# Package ‘muHVT’

**March 3, 2023**

## Type
Package

## Date
2023-03-03

## Title
Constructing Hierarchical Voronoi Tessellations and Overlay Heatmap for Data Analysis

## Version
3.0.0

## Description
The muHVT package is a collection of R functions to facilitate building topology preserving maps for rich multivariate data. See [https://en.wikipedia.org/wiki/Voronoi_diagram](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, magrittr, polyclip, rgeos, ggplot2, tidyR, scales, cluster, reshape2, plyr

## 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, data.table

## VignetteBuilder
knitr

## NeedsCompilation
no

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

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

Description

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

Usage

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

Arguments

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

Author(s)
Shubhra Prakash <shubhra.prakash@mu-sigma.com>

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

```r
data(USArrests)
hvt.results <- list()
hvt.results <- HVT(USArrests, n_cells = 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)
```

---

**HVT**

*Constructing Hierarchical Voronoi Tessellations*

**Description**

Main function to construct hierarchical voronoi tessellations.

**Usage**

```r
HVT(
  dataset,
  min_compression_perc = NA,
  n_cells = NA,
  depth = 1,
  quant.err = 0.2,
  projection.scale = 10,
  normalize = FALSE,
  distance_metric = c("L1_Norm", "L2_Norm"),
  error_metric = c("mean", "max"),
  quant_method = c("kmeans", "kmedoids"),
  scale_summary = NA,
  diagnose = FALSE,
  hvt_validation = FALSE,
  train_validation_split_ratio = 0.8
)
```

**Arguments**

- **dataset**
  - Data frame. A data frame with different columns is given as input.

- **min_compression_perc**
  - Numeric. An integer indicating the minimum percent compression rate to be achieved for the dataset

- **n_cells**
  - Numeric. An integer indicating the number of cells per hierarchy (level)

- **depth**
  - Numeric. An integer indicating the number of levels. (1 = No hierarchy, 2 = 2 levels, etc ... )
quant.err Numeric. A number indicating the quantization error threshold.

projection.scale Numeric. A number indicating the scale factor for the tessellations so as to visualize the sub-tessellations well enough.

normalize Logical. A logical value indicating if the columns in your dataset should be normalized. Default value is TRUE.

distance_metric character. The distance metric can be "L1_Norm" or "Manhattan". L1_Norm is selected by default.

error_metric character. The error metric can be "mean" or "max". mean is selected by default

quant_method character. The quant_method can be "kmeans" or "kmedoids". kmeans is selected by default

scale_summary List. A list with mean and standard deviation values for all the features in the dataset. Pass the scale summary when the input dataset is already scaled or normalize is set to False.

diagnose 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 hvq function is called from this function. The output of the hvq 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>, Shantanu Vaidya <shantanu.vaidya@mu-sigma.com>
See Also

plotHVT
hvtHmap

Examples

data(USArrests)
hvt.results <- list()
hvt.results <- HVT(USArrests, min_compression_perc = 70, quant.err = 0.2,
distance_metric = "L1_Norm", error_metric = "mean",
projection.scale = 10, normalize = TRUE,
quant_method="kmeans")
plotHVT(hvt.results, line.width = c(0.8), color.vec = c('#141B41'),
maxDepth = 1)

hvt.results <- list()
hvt.results <- HVT(USArrests, n_cells = 15, depth = 3, quant.err = 0.2,
distance_metric = "L1_Norm", error_metric = "mean",
projection.scale = 10, normalize = TRUE,
quant_method="kmeans")
plotHVT(hvt.results, line.width = c(1.2,0.8,0.4), color.vec = c('#141B41','#0582CA','#8BA0B4'),
maxDepth = 3)
Arguments

- **hvt.results** List. A list of hvt.results obtained from the HVT function.
- **dataset** Data frame. The input data set.
- **child.level** Numeric. Indicating the level for which the heat map is to be plotted.
- **hmap.cols** Numeric or Character. The column number of column name from the dataset indicating the variables for which the heat map is to be plotted.
- **color.vec** Vector. A color vector such that length(color.vec) = (child.level - 1). (default = NULL)
- **line.width** Vector. A line width vector such that length(line.width) = (child.level - 1). (default = NULL)
- **centroid.size** Numeric. Indicating the centroid size of the first level. (default = 3)
- **pch** Numeric. Indicating the centroid’s symbol type. (default = 21)
- **palette.color** 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)
- **previous_level_heatmap** 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.
- **show.points** Logical. Indicating if the centroids should be plotted on the tessellations. (default = FALSE)
- **asp** Numeric. Indicating the aspect ratio type. For flexible aspect ratio set, asp = NA. (default = 1)
- **ask** Logical. If TRUE (and the R session is interactive) the user is asked for input, before a new figure is drawn. (default = TRUE)
- **tess.label** Vector. A vector for labelling the tessellations. (default = NULL)
- **quant.error.hmap** Numeric. A number indicating the quantization error threshold.
- **n_cells.hmap** Numeric. An integer indicating the number of cells/clusters per hierarchy (level)
- **label.size** Numeric. The size by which the tessellation labels should be scaled. (default = 0.5)
- **...** 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, n_cells = 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, n_cells.hmap = 6)

mlayerHVT

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

mlayerHVT(
data,
hvt_mapA,
hvt_mapB,
hvt_mapC,
mad.threshold = 0.2,
normalize = T,
distance_metric = "L1_Norm",
error_metric = "max",
child.level = 1,
line.width = c(0.6, 0.4, 0.2),
color.vec = c("#141B41", ",#6369D1", ",#D8D2E1"),
yVar = NULL,
...
)

Arguments

data Data Frame. A dataframe containing test dataset. The dataframe should have at least one variable used while training. The variables from this dataset can also be used to overlay as heatmap

hvt_mapA A list of hvt.results.model obtained from HVT function while performing hierarchical vector quantization on train data

hvt_mapB A list of removed outlier rows using removedOutliers function

hvt_mapC A list of hvt.results.model obtained from HVT function while performing hierarchical vector quantization on train data without outlier(s)

mad.threshold A numeric values indicating the permissible Mean Absolute Deviation

normalize Logical. A logical value indicating if the columns in your dataset should be normalized. Default value is TRUE.

distance_metric character. The distance metric can be 'Euclidean' or "Manhattan". Euclidean is selected by default.

error_metric character. The error metric can be "mean" or "max". mean is selected by default

child.level A number indicating the level for which the heat map is to be plotted.(Only used if hmap.cols is not NULL)

line.width Vector. A line width vector

color.vec Vector. A color vector

yVar character. Name of the dependent variable(s)

... color.vec and line.width can be passed from here

Author(s)

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

See Also

HVT
hvtHmap
Examples

data(USArrests)

# Split in train and test
train <- USArrests[1:40,]
test <- USArrests[41:50,]

hvt_mapA <- list()
hvt_mapA <- HVT(train, min_compression_perc = 70, quant.err = 0.2,
                 distance_metric = "L1_Norm", error_metric = "mean",
                 projection.scale = 10, normalize = TRUE,
                 quant_method="kmeans")

identified_outlier_cells <- c(2, 10)
output_list <- removeOutliers(identified_outlier_cells, hvt_mapA)
hvt_mapB <- output_list[[1]]
dataset_without_outliers <- output_list[[2]]

mapA_scale_summary = hvt_mapA[[3]]$scale_summary
hvt_mapC <- list()
hvt_mapC <- HVT(dataset_without_outliers, n_cells = 15,
                depth = 2, quant.err = 0.2, distance_metric = "L1_Norm",
                error_metric = "max", quant_method = "kmeans",
                projection.scale = 10, normalize = FALSE, scale_summary = mapA_scale_summary)

predictions <- list()
predictions <- mlayerHVT(test, hvt_mapA, hvt_mapB, hvt_mapC)

---

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])
```

Description

Plot the identified outlier cell(s) in the voronoi tessellations map. Plotting function to construct hierarchical voronoi tessellations and highlight the cells using the compressed HVT map.

Usage

```
plotCells(
  plot.cells,  # Vector. A vector indicating the cells to be highlighted in the map
  hvt.map,     # List. A list containing the output of HVT function which has the details of the tessellations to be plotted.
  line.width = c(0.6),  # Numeric Vector. A vector indicating the line widths of the tessellation boundaries for each level.
  color.vec = c("#141B41"),
  pch1 = 21,
  centroid.size = 0.5,
  title = NULL,
  maxDepth = 1
)
```

Arguments

- `plot.cells`
- `hvt.map`
- `line.width`
### Arguments

- **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 to `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)

Shantanu Vaidya <shantanu.vaidya@mu-sigma.com>

### See Also

- `HVT`
- `hvtHmap`

---

**plotDiag**  
Make the diagnostic plots for hierarchical voronoi tessellations model.

### Description

Make the diagnostic plots for hierarchical voronoi tessellations model.

### Usage

```r
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

Plot the hierarchical tessellations. Main plotting function to construct hierarchical voronoi tessellations.

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, n_cells = 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("#1B41B41"),
  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**

```r
predictHVT(
  data,  # List. A dataframe containing test dataset. The dataframe should have atleast one
  hvt.results.model,  # A list of hvt.results.model obtained from HVT function while performing hierarchical vector quantization on train data
  child.level = 1,  # A number indicating the level for which the heat map is to be plotted.(Only used if hmap.cols is not NULL)
  mad.threshold = 0.2,  # A numeric values indicating the permissible Mean Absolute Deviation
  line.width = c(0.6, 0.4, 0.2),  # Vector. A line width vector
  color.vec = c("#141B41", "#6369D1", "#D8D2E1"),  # Vector. A color vector
  normalize = T,  # Logical. A logical value indicating if the columns in your dataset should be normalized. Default value is TRUE.
  distance_metric = "L1_Norm",  # character. The distance metric can be 'Euclidean" or "Manhattan". Euclidean is selected by default.
  error_metric = "max",  # character. The error metric can be "mean" or "max". mean is selected by default
  yVar = NULL,  # character. Name of the dependent variable(s)
  ...  # color.vec and line.width can be passed from here
)
```

**Arguments**

- **data**
  - 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

- **hvt.results.model**
  - A list of hvt.results.model obtained from HVT function while performing hierarchical vector quantization on train data

- **child.level**
  - A number indicating the level for which the heat map is to be plotted.(Only used if hmap.cols is not NULL)

- **mad.threshold**
  - A numeric values indicating the permissible Mean Absolute Deviation

- **line.width**
  - Vector. A line width vector

- **color.vec**
  - Vector. A color vector

- **normalize**
  - Logical. A logical value indicating if the columns in your dataset should be normalized. Default value is TRUE.

- **distance_metric**
  - character. The distance metric can be 'Euclidean" or "Manhattan". Euclidean is selected by default.

- **error_metric**
  - character. The error metric can be "mean" or "max". mean is selected by default

- **yVar**
  - character. Name of the dependent variable(s)

- **...**
  - color.vec and line.width 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, n_cells = 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>
**Description**

Remove identified outlier cell(s) from the dataset

**Usage**

```
removeOutliers(outlier_cells, hvt_results)
```

**Arguments**

- `outlier_cells` Vector. A vector with the cell number of the identified outliers
- `hvt_results` List. A list having the results of the compressed map i.e. output of HVT function

**Details**

This function is used to remove the identified outlier cell(s) from the dataset. It is recommended to run the HVT function before running this function. It takes input in the form of cell number of the outlier cell(s) identified using the output of the HVT function and the compressed map (hvt_mapA) generated using the HVT function. The output of this function is a list of two items: a new map having the data of removed outlier cell(s) and the subset of dataset without outliers.

**Value**

A list of two items: a map having the data of removed outlier cells and the subset of the dataset without outlier(s) which has to be passed as input argument to HVT function to generate another map

- `[[1]]` Dataframe. Information about the removed outlier cell(s)
- `[[2]]` Dataframe. Subset of dataset without the outlier cell(s)

**Author(s)**

Shantanu Vaidya <shantanu.vaidya@mu-sigma.com>

**See Also**

- `plotHVT`
- `HVT`
- `mlayerHVT`
Examples

data(USArrests)
hvt_mapA <- list()
hvt_mapA <- HVT(USArrests, min_compression_perc = 70, quant.err = 0.2,
  distance_metric = "L1_Norm", error_metric = "mean",
  projection.scale = 10, normalize = TRUE,
  quant_method="kmeans")
plotHVT(hvt_mapA, line.width = c(0.8), color.vec = c('#141B41'),
  maxDepth = 1)

identified_outlier_cells <- c(2, 10)
output_list <- removeOutliers(identified_outlier_cells, hvt_mapA)
hvt_mapB <- output_list[[1]]
dataset_without_outliers <- output_list[[2]]
Index

* diagnostics
  plotDiag, 12
  qeHistPlot, 15
* hplot
  HVT, 4
  hvtHmap, 6
  plotCells, 11
  plotHVT, 13
* mlayer
  mlayerHVT, 8
* predict
  predictHVT, 14
exploded_hmap, 2
get_cell_id, 3
HVT, 4, 9, 12, 13, 15, 16
hvtHmap, 6, 6, 9, 12, 13, 15
mlayerHVT, 8, 16
multiNormalDist, 10
plotCells, 11
plotDiag, 12
plotHVT, 6, 8, 12, 13, 16
points, 12, 13
predictHVT, 14
qeHistPlot, 15
removeOutliers, 16