

Package ‘muHVT’

October 13, 2022

Type Package

Date 2022-04-20

Title Constructing Hierarchical Voronoi Tessellations and Overlay
Heatmap for Data Analysis

Version 2.0.0

Description Constructing hierarchical voronoi tessellations for a given data set and overlay heatmap for variables at various levels of the tessellations for in-depth data analysis. See <https://en.wikipedia.org/wiki/Voronoi_diagram> for more information. Credits to Mu Sigma for their continuous support throughout the development of the package.

License Apache License 2.0

Encoding UTF-8

Imports MASS, deldir, grDevices, splancs, sp, conf.design, Hmisc,
stats, dplyr, purrr, gtools, magrittr, plyr, polyclip, rgeos,
ggplot2, tidyr, scales, patchwork, cluster, reshape2, tidyverse

Depends R (>= 3.6.0)

BugReports <https://github.com/Mu-Sigma/muHVT/issues>

URL <https://github.com/Mu-Sigma/muHVT>

RoxygenNote 7.1.2

Suggests knitr, rmarkdown, testthat, geozoo, kableExtra, plotly

VignetteBuilder knitr

NeedsCompilation no

Author Zubin Dowlaty [aut],
Shubhra Prakash [ctb],
Sangeet Moy Das [ctb],
Sunuganty Achyut Raj [ctb],
Praditi Shah [ctb],
Avinash Joshi [ctb],
Meet Dave [ctb],
Mu Sigma, Inc. [cre]

Maintainer ``Mu Sigma, Inc." <ird.experiencelab@mu-sigma.com>

Repository CRAN

Date/Publication 2022-04-27 12:40:02 UTC

R topics documented:

exploded_hmap	2
get_cell_id	3
HVT	4
hvtHmap	6
multiNormalDist	8
plotDiag	9
plotHVT	9
predictHVT	10
qeHistPlot	12

Index	13
--------------	-----------

exploded_hmap	<i>Function to construct an interactive 3D heatmap overlay for each HVT Level</i>
---------------	---

Description

Function to construct an interactive 3D heatmap overlay for each HVT Level

Usage

```
exploded_hmap(
  hvt.results,
  child.level = NULL,
  hmap.cols = NULL,
  nclust.hmap = NULL,
  separation_width = 7,
  layer_opacity = c(0.5, 0.75, 0.99),
  dim_size = 1000,
  ...
)
```

Arguments

<code>hvt.results</code>	List. A list of <code>hvt.results</code> obtained from the HVT function.
<code>child.level</code>	Numeric. Indicating the level for which the heat map is to be plotted.
<code>hmap.cols</code>	Numeric or Character. The column number of column name from the dataset indicating the variables for which the heat map is to be plotted.
<code>nclust.hmap</code>	Numeric. An integer indicating the number of clusters per hierarchy (level)
<code>separation_width</code>	Numeric. An integer indicating the width between two Levels
<code>layer_opacity</code>	Numeric. A vector indicating the opacity of each layer/ level
<code>dim_size</code>	Numeric. An integer indicating the dimension size used to create the matrix for the plot
<code>...</code>	<code>color.vec</code> and <code>line.width</code> can be passed from here

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>

get_cell_id

Function to generate cell ID based on 1D sammons projection

Description

To generate cell id for the multivariate data, the data is being projected from n-dimensions to 1-dimension and the cell id is being assigned by ordering these values and finding the corresponding indexes. The output CellID gets appended to the HVT model.

Usage

```
get_cell_id(hvt.results)
```

Arguments

`hvt.results` List. A list of hvt.results obtained from the HVT function.

Details

Cell ID

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>

Examples

```
data(USArrests)
hvt.results <- list()
hvt.results <- HVT(USArrests, nclust = 15, depth = 1, quant.err = 0.2,
  distance_metric = "L1_Norm", error_metric = "mean",
  projection.scale = 10, normalize = TRUE,
  quant_method="kmeans",diagnose=TRUE)
plotHVT(hvt.results, line.width = c(0.8), color.vec = c('#141B41'),
  maxDepth = 1)
get_cell_id (hvt.results)
```

Description

Main function to construct hierarchical voronoi tessellations.

Usage

```
HVT(
  dataset,
  nclust = 15,
  depth = 3,
  quant.err = 0.2,
  projection.scale = 10,
  normalize = TRUE,
  distance_metric = c("L1_Norm", "L2_Norm"),
  error_metric = c("mean", "max"),
  quant_method = "kmeans",
  diagnose = TRUE,
  hvt_validation = FALSE,
  train_validation_split_ratio = 0.8
)
```

Arguments

<code>dataset</code>	Data frame. A data frame with different columns is given as input.
<code>nclust</code>	Numeric. An integer indicating the number of clusters per hierarchy (level)
<code>depth</code>	Numeric. An integer indicating the number of levels. (1 = No hierarchy, 2 = 2 levels, etc ...)
<code>quant.err</code>	Numeric. A number indicating the quantization error threshold.
<code>projection.scale</code>	Numeric. A number indicating the scale factor for the tessellations so as to visualize the sub-tessellations well enough.
<code>normalize</code>	Logical. A logical value indicating if the columns in your dataset should be normalized. Default value is TRUE.
<code>distance_metric</code>	character. The distance metric can be 'L1_Norm' or 'Manhattan'. L1_Norm is selected by default.
<code>error_metric</code>	character. The error metric can be "mean" or "max". mean is selected by default
<code>quant_method</code>	character. The quant_method can be "kmeans" or "kmedoids". kmeans is selected by default
<code>diagnose</code>	Logical. A logical value indicating if the diagnose is required. Default value is TRUE.

`hvt_validation` Logical. A logical value indicating if the MAD values are to tested for validation set. Default value is FALSE.

`train_validation_split_ratio`
 Numeric. A numeric value indicating the train and validation split ratio.

Details

This is the main function to construct hierarchical voronoi tessellations. The `hvt` function is called from this function. The output of the `hvt` function is hierarchical clustered data which will be the input for constructing tessellations. The data is then represented in 2d coordinates and the tessellations are plotted using these coordinates as centroids. For subsequent levels, transformation is performed on the 2d coordinates to get all the points within its parent tile. Tessellations are plotted using these transformed points as centroids. The lines in the tessellations are chopped in places so that they do not protrude outside the parent polygon. This is done for all the subsequent levels.

Value

A list that contains the hierarchical tessellation information. This list has to be given as input argument to plot the tessellations.

[[1]] List. Information about the tessellation co-ordinates - level wise

[[2]] List. Information about the polygon co-ordinates - level wise

[[3]] List. Information about the hierarchical vector quantized data - level wise

[[4]] List. Information about the model diagnosis- selected level

[[5]] List. Information about the MAD values and percentage anomalies for validation dataset

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>, Sangeet Moy Das <sangeet.das@mu-sigma.com>

See Also

[plotHVT](#)
[hvtHmap](#)

Examples

```
data(USArrests)
hvt.results <- list()
hvt.results <- HVT(USArrests, nclust = 15, depth = 1, quant.err = 0.2,
  distance_metric = "L1_Norm", error_metric = "mean",
  projection.scale = 10, normalize = TRUE,
  quant_method="kmeans",diagnose=TRUE)
plotHVT(hvt.results, line.width = c(0.8), color.vec = c('#141B41'),
  maxDepth = 1)

hvt.results <- list()
hvt.results <- HVT(USArrests, nclust = 15, depth = 3, quant.err = 0.2,
  distance_metric = "L1_Norm", error_metric = "mean",
```

```

        projection.scale = 10, normalize = TRUE,
        quant_method="kmeans",diagnose=TRUE)
plotHVT(hvt.results, line.width = c(1.2,0.8,0.4), color.vec = c('#141B41', '#0582CA', '#8BA0B4'),
        maxDepth = 3)

```

hvtHmap

Heat Map over Hierarchical Voronoi Tessellations

Description

Main function to construct heatmap overlay for hierarchical voronoi tessellations.

Usage

```

hvtHmap(
  hvt.results,
  dataset,
  child.level,
  hmap.cols,
  color.vec = NULL,
  line.width = NULL,
  centroid.size = 3,
  pch = 21,
  palette.color = 6,
  previous_level_heatmap = T,
  show.points = F,
  asp = 1,
  ask = T,
  tess.label = NULL,
  quant.error.hmap = NULL,
  nclust.hmap = NULL,
  label.size = 0.5,
  ...
)

```

Arguments

<code>hvt.results</code>	List. A list of <code>hvt.results</code> obtained from the HVT function.
<code>dataset</code>	Data frame. The input data set.
<code>child.level</code>	Numeric. Indicating the level for which the heat map is to be plotted.
<code>hmap.cols</code>	Numeric or Character. The column number of column name from the dataset indicating the variables for which the heat map is to be plotted.
<code>color.vec</code>	Vector. A color vector such that <code>length(color.vec) = (child.level - 1)</code> . (default = NULL)
<code>line.width</code>	Vector. A line width vector such that <code>length(line.width) = (child.level - 1)</code> . (default = NULL)

<code>centroid.size</code>	Numeric. Indicating the centroid size of the first level. (default = 3)
<code>pch</code>	Numeric. Indicating the centroid's symbol type. (default = 21)
<code>palette.color</code>	Numeric. Indicating the heat map color palette. 1 - rainbow, 2 - heat.colors, 3 - terrain.colors, 4 - topo.colors, 5 - cm.colors, 6 - seas color. (default = 6)
<code>previous_level_heatmap</code>	Logical. If TRUE, the heatmap of previous level will be overlayed on the heatmap of selected level. If #' FALSE, the heatmap of only selected level will be plotted
<code>show.points</code>	Logical. Indicating if the centroids should be plotted on the tessellations. (default = FALSE)
<code>asp</code>	Numeric. Indicating the aspect ratio type. For flexible aspect ratio set, asp = NA. (default = 1)
<code>ask</code>	Logical. If TRUE (and the R session is interactive) the user is asked for input, before a new figure is drawn. (default = TRUE)
<code>tess.label</code>	Vector. A vector for labelling the tessellations. (default = NULL)
<code>quant.error.hmap</code>	Numeric. A number indicating the quantization error threshold.
<code>nclust.hmap</code>	Numeric. An integer indicating the number of clusters per hierarchy (level)
<code>label.size</code>	Numeric. The size by which the tessellation labels should be scaled. (default = 0.5)
<code>...</code>	The ellipsis is passed to it as additional argument. (Used internally)

Details

The output of the HVT function has all the required information about the HVT. Now a heat map is overlayed over this HVT. The user defines the level and also those variables of the data for which the heat map is to be plotted. Now for each variable a separate heat map is plotted. The plot area is divided into 2 screens where the first screen is relatively large and will have the heat map. The second screen is small and contains the gradient scale. To plot the heat map, the data is first normalized. The gradient scale is divided into 'n' regions(500 is the set default). Using the normalized data, the different regions into which the data items fall are found. Each data item is now having a region on the gradient scale. This color is filled in the tile corresponding to the data item. This procedure is done for all the tiles for that level to get the complete heat map. Once the heat map is ready, the higher level tessellations are plotted to represent the hierarchies. The size of the centroids, the thickness of the lines and the color of the tessellation lines can be given as input by the user. Appropriate values for these parameters should be given to identify the hierarchies properly. In the second screen the gradient scale is plotted. The heat maps and hierarchical tessellations are obtained for all the desired variables.

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>, Sangeet Moy Das <sangeet.das@mu-sigma.com>

See Also

[plotHVT](#)

Examples

```
data(USArrests)
hvt.results <- list()
hvt.results <- HVT(USArrests, nclust = 15, depth = 1, quant.err = 0.2,
                  distance_metric = "L1_Norm", error_metric = "mean",
                  projection.scale = 10, normalize = TRUE,
                  quant_method="kmeans",diagnose=TRUE)
hvtHmap(hvt.results, USArrests, child.level = 1,hmap.cols = 'Murder',
        line.width = c(0.2), color.vec = c('#141B41'),palette.color = 6,
        quant.error.hmap = 0.2,nclust.hmap = 6)
```

multiNormalDist

Multivariate normal distribution

Description

Function to generate multivariate normal distribution where each variable has a standard normal distribution $N(0,1)$

Usage

```
multiNormalDist(sample.size = 2500, ncol = 5)
```

Arguments

sample.size	Numeric. Indicating the sample size for distribution
ncol	Numeric. Indicating the number of columns

Details

The multivariate normal distribution, multivariate Gaussian distribution, or joint normal distribution is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. It is a vector in multiple normally distributed variables, such that any linear combination of the variables is also normally distributed.

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>

Examples

```
multiNormalDist(2500,2)
x=multiNormalDist(2500,2)
hist(x[,1])
```

plotDiag	<i>Make the diagnostic plots for hierarchical voronoi tessellations model.</i>
----------	--

Description

Make the diagnostic plots for hierarchical voronoi tessellations model.

Usage

```
plotDiag(model_obj)
```

Arguments

model_obj List. A list of model_obj obtained from the HVT function or prediction object

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>

See Also

[plotHVT](#)

plotHVT	<i>Plot the hierarchical tessellations. Main plotting function to construct hierarchical voronoi tessellations.</i>
---------	---

Description

Plot the hierarchical tessellations.

Main plotting function to construct hierarchical voronoi tessellations.

Arguments

hvt.results	List. A list containing the output of HVT function which has the details of the tessellations to be plotted.
line.width	Numeric Vector. A vector indicating the line widths of the tessellation boundaries for each level.
color.vec	Vector. A vector indicating the colors of the boundaries of the tessellations at each level.
pch1	Numeric. Symbol type of the centroids of the tessellations (parent levels). Refer points . (default = 21)
centroid.size	Numeric. Size of centroids of first level tessellations. (default = 3)
title	String. Set a title for the plot. (default = NULL)
maxDepth	Numeric. An integer indicating the number of levels. (default = NULL)

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>, Sangeet Moy Das <sangeet.das@mu-sigma.com>

See Also

[HVT](#)
[hvtHmap](#)

Examples

```
data("USArrests", package="datasets")

hvt.results <- list()
hvt.results <- HVT(USArrests, nclust = 15, depth = 1, quant.err = 0.2,
  distance_metric = "L1_Norm", error_metric = "mean",
  projection.scale = 10, normalize = TRUE,
  quant_method="kmeans", diagnose=TRUE)
plotHVT(hvt.results, line.width = c(0.8), color.vec = c('#141B41'),
  maxDepth = 1)
```

predictHVT

Predict which cell and what level each point in the test dataset belongs to

Description

Predict which cell and what level each point in the test dataset belongs to

Usage

```
predictHVT(
  data,
  hvt.results.model,
  child.level = 1,
  mad.threshold = 0.2,
  line.width = c(0.6, 0.4, 0.2),
  color.vec = c("#141B41", "#6369D1", "#D8D2E1"),
  normalize = T,
  distance_metric = "L1_Norm",
  error_metric = "max",
  yVar = NULL,
  ...
)
```

Arguments

<code>data</code>	List. A dataframe containing test dataset. The dataframe should have atleast one variable used while training. The variables from this dataset can also be used to overlay as heatmap
<code>hvt.results.model</code>	A list of <code>hvt.results.model</code> obtained from HVT function while performing hierarchical vector quantization on train data
<code>child.level</code>	A number indicating the level for which the heat map is to be plotted.(Only used if <code>hmap.cols</code> is not NULL)
<code>mad.threshold</code>	A numeric values indicating the permissible Mean Absolute Deviation
<code>line.width</code>	Vector. A line width vector
<code>color.vec</code>	Vector. A color vector
<code>normalize</code>	Logical. A logical value indicating if the columns in your dataset should be normalized. Default value is TRUE.
<code>distance_metric</code>	character. The distance metric can be 'Euclidean' or 'Manhattan'. Euclidean is selected by default.
<code>error_metric</code>	character. The error metric can be "mean" or "max". mean is selected by default
<code>yVar</code>	character. Name of the dependent variable(s)
<code>...</code>	<code>color.vec</code> and <code>line.width</code> can be passed from here

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>, Sangeet Moy Das <sangeet.das@mu-sigma.com>

See Also

[HVT](#)
[hvtHmap](#)

Examples

```
data(USArrests)
#Split in train and test

train <- USArrests[1:40,]
test <- USArrests[41:50,]

hvt.results <- list()
hvt.results <- HVT(train, nclust = 15, depth = 1, quant.err = 0.2,
  distance_metric = "L1_Norm", error_metric = "mean",
  projection.scale = 10, normalize = TRUE,
  quant_method="kmeans", diagnose=TRUE)

predictions <- predictHVT(test,hvt.results, child.level=2, mad.threshold = 0.2)
print(predictions$scoredPredictedData)
```

qeHistPlot

plotDiag Make the diagnostic plots for hierarchical voronoi tessellations model.

Description

plotDiag

Make the diagnostic plots for hierarchical voronoi tessellations model.

Usage

```
qeHistPlot(hvt.results, hvt.predictions)
```

Arguments

`hvt.results` List. A list of hvt.results obtained from the HVT function.

`hvt.predictions`

List. A list of hvt.predictions obtained from the Predict function.

Author(s)

Shubhra Prakash <shubhra.prakash@mu-sigma.com>

See Also

[plotHVT](#)

Index

- * **diagnostics**
 - plotDiag, 9
 - qeHistPlot, 12
- * **hplot**
 - HVT, 4
 - hvtHmap, 6
 - plotHVT, 9
- * **predict**
 - predictHVT, 10
- exploded_hmap, 2
- get_cell_id, 3
- HVT, 4, 10, 11
- hvtHmap, 5, 6, 10, 11
- multiNormalDist, 8
- plotDiag, 9
- plotHVT, 5, 7, 9, 9, 12
- points, 9
- predictHVT, 10
- qeHistPlot, 12