained.

Value

Function qtclust by default returns objects of class "kccasimple". If argument kcca is TRUE, function kcca() is run afterwards (initialized on the QT cluster solution). Data points not clustered by the QT cluster algorithm are omitted from the kcca() iterations, but filled back into the return object. All plot methods defined for objects of class "kcca" can be used.

Author(s)

Friedrich Leisch

References

Heyer, L. J., Kruglyak, S., Yooseph, S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research* 9, 1106–1115.

Theresa Scharl and Friedrich Leisch. The stochastic QT-clust algorithm: evaluation of stability and variance on time-course microarray data. In Alfredo Rizzi and Maurizio Vichi, editors, *Compstat 2006 – Proceedings in Computational Statistics*, pages 1015-1022. Physica Verlag, Heidelberg, Germany, 2006.

Examples

```
x <- matrix(10*runif(1000), ncol=2)

## maximum distance of point to cluster center is 3
c11 <- qtclust(x, radius=3)

## maximum distance of point to cluster center is 1
## -> more clusters, longer runtime
c12 <- qtclust(x, radius=1)

opar <- par(c("mfrow", "mar"))
par(mfrow=c(2,1), mar=c(2.1,2.1,1,1))
plot(x, col=predict(c11), xlab="", ylab="")
plot(x, col=predict(c12), xlab="", ylab="")
par(opar)
```

 randIndex

Compare Partitions

Description

Compute the (adjusted) Rand, Jaccard and Fowlkes-Mallows index for agreement of two partitions.

Usage

```

comPart(x, y, type=c("ARI","RI","J","FM"))
## S4 method for signature 'flexclust,flexclust'
comPart(x, y, type)
## S4 method for signature 'numeric,numeric'
comPart(x, y, type)
## S4 method for signature 'flexclust,numeric'
comPart(x, y, type)
## S4 method for signature 'numeric,flexclust'
comPart(x, y, type)

randIndex(x, y, correct=TRUE, original=!correct)
## S4 method for signature 'table,missing'
randIndex(x, y, correct=TRUE, original=!correct)
## S4 method for signature 'ANY,ANY'
randIndex(x, y, correct=TRUE, original=!correct)

```

Arguments

| | |
|-------------------|--|
| x | Either a 2-dimensional cross-tabulation of cluster assignments (for randIndex only), an object inheriting from class "flexclust", or an integer vector of cluster memberships. |
| y | An object inheriting from class "flexclust", or an integer vector of cluster memberships. |
| type | character vector of abbreviations of indices to compute. |
| correct, original | Logical, correct the Rand index for agreement by chance? |

Value

A vector of indices.

Rand Index

Let A denote the number of all pairs of data points which are either put into the same cluster by both partitions or put into different clusters by both partitions. Conversely, let D denote the number of all pairs of data points that are put into one cluster in one partition, but into different clusters by the other partition. The partitions disagree for all pairs D and agree for all pairs A . We can measure the agreement by the Rand index $A/(A + D)$ which is invariant with respect to permutations of cluster labels.

The index has to be corrected for agreement by chance if the sizes of the clusters are not uniform (which is usually the case), or if there are many clusters, see Hubert & Arabie (1985) for details.

Jaccard Index

If the number of clusters is very large, then usually the vast majority of pairs of points will not be in the same cluster. The Jaccard index tries to account for this by using only pairs of points that are in the same cluster in the definition of A .

Fowlkes-Mallows

Let A again be the pairs of points that are in the same cluster in both partitions. Fowlkes-Mallows divides this number by the geometric mean of the sums of the number of pairs in each cluster of the two partitions. This gives the probability that a pair of points which are in the same cluster in one partition are also in the same cluster in the other partition.

Author(s)

Friedrich Leisch

References

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2, 193–218, 1985.

Marina Meila. Comparing clusterings - an axiomatic view. In Stefan Wrobel and Luc De Raedt, editors, *Proceedings of the International Machine Learning Conference (ICML)*. ACM Press, 2005.

Examples

```
## no class correlations: corrected Rand almost zero
g1 <- sample(1:5, size=1000, replace=TRUE)
g2 <- sample(1:5, size=1000, replace=TRUE)
tab <- table(g1, g2)
randIndex(tab)

## uncorrected version will be large, because there are many points
## which are assigned to different clusters in both cases
randIndex(tab, correct=FALSE)
comPart(g1, g2)

## let pairs (g1=1,g2=1) and (g1=3,g2=3) agree better
k <- sample(1:1000, size=200)
g1[k] <- 1
g2[k] <- 1
k <- sample(1:1000, size=200)
g1[k] <- 3
g2[k] <- 3
tab <- table(g1, g2)

## the index should be larger than before
randIndex(tab, correct=TRUE, original=TRUE)
comPart(g1, g2)
```

randomTour

Plot a Random Tour

Description

Create a series of projection plots corresponding to a random tour through the data.

Usage

```

randomTour(object, ...)

## S4 method for signature 'ANY'
randomTour(object, ...)
## S4 method for signature 'matrix'
randomTour(object, ...)
## S4 method for signature 'flexclust'
randomTour(object, data=NULL, col=NULL, ...)

randomTourMatrix(x, directions=10,
                 steps=100, sec=4, sleep = sec/steps,
                 axiscol=2, axislab=colnames(x),
                 center=NULL, radius=1, minradius=0.01, asp=1,
                 ...)

```

Arguments

| | |
|------------|--|
| object, x | A matrix or an object of class "flexclust". |
| data | Data to include in plot. |
| col | Plotting colors for data points. |
| directions | Integer value, how many different directions are toured. |
| steps | Integer, number of steps in each direction. |
| sec | Numerical, lower bound for the number of seconds each direction takes. |
| sleep | Numerical, sleep for as many seconds after each picture has been plotted. |
| axiscol | If not NULL, then arrows are plotted for projections of the original coordinate axes in these colors. |
| axislab | Optional labels for the projected axes. |
| center | Center of the coordinate system to use in projected space. Default is the center of the plotting region. |
| radius | Relative size of the arrows. |
| minradius | Minimum radius of arrows to include. |
| asp, ... | Passed on to randomTourMatrix and from there to plot. |

Details

Two random locations are chosen, and data then projected onto hyperplanes which are orthogonal to step vectors interpolating the two locations. The first two coordinates of the projected data are plotted. If `directions` is larger than one, then after the first steps plots one more random location is chosen, and the procedure is repeated from the current position to the new location, etc..

The whole procedure is similar to a grand tour, but no attempt is made to optimize subsequent directions, `randomTour` simply chooses a random direction in each iteration. Use `rggobi` for the real thing.

Obviously the function needs a reasonably fast computer and graphics device to give a smooth impression, for `x11` it may be necessary to use `type="Xlib"` rather than `cairo`.

Author(s)

Friedrich Leisch

Examples

```

if(interactive()){
  par(ask=FALSE)
  randomTour(iris[,1:4], axiscol=2:5)
  randomTour(iris[,1:4], col=as.numeric(iris$Species), axiscol=4)

  x <- matrix(runif(300), ncol=3)
  x <- rbind(x, x+1, x+2)
  cl <- cclust(x, k=3, save.data=TRUE)

  randomTour(cl, center=0, axiscol="black")

  ## now use predicted cluster membership for new data as colors
  randomTour(cl, center=0, axiscol="black",
             data=matrix(rnorm(3000, mean=1, sd=2), ncol=3))
}

```

relabel

*Relabel Cluster Results.***Description**

The clusters are relabelled to obtain a unique labeling.

Usage

```

relabel(object, by, ...)
## S4 method for signature 'kccasimple,character'
relabel(object, by, which = NULL, ...)
## S4 method for signature 'kccasimple,integer'
relabel(object, by, ...)
## S4 method for signature 'kccasimple,missing'
relabel(object, by, ...)
## S4 method for signature 'stepFlexclust,integer'
relabel(object, by = "series", ...)
## S4 method for signature 'stepFlexclust,missing'
relabel(object, by, ...)

```

Arguments

| | |
|--------|---|
| object | An object of class "kccasimple" or "stepFlexclust". |
| by | If a character vector, it needs to be one of "mean", "median", "variable", "manual", "centers", "shadow", "symmshadow" or "series". If missing, "mean" or "series" is used depending on if object is of class "kccasimple" or "stepFlexclust". If an integer vector, it needs to indicate the new ordering. |

| | |
|-------|---|
| which | Either an integer vector indicating the ordering or a vector of length one indicating the variable used for ordering. |
| ... | Currently not used. |

Details

If `by` is a character vector with value "mean" or "median", the clusters are ordered by the mean or median values over all variables for each cluster. If `by = "manual"` which needs to be a vector indicating the ordering. If `by = "variable"` which needs to be indicate the variable which is used to determine the ordering. If `by` is "centers", "shadow" or "symmshadow", cluster similarities are calculated using `clusterSim` and used to determine an ordering using `seriate` from package **seriation**.

If `by = "series"` the relabeling is performed over a series of clustering to minimize the misclassification.

Author(s)

Friedrich Leisch

See Also

[clusterSim](#), [seriate](#)

shadow

Cluster Shadows and Silhouettes

Description

Compute and plot shadows and silhouettes.

Usage

```
## S4 method for signature 'kccasimple'
shadow(object, ...)
## S4 method for signature 'kcca'
Silhouette(object, data=NULL, ...)
```

Arguments

| | |
|--------|---|
| object | An object of class "kcca" or "kccasimple". |
| data | Data to compute silhouette values for. If the cluster object was created with <code>save.data=TRUE</code> , then these are used by default. |
| ... | Currently not used. |

Details

The shadow value of each data point is defined as twice the distance to the closest centroid divided by the sum of distances to closest and second-closest centroid. If the shadow values of a point is close to 0, then the point is close to its cluster centroid. If the shadow value is close to 1, it is almost equidistant to the two centroids. Thus, a cluster that is well separated from all other clusters should have many points with small shadow values.

The silhouette value of a data point is defined as the scaled difference between the average dissimilarity of a point to all points in its own cluster to the smallest average dissimilarity to the points of a different cluster. Large silhouette values indicate good separation.

The main difference between silhouette values and shadow values is that we replace average dissimilarities to points in a cluster by dissimilarities to point averages (=centroids). See Leisch (2009) for details.

Author(s)

Friedrich Leisch

References

Friedrich Leisch. Neighborhood graphs, stripes and shadow plots for cluster visualization. *Statistics and Computing*, 2009. Accepted for publication on 2009-06-16.

See Also

[silhouette](#)

Examples

```
data(Nclus)
set.seed(1)
c5 <- cclust(Nclus, 5, save.data=TRUE)
c5
plot(c5)

## high shadow values indicate clusters with *bad* separation
shadow(c5)
plot(shadow(c5))

## high Silhouette values indicate clusters with *good* separation
Silhouette(c5)
plot(Silhouette(c5))
```

shadowStars

*Shadow Stars***Description**

Shadow star plots and corresponding panel functions.

Usage

```
shadowStars(object, which=1:2, project=NULL,
            width=1, varwidth=FALSE,
            panel=panelShadowStripes,
            box=NULL, col=NULL, add=FALSE, ...)
```

```
panelShadowStripes(x, col, ...)
panelShadowViolin(x, ...)
panelShadowBP(x, ...)
panelShadowSkeleton(x, ...)
```

Arguments

| | |
|----------|--|
| object | An object of class "kcca". |
| which | Index numbers of dimensions of (projected) input space to plot. |
| project | Projection object for which a predict method exists, e.g., the result of prcomp . |
| width | Width of vertices connecting the cluster centroids. |
| varwidth | Logical, shall all vertices have the same width or should the width be proportional to number of points shown on the vertex? |
| panel | Function used to draw vertices. |
| box | Color of rectangle drawn around each vertex. |
| col | A vector of colors for the clusters. |
| add | Logical, start a new plot? |
| ... | Passed on to panel function. |
| x | Shadow values of data points corresponding to the vertex. |

Details

The shadow value of each data point is defined as twice the distance to the closest centroid divided by the sum of distances to closest and second-closest centroid. If the shadow values of a point is close to 0, then the point is close to its cluster centroid. If the shadow value is close to 1, it is almost equidistant to the two centroids. Thus, a cluster that is well separated from all other clusters should have many points with small shadow values.

The neighborhood graph of a cluster solution connects two centroids by a vertex if at least one data point has the two centroids as closest and second closest. The width of the vertex is proportional to

the sum of shadow values of all points having these two as closest and second closest. A shadow star depicts the distribution of shadow values on the vertex, see Leisch (2009) for details.

Currently four panel functions are available:

`panelShadowStripes`: line segment for each shadow value.

`panelShadowViolin`: violin plot of shadow values.

`panelShadowBP`: box-percentile plot of shadow values.

`panelShadowSkeleton`: average shadow value.

Author(s)

Friedrich Leisch

References

Friedrich Leisch. Neighborhood graphs, stripes and shadow plots for cluster visualization. *Statistics and Computing*, 2009. Accepted for publication on 2009-06-16.

See Also

[shadow](#)

Examples

```
data(Nclus)
set.seed(1)
c5 <- cclust(Nclus, 5, save.data=TRUE)
c5
plot(c5)

shadowStars(c5)
shadowStars(c5, varwidth=TRUE)

shadowStars(c5, panel=panelShadowViolin)
shadowStars(c5, panel=panelShadowBP)

## always use varwidth=TRUE with panelShadowSkeleton, otherwise a few
## large shadow values can lead to misleading results:
shadowStars(c5, panel=panelShadowSkeleton)
shadowStars(c5, panel=panelShadowSkeleton, varwidth=TRUE)
```

`slsaplot`*Segment Level Stability Across Solutions Plot.*

Description

Create a segment level stability across solutions plot, possibly using an additional variable for coloring the nodes.

Usage

```
slsaplot(object, nodecol = NULL, ...)
```

Arguments

| | |
|----------------------|--|
| <code>object</code> | An object returned by <code>stepFlexclust</code> . |
| <code>nodecol</code> | A numeric vector of length equal to the number of observations clustered in <code>object</code> which represents an additional variable where a cluster-specific mean is calculated and used to color the nodes. |
| <code>...</code> | Additional graphical parameters to modify the plot. |

Details

For more details see Dolnicar and Leisch (2017) and Dolnicar et al. (2018).

Value

List of length equal to the number of different cluster solutions minus one containing numeric vectors of the entropy values used by default to color the nodes.

Author(s)

Friedrich Leisch

References

Dolnicar S. and Leisch F. (2017) "Using Segment Level Stability to Select Target Segments in Data-Driven Market Segmentation Studies" *Marketing Letters*, 28 (3), pp. 423–436.

Dolnicar S., Gruen B., and Leisch F. (2018) *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Springer Singapore.

See Also

[stepFlexclust](#), [relabel](#), [slswFlexclust](#)

Examples

```

data("Nclus")
cl25 <- stepFlexclust(Nclus, k=2:5)
slsaplot(cl25)
cl25 <- relabel(cl25)
slsaplot(cl25)

```

slswFlexclust

*Segment Level Stability Within Solution.***Description**

Assess segment level stability within solution.

Usage

```

slswFlexclust(x, object, ...)
## S4 method for signature 'resampleFlexclust,missing'
plot(x, y, ...)
## S4 method for signature 'resampleFlexclust'
boxplot(x, which=1, ylab=NULL, ...)
## S4 method for signature 'resampleFlexclust'
densityplot(x, data, which=1, ...)
## S4 method for signature 'resampleFlexclust'
summary(object)

```

Arguments

| | |
|--------|---|
| x | A numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns) passed to stepFlexclust . |
| object | Object of class "kcca" for slswFlexclust and "resampleFlexclust" for the summary method. |
| y | Missing. |
| which | Integer or character indicating which validation measure is used for plotting. |
| ylab | Axis label. |
| data | Not used. |
| ... | Additional arguments; for details see below. |

Details

Additional arguments in slswFlexclust are argument nsamp which is by default equal to 100 and allows to change the number of bootstrap pairs drawn. Argument seed allows to set a random seed and argument multicore is by default TRUE and indicates if bootstrap samples should be drawn in parallel. Argument verbose is by default equal to FALSE and if TRUE progress information is shown during computations.

There are plotting as well as printing and summary methods implemented for objects of class "resampleFlexclust". In addition to a standard plot method also methods for densityplot and boxplot are provided.

For more details see Dolnicar and Leisch (2017) and Dolnicar et al. (2018).

Value

An object of class "resampleFlexclust".

Author(s)

Friedrich Leisch

References

Dolnicar S. and Leisch F. (2017) "Using Segment Level Stability to Select Target Segments in Data-Driven Market Segmentation Studies" Marketing Letters, 28 (3), pp. 423–436.

Dolnicar S., Gruen B., and Leisch F. (2018) Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful. Springer Singapore.

See Also

[slsaplot](#)

Examples

```
data("Nclus")
c13 <- kcca(Nclus, k = 3)
slsw.c13 <- slswFlexclust(Nclus, c13, nsamp = 20)
plot(Nclus, col = clusters(c13))
plot(slsw.c13)
densityplot(slsw.c13)
boxplot(slsw.c13)
```

stepFlexclust

Run Flexclust Algorithms Repeatedly

Description

Runs clustering algorithms repeatedly for different numbers of clusters and returns the minimum within cluster distance solution for each.

Usage

```

stepFlexclust(x, k, nrep=3, verbose=TRUE, FUN = kcca, drop=TRUE,
              group=NULL, simple=FALSE, save.data=FALSE, seed=NULL,
              multicore=TRUE, ...)

stepcclust(...)

## S4 method for signature 'stepFlexclust,missing'
plot(x, y,
     type=c("barplot", "lines"), totaldist=NULL,
     xlab=NULL, ylab=NULL, ...)

## S4 method for signature 'stepFlexclust'
getModel(object, which=1)

```

Arguments

| | |
|------------|---|
| x, ... | Passed to kcca or cclust . |
| k | A vector of integers passed in turn to the k argument of kcca |
| nrep | For each value of k run kcca nrep times and keep only the best solution. |
| FUN | Cluster function to use, typically kcca or cclust . |
| verbose | If TRUE, show progress information during computations. |
| drop | If TRUE and K is of length 1, then a single cluster object is returned instead of a "stepFlexclust" object. |
| group | An optional grouping vector for the data, see kcca for details. |
| simple | Return an object of class "kccasimple"? |
| save.data | Save a copy of x in the return object? |
| seed | If not NULL, a call to <code>set.seed()</code> is made before any clustering is done. |
| multicore | If TRUE, use <code>mclapply()</code> from package parallel for parallel processing. |
| y | Not used. |
| type | Create a barplot or lines plot. |
| totaldist | Include value for 1-cluster solution in plot? Default is TRUE if K contains 2, else FALSE. |
| xlab, ylab | Graphical parameters. |
| object | Object of class "stepFlexclust". |
| which | Number of model to get. If character, interpreted as number of clusters. |

Details

stepcclust is a simple wrapper for `stepFlexclust(..., FUN=cclust)`.

Author(s)

Friedrich Leisch

Examples

```

data("Nclus")
plot(Nclus)

## multicore off for CRAN checks
c11 <- stepFlexclust(Nclus, k=2:7, FUN=cclust, multicore=FALSE)
c11

plot(c11)

# two ways to do the same:
getModel(c11, 4)
c11[[4]]

opar <- par("mfrow")
par(mfrow=c(2, 2))
for(k in 3:6){
  image(getModel(c11, as.character(k)), data=Nclus)
  title(main=paste(k, "clusters"))
}
par(opar)

```

stripes

*Stripes Plot***Description**

Plot distance of data points to cluster centroids using stripes.

Usage

```

stripes(object, groups=NULL, type=c("first", "second", "all"),
        beside=(type!="first"), col=NULL, gp.line=NULL, gp.bar=NULL,
        gp.bar2=NULL, number=TRUE, legend=!is.null(groups),
        ylim=NULL, ylab="distance from centroid",
        margins=c(2,5,3,2), ...)

```

Arguments

| | |
|--------------------------|--|
| object | An object of class "kcca". |
| groups | Grouping variable to color-code the stripes. By default cluster membership is used as groups. |
| type | Plot distance to closest, closest and second-closest or to all centroids? |
| beside | Logical, make different stripes for different clusters? |
| col | Vector of colors for clusters or groups. |
| gp.line, gp.bar, gp.bar2 | Graphical parameters for horizontal lines and background rectangular areas, see gpar . |

| | |
|------------|---|
| number | Logical, write cluster numbers on x-axis? |
| legend | Logical, plot a legend for the groups? |
| ylim, ylab | Graphical parameters for y-axis. |
| margins | Margin of the plot. |
| ... | Further graphical parameters. |

Details

A simple, yet very effective plot for visualizing the distance of each point from its closest and second-closest cluster centroids is a stripes plot. For each of the k clusters we have a rectangular area, which we optionally vertically divide into k smaller rectangles (`beside=TRUE`). Then we draw a horizontal line segment for each data point marking the distance of the data point from the corresponding centroid.

Author(s)

Friedrich Leisch

References

Friedrich Leisch. Neighborhood graphs, stripes and shadow plots for cluster visualization. *Statistics and Computing*, 20(4), 457–469, 2010.

Examples

```

bw05 <- bundestag(2005)
bavaria <- bundestag(2005, state="Bayern")

set.seed(1)
c4 <- cclust(bw05, k=4, save.data=TRUE)
plot(c4)

stripes(c4)
stripes(c4, beside=TRUE)

stripes(c4, type="sec")
stripes(c4, type="sec", beside=FALSE)
stripes(c4, type="all")

stripes(c4, groups=bavaria)

## ugly, but shows how colors of all parts can be changed
library("grid")
stripes(c4, type="all",
        gp.bar=gpar(col="red", lwd=3, fill="white"),
        gp.bar2=gpar(col="green", lwd=3, fill="black"))

```

vacmot

Vacation Motives of Australians

Description

In 2006 a sample of 1000 respondents representative for the adult Australian population was asked about their environmental behaviour when on vacation. In addition the survey also included a list of statements about vacation motives like "I want to rest and relax," "I use my holiday for the health and beauty of my body," and "Cultural offers and sights are a crucial factor.". Answers are binary ("applies", "does not apply").

Usage

```
data(vacmot)
```

Format

Data frame `vacmot` has 1000 observations on 20 binary variables on travel motives. Data frame `vacmotdesc` has 1000 observation on sociodemographic descriptor variables, mean moral obligation to protect the environment score, mean NEP score, and mean environmental behaviour score, see Dolnicar & Leisch (2008) for details. In addition integer vector `vacmot6` contains the 6 cluster partition presented in Dolnicar & Leisch (2008).

Source

The data set was collected by the Institute for Innovation in Business and Social Research, University of Wollongong (NSW, Australia).

References

Sara Dolnicar and Friedrich Leisch. An investigation of tourists' patterns of obligation to protect the environment. *Journal of Travel Research*, 46:381-391, 2008.

Sara Dolnicar and Friedrich Leisch. Using graphical statistics to better understand market segmentation solutions. *International Journal of Market Research*, 56(2):97-120, 2014.

Examples

```
data(vacmot)
summary(vacmotdesc)
dotchart(sort(colMeans(vacmot)))

## reproduce Figure 6 from Dolnicar & Leisch (2008)
c16 <- kcca(vacmot, k=vacmot6, control=list(iter=0))
barchart(c16)
```

volunteers

Motivation of Australian Volunteers

Description

Part of an Australian survey on motivation of volunteers to work for non-profit organisations like Red Cross, State Emergency Service, Rural Fire Service, Surf Life Saving, Rotary, Parents and Citizens Associations, etc..

Usage

```
data(volunteers)
```

Format

A data frame with 1415 observations on the following 21 variables: age and gender of respondents plus 19 binary motivation items (1 applies/ 0 does not apply).

GENDER Gender of respondent.

AGEG Age group, a factor with categorized age of respondents.

meet.people I can meet different types of people.

no.one.else There is no-one else to do the work.

example It sets a good example for others.

socialise I can socialise with people who are like me.

help.others It gives me the chance to help others.

give.back I can give something back to society.

career It will help my career prospects.

lonely It makes me feel less lonely.

active It keeps me active.

community It will improve my community.

cause I can support an important cause.

faith I can put faith into action.

services I want to maintain services that I may use one day.

children My children are involved with the organisation.

good.job I feel like I am doing a good job.

benefited I know someone who has benefited from the organisation.

network I can build a network of contacts.

recognition I can gain recognition within the community.

mind.off It takes my mind off other things.

Source

The volunteering data was collected by the Institute for Innovation in Business and Social Research, University of Wollongong (NSW, Australia), using funding from Bushcare Wollongong and the Australian Research Council under the ARC Linkage Grant scheme (LP0453682).

References

Melanie Randle and Sara Dolnicar. Not Just Any Volunteers: Segmenting the Market to Attract the High-Contributors. *Journal of Non-profit and Public Sector Marketing*, 21(3), 271-282, 2009.

Melanie Randle and Sara Dolnicar. Self-congruity and volunteering: A multi-organisation comparison. *European Journal of Marketing*, 45(5), 739-758, 2011.

Melanie Randle, Friedrich Leisch, and Sara Dolnicar. Competition or collaboration? The effect of non-profit brand image on volunteer recruitment strategy. *Journal of Brand Management*, 20(8):689-704, 2013.

Index

- * **classes**
 - flexclustControl-class, 23
- * **cluster**
 - bclust, 8
 - bootFlexclust, 10
 - cclust, 15
 - clusterSim, 17
 - conversion, 18
 - dist2, 21
 - distances, 22
 - info, 28
 - kcca, 29
 - kcca2df, 32
 - parameters, 35
 - qtclust, 42
 - randIndex, 43
 - slsaplot, 52
 - slswFlexclust, 53
 - stepFlexclust, 54
- * **color**
 - flxColors, 25
- * **datagen**
 - priceFeature, 37
- * **datasets**
 - achieve, 3
 - auto, 3
 - birth, 10
 - bundestag, 12
 - dentitio, 20
 - milk, 32
 - Nclus, 33
 - nutrient, 34
 - vacmot, 58
 - volunteers, 59
- * **hplot**
 - barplot-methods, 5
 - bwplot-methods, 14
 - histogram-methods, 26
 - image-methods, 27
 - pairs-methods, 34
 - plot-methods, 35
 - projAxes, 38
 - propBarchart, 40
 - randomTour, 45
 - shadow, 48
 - shadowStars, 50
 - slsaplot, 52
 - slswFlexclust, 53
 - stripes, 56
- * **methods**
 - barplot-methods, 5
 - bwplot-methods, 14
 - histogram-methods, 26
 - image-methods, 27
 - pairs-methods, 34
 - plot-methods, 35
 - predict-methods, 37
 - randomTour, 45
 - relabel, 47
 - shadow, 48
 - shadowStars, 50
- * **multivariate**
 - bclust, 8
 - dist2, 21
- [[, stepFlexclust, ANY, missing-method (stepFlexclust), 54
- achieve, 3
- arrows, 39
- as.kcca (conversion), 18
- auto, 3
- barchart, 41
- barchart, bclust-method (barplot-methods), 5
- barchart, hclust-method (barplot-methods), 5
- barchart, kcca-method (barplot-methods), 5

- barchart, kccasimple-method
(barplot-methods), 5
- barplot, kcca-method (barplot-methods), 5
- barplot, kccasimple-method
(barplot-methods), 5
- barplot-methods, 5
- bclust, 8
- birth, 10
- bootFlexclust, 10
- bootFlexclust-class (bootFlexclust), 10
- boxplot, bootFlexclust-method
(bootFlexclust), 10
- boxplot, resampleFlexclust-method
(slswFlexclust), 53
- btw2002 (bundestag), 12
- btw2005 (bundestag), 12
- btw2009 (bundestag), 12
- bundestag, 12
- bwplot, 41
- bwplot, bclust-method (bwplot-methods),
14
- bwplot, kcca-method (bwplot-methods), 14
- bwplot, kccasimple-method
(bwplot-methods), 14
- bwplot-methods, 14
- cclust, 8, 9, 15, 24, 31, 55
- cclustControl (flexclustControl-class),
23
- cclustControl-class
(flexclustControl-class), 23
- centAngle (distances), 22
- centMean (distances), 22
- centMedian (distances), 22
- centOptim (distances), 22
- centOptim01 (distances), 22
- clusters, bclust, missing-method
(bclust), 8
- clusters, flexclust, ANY-method
(predict-methods), 37
- clusters, flexclust, missing-method
(predict-methods), 37
- clusterSim, 17, 48
- clusterSim, kcca-method (clusterSim), 17
- clusterSim, kccasimple-method
(clusterSim), 17
- coerce, 19
- coerce, kccasimple, kmeans-method
(conversion), 18
- coerce, list, cclustControl-method
(flexclustControl-class), 23
- coerce, list, flexclustControl-method
(flexclustControl-class), 23
- coerce, NULL, cclustControl-method
(flexclustControl-class), 23
- coerce, NULL, flexclustControl-method
(flexclustControl-class), 23
- comPart (randIndex), 43
- comPart, flexclust, flexclust-method
(randIndex), 43
- comPart, flexclust, numeric-method
(randIndex), 43
- comPart, numeric, flexclust-method
(randIndex), 43
- comPart, numeric, numeric-method
(randIndex), 43
- conversion, 18
- Cutree (conversion), 18
- cutree, 19
- data.matrix, 29
- densityplot, bootFlexclust-method
(bootFlexclust), 10
- densityplot, resampleFlexclust-method
(slswFlexclust), 53
- dentitio, 20
- dist, 8, 21
- dist2, 21
- distances, 22, 31
- distAngle (distances), 22
- distCanberra (distances), 22
- distCor (distances), 22
- distEuclidean (distances), 22
- distJaccard (distances), 22
- distManhattan (distances), 22
- distMax (distances), 22
- distMinkowski (distances), 22
- flexclust-class (kcca), 29
- flexclustControl
(flexclustControl-class), 23
- flexclustControl-class, 23
- flxColors, 25
- flxPalette (flxColors), 25
- getModel (stepFlexclust), 54
- getModel, stepFlexclust-method
(stepFlexclust), 54

- gpar, 56
- grep, 12
- groupBWplot (propBarchart), 40
- hcl, 25
- hclust, 8, 9, 19
- histogram, 26
- histogram,kccasimple,data.frame-method (histogram-methods), 26
- histogram,kccasimple,matrix-method (histogram-methods), 26
- histogram,kccasimple,missing-method (histogram-methods), 26
- histogram-methods, 26
- image,kcca-method (image-methods), 27
- image,kccasimple-method (image-methods), 27
- image-methods, 27
- info, 28, 28
- info,flexclust,character-method (info), 28
- infoCheck, 28
- kcca, 16, 24, 27, 29, 42, 55
- kcca-class (kcca), 29
- kcca2df, 32
- kccaFamily, 22
- kccaFamily (kcca), 29
- kccaFamily-class (kcca), 29
- kccasimple-class (kcca), 29
- kmeans, 18
- kruskal.test, 41
- makeCluster, 11
- milk, 32
- Nclus, 33
- nutrient, 34
- p.adjust, 40
- pairs,kcca-method (pairs-methods), 34
- pairs,kccasimple-method (pairs-methods), 34
- pairs-methods, 34
- pam, 18
- panelShadowBP (shadowStars), 50
- panelShadowSkeleton (shadowStars), 50
- panelShadowStripes (shadowStars), 50
- panelShadowViolin (shadowStars), 50
- parameters, 35
- parameters,bclust-method (bclust), 8
- parameters,kccasimple-method (parameters), 35
- placeLabels (projAxes), 38
- placeLabels,projAxes-method (projAxes), 38
- plot,bclust,missing-method (bclust), 8
- plot,bootFlexclust,missing-method (bootFlexclust), 10
- plot,kcca,missing-method (plot-methods), 35
- plot,kccasimple,missing-method (plot-methods), 35
- plot,projAxes,missing-method (projAxes), 38
- plot,resampleFlexclust,missing-method (slswFlexclust), 53
- plot,shadow,ANY-method (shadow), 48
- plot,Silhouette,ANY-method (shadow), 48
- plot,stepFlexclust,missing-method (stepFlexclust), 54
- plot-methods, 35
- plot.priceFeature (priceFeature), 37
- prcomp, 34, 36, 39, 50
- predict,kccasimple-method (predict-methods), 37
- predict-methods, 37
- priceFeature, 37
- projAxes, 38
- projAxes-class (projAxes), 38
- prop.test, 41
- propBarchart, 40
- propBarchart-class (propBarchart), 40
- qtclust, 42
- randIndex, 43
- randIndex,ANY,ANY-method (randIndex), 43
- randIndex,table,missing-method (randIndex), 43
- randomTour, 45
- randomTour,ANY-method (randomTour), 45
- randomTour,flexclust-method (randomTour), 45
- randomTour,matrix-method (randomTour), 45
- randomTourMatrix (randomTour), 45
- relabel, 47, 52

- relabel,kccasimple,character-method
(relabel), 47
- relabel,kccasimple,integer-method
(relabel), 47
- relabel,kccasimple,missing-method
(relabel), 47
- relabel,stepFlexclust,character-method
(relabel), 47
- relabel,stepFlexclust,integer-method
(relabel), 47
- relabel,stepFlexclust,missing-method
(relabel), 47

- seriate, 48
- shadow, 48, 51
- shadow,kccasimple-method (shadow), 48
- shadowStars, 50
- show,bootFlexclust-method
(bootFlexclust), 10
- show,kccasimple-method (kcca), 29
- show,propBarchart-method
(propBarchart), 40
- show,resampleFlexclust-method
(slswFlexclust), 53
- show,shadow-method (shadow), 48
- show,Silhouette-method (shadow), 48
- show,stepFlexclust-method
(stepFlexclust), 54
- Silhouette (shadow), 48
- silhouette, 49
- Silhouette,kcca-method (shadow), 48
- slsaplot, 52, 54
- slswFlexclust, 52, 53
- solve_LSAP, 31
- stepcclust (stepFlexclust), 54
- stepFlexclust, 11, 31, 52, 53, 54
- stepFlexclust-class (stepFlexclust), 54
- stripes, 56
- summary,bootFlexclust-method
(bootFlexclust), 10
- summary,kccasimple-method (kcca), 29
- summary,propBarchart-method
(propBarchart), 40
- summary,resampleFlexclust-method
(slswFlexclust), 53

- text, 39
- title, 41

- vacmot, 58
- vacmot6 (vacmot), 58
- vacmotdesc (vacmot), 58
- volunteers, 59

- x11, 46